



# Word Sense Alignment of Lexical Resources

Vom Fachbereich Informatik  
der Technischen Universität Darmstadt  
genehmigte

## **Dissertation**

zur Erlangung des akademischen Grades Dr.-Ing.

vorgelegt von  
**Michael Matuschek, M.Sc.**  
geboren in Hilden

Tag der Einreichung: 5. August 2014  
Tag der Disputation: 29. September 2014

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt  
Assoc. Prof. Roberto Navigli, Ph.D., Rom  
Prof. Dr. Karsten Weihe, Darmstadt

Darmstadt 2015  
D17

Please cite this document as

URN: urn:nbn:de:tuda-tuprints-43555

URL: <http://tuprints.ulb.tu-darmstadt.de/id/eprint/4355>

This document is provided by tuprints,  
E-Publishing-Service of the TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

[tuprints@ulb.tu-darmstadt.de](mailto:tuprints@ulb.tu-darmstadt.de)



This work is published under the following Creative Commons license:

Attribution – Non Commercial – No Derivative Works 3.0 Germany

<http://creativecommons.org/licenses/by-nc-nd/2.0/de/deed.en>

# Ehrenwörtliche Erklärung <sup>1</sup>

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades “Dr.-Ing.” mit dem Titel “Word Sense Alignment of Lexical Resources” selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 5. August 2014

---

Michael Matuschek, M.Sc.

---

<sup>1</sup>Gemäß §9 Abs. 1 der Promotionsordnung der TU Darmstadt



## Wissenschaftlicher Werdegang des Verfassers<sup>2</sup>

- 10/03 – 03/08 Studium der Informatik an der Heinrich-Heine-Universität  
Düsseldorf
- 06/06 Abschluss als Bachelor of Science  
Bachelor-Thesis: “Extraktion relationaler Daten für das Semantic  
Web in Oracle 10g”  
Gutachter: Prof. Dr. Stefan Conrad, Prof. Dr. Michael Leuschel
- 04/08 Abschluss als Master of Science  
Master-Thesis: “Temporal Aspects in Data Mining”  
Gutachter: Prof. Dr. Stefan Conrad, Prof. Dr. Martin Mauve
- seit 10/09 Wissenschaftlicher Mitarbeiter am Ubiquitous Knowledge  
Processing Lab, Technische Universität Darmstadt

---

<sup>2</sup>Gemäß §20 Abs. 3 der Promotionsordnung der TU Darmstadt



## Abstract

Lexical-semantic resources (LSRs) are a cornerstone for many areas of Natural Language Processing (NLP) such as word sense disambiguation or information extraction. LSRs exist in many varieties, focusing on different information types and languages, or being constructed according to different paradigms. However, the large number of different LSRs is still not able to meet the growing demand for large-scale resources for different languages and application purposes. Thus, the orchestrated usage of different LSRs is necessary in order to cover more words and senses, and also to have access to a richer knowledge representation when word senses are covered in more than one resource. In this thesis, we address the task of finding equivalent senses in these resources, which is known as *Word Sense Alignment* (WSA), and report various contributions to this area.

First, we give a formal definition of WSA and describe suitable evaluation metrics and baselines for this task. Then, we position WSA in the broad area of semantic processing by comparing it to related tasks from NLP and other fields, establishing that WSA indeed displays a unique set of properties and challenges which need to be addressed.

After that, we discuss the resources we employ for WSA, distinguishing between expert-built and collaboratively constructed resources. We give a brief description and refer to related work for each resource, and we discuss the collaboratively constructed, multilingual resource OmegaWiki in greater detail, as it has not been exhaustively covered in previous work and also presents a unique, concept-centered and language-agnostic structure, which makes it interesting for NLP applications. At the same time, we shed light on disadvantages of this approach and gaps in OmegaWiki's content. After the presentation of the resources, we perform a comparative analysis of them which focuses on their suitability for different approaches to WSA. In particular, we analyze their glosses as well as their structure and point out flaws and differences between them. Based on this, we motivate the selection of resource pairs we investigate and describe the WSA gold standard datasets they participate in. On top of the ones presented in previous work, we discuss four new datasets we created, filling gaps in the body of WSA research.

We then go on to present an alignment between Wiktionary and OmegaWiki, using a similarity-based framework. For the first time, it is applied to two collaboratively constructed resources. We improve this framework by adding a machine translation component, which we use to align WordNet and the German part of OmegaWiki. A cross-validation experiment with the English OmegaWiki (i.e. for the monolingual case) shows that both configurations perform comparably as only few errors are introduced by the translation component. This confirms the general validity of the idea.

Building on the observation that similarity-based approaches suffer from the insufficient lexical overlap between different glosses, we also present the novel alignment algorithm Dijkstra-WSA. It works on graph representations of LSRs induced, for instance, by semantic relations or links, and exploits the intuition that related senses are concentrated in adjacent regions of the resources. This algorithm performs competitively on six out of eight evaluation datasets, and we also present a combination with the similarity-based approach mentioned above in a backoff con-

figuration. This approach achieves a significant improvement over previous work on all considered datasets.

To further exploit the insight that text similarity-based and graph-based approaches complement each other, we also combine these notions in a machine learning framework. This way, we achieve a further overall improvement in terms of F-measure for four out of eight considered datasets, while for three others we could achieve a significant improvement in alignment precision and accuracy. We investigate different machine learning classifiers and conclude that Bayesian Networks show the most robust results across datasets. While we also discuss additional machine learning features, none of these lead to further improvements, which we consider proof that structure and glosses of the LSRs are sufficiently informative for finding equivalent senses in LSRs. Moreover, we discuss different approaches to aligning more than two resources at once (N-way alignment), which however do not yield satisfactory results. We also analyze the reasons for that and identify a great demand for future research.

The unified LSR UBY provides the greater context for this thesis. Its representation format UBY-LMF (based on the *Lexical Markup Framework* standard) reflects the structure and content of many different LSRs with the greatest possible level of accuracy, making them interoperable and accessible. We demonstrate how the standardization is operationalized, where OmegaWiki serves as a showcase for presenting the properties of UBY-LMF, including the representation of the sense alignments. We also discuss the final, instantiated resource UBY, as well as the Java-based API, which allows easy programmatic access to it, a web interface for conveniently browsing UBY's contents, and the alignment framework we used for our experiments, whose implementation was enabled by the standardization efforts and the API.

To demonstrate that sense alignments are indeed beneficial for NLP, we discuss different applications which make use of them. The clustering of fine-grained Germanet and WordNet senses by exploiting 1:n alignments to OmegaWiki, Wiktionary and Wikipedia significantly improves word sense disambiguation accuracy on standard evaluation datasets for German and English, while this approach is language-independent and does not require external knowledge or resource-specific feature engineering. The second scenario is computer-aided translation. We argue that the multilingual resources OmegaWiki and Wiktionary can be a useful source of knowledge, and especially translations, for this kind of applications. In this context, we also further discuss the results of the alignment we produce between them, and we give examples of the additional knowledge that becomes available through their combined usage.

Finally, we point out many directions for future work, not only for WSA, but also for the design of aligned resources such as UBY and the applications that benefit from them.





## Zusammenfassung

Lexikalisch-semantische Ressourcen (LSRs) sind ein Grundbaustein für viele Bereiche des Natural Language Processing (NLP), wie z.B. Lesartendisambiguierung oder Informationsextraktion. Es gibt LSRs in vielen Varianten, mit Schwerpunkten auf verschiedenen Informationstypen und Sprachen. Nichtsdestotrotz kann die große Zahl verschiedener LSRs den wachsenden Bedarf an umfangreichen Ressourcen für verschiedene Sprachen und Anwendungen nur unzureichend decken. Aus diesem Grund ist die kombinierte Nutzung verschiedener LSR nötig, um mehr Wörter und Bedeutungen abzudecken, und auch um Zugriff zu umfangreichem Wissen zu haben, wenn eine Wortbedeutung in mehreren Ressourcen vertreten ist. In dieser Arbeit adressieren wir die Aufgabenstellung, äquivalente Wortbedeutungen in diesen Ressourcen zu identifizieren. Dies bezeichnet man als *Word Sense Alignment* (WSA), und wir berichten über zahlreiche Beiträge zu diesem Forschungsfeld.

Zunächst definieren wir WSA und beschreiben mögliche Evaluationsmetriken und Baselines für diese Aufgabe. Danach verorten wir WSA im weiten Feld der semantischen Sprachverarbeitung, indem wir es zu verwandten Problemen in NLP sowie in anderen Bereichen in Bezug setzen. Dabei stellen wir fest, dass WSA einzigartige Anforderungen mit sich bringt, die berücksichtigt werden müssen.

Im Anschluss diskutieren wir die Ressourcen, die wir für WSA einsetzen, und unterscheiden dabei zwischen von Experten erstellten und kollaborativ erstellten Ressourcen. Während wir für die meisten Ressourcen einen kurzen Überblick geben, besprechen wir die kollaborative, mehrsprachige Ressource OmegaWiki ausführlich, da diese in früheren Arbeiten keine umfangreiche Beachtung fand und darüber hinaus eine einmalige, konzeptorientierte und sprachunabhängige Struktur hat, die sie für NLP-Anwendungen interessant macht. Wir weisen jedoch ebenso auf nachteilige Eigenschaften und Lücken in OmegaWiki hin, die daraus resultieren. Nach der Vorstellung der Ressourcen führen wir eine vergleichende Analyse durch, welche sich auf die Eignung verschiedener LSRs für unterschiedliche WSA-Ansätze konzentriert. Dabei analysieren wir insbesondere die Beschreibungen der Wortbedeutungen und die Struktur der Ressourcen, wobei wir Schwächen einzelner Ressourcen sowie Unterschiede zwischen diesen aufarbeiten. Basierend auf dieser Analyse motivieren wir die Auswahl von Ressourcenpaaren, die wir untersuchen. Wir beschreiben ebenso die WSA-Goldstandards bzw. Evaluationsdatensätze, an denen sie beteiligt sind. Neben denen, die bereits in früheren Arbeiten vorgestellt wurden, diskutieren wir auch vier von uns neu erstellte Datensätze.

Danach präsentieren wir ein Alignment zwischen Wiktionary und OmegaWiki, wobei wir auf einem ähnlichkeitsbasierten Ansatz aufbauen, welcher hier erstmals auf zwei kollaborativ erstellte Ressourcen angewendet wird. Wir erweitern diesen Ansatz um eine maschinelle Übersetzungskomponente, welche genutzt wird um WordNet und den deutschen Teil von OmegaWiki zu alignieren. Ein Vergleichsexperiment mit dem englischen OmegaWiki (d.h. für den monolingualen Fall) zeigt, dass beide Konfigurationen vergleichbare Ergebnisse erzielen, da die Übersetzungskomponente nur wenige Fehler macht. Dies bestätigt die Effektivität unseres Ansatzes.

Basierend auf der Beobachtung, dass ähnlichkeitsbasierte Verfahren an ihre Grenzen stoßen, falls die Überlappung zwischen Bedeutungsbeschreibungen unzureichend ist, stellen wir einen neuen Alignment-Algorithmus namens Dijkstra-WSA vor. Er

arbeitet auf Graphrepräsentationen der LSRs, die bspw. von semantischen Relationen oder Links induziert werden, und beruht auf der Intuition, dass verwandte Bedeutungen in benachbarten Regionen konzentriert sind. Der Algorithmus zeigt überzeugende Ergebnisse für sechs von acht Evaluationsdatensätzen, und wir präsentieren auch eine Kombination mit dem ähnlichkeitsbasierten Ansatz, welche eine signifikante Verbesserung zu früheren Arbeiten auf allen Datensätzen bewirkt.

Um die Erkenntnis, dass sich ähnlichkeitsbasierte und graphbasierte Verfahren ergänzen, besser auszunutzen, kombinieren wir beide Ansätze auch mit Hilfe von maschinellen Lernverfahren, womit wir eine weitere Verbesserung der Gesamtergebnisse (hinsichtlich F-Measure) für vier von acht Datensätzen erreichen, während wir für drei weitere einen signifikanten Anstieg in Precision und Accuracy feststellen. Wir untersuchen verschiedene maschinelle Lernverfahren, wobei Bayes'sche Netze die beste Gesamtleistung zeigen, und obwohl wir weitere Merkmale für das maschinelle Lernen untersuchen ist keine weitere Verbesserung der Ergebnisse möglich. Wir werten dies als Hinweis, dass die Struktur und die Beschreibungen der Wortbedeutungen ausreichend informativ sind, um äquivalente Bedeutungen in LSRs zu identifizieren. Weiterhin untersuchen wir verschiedene Ansätze, um mehr als zwei Ressourcen gleichzeitig zu alignieren, wobei wir jedoch keine befriedigenden Ergebnisse erzielen. Wir analysieren die Gründe hierfür und identifizieren zahlreiche Ansätze für zukünftige Arbeiten.

Die integrierte Ressource UBY bildet den größeren Rahmen für diese Arbeit. Das zugrunde liegende Repräsentationsformat UBY-LMF (basierend auf dem *Lexical Markup Framework*-Standard) spiegelt die Struktur und den Inhalt vieler verschiedener LSRs im größtmöglichen Detailgrad wider, wodurch sie interoperabel und besser zugänglich werden. Wir demonstrieren die praktische Anwendbarkeit des Formats anhand von OmegaWiki und präsentieren an diesem Beispiel die wichtigsten Eigenschaften von UBY-LMF, insbesondere die Repräsentation von Alignments. Wir stellen auch die finale, instantiierte Ressource UBY vor, ebenso wie die Java-basierte API, die programmatischen Zugang dazu ermöglicht, ein Web-Interface um die Inhalte von UBY im Browser zu untersuchen und das Alignment-Framework für unsere Experimente, dessen Implementierung durch die Standardisierung und die API erst ermöglicht wurde.

Um zu zeigen, dass WSA tatsächlich nützlich für NLP ist, stellen wir verschiedene Anwendungen vor, die darauf zurückgreifen. Das Clustering feingranularer GermaNet- und WordNet-Bedeutungen durch Ausnutzen von 1:n-Alignments zu OmegaWiki, Wiktionary und Wikipedia führt zu einem signifikanten Anstieg der Genauigkeit von Lesartendisambiguierung auf Standard-Evaluationsdatensätzen für Deutsch und Englisch, wobei dieser Ansatz sprachunabhängig ist und keinen speziellen Anpassungsaufwand für die jeweiligen Ressourcen erfordert. Das zweite Szenario ist computerunterstützte Übersetzung, und wir zeigen, dass mehrsprachige Ressourcen wie OmegaWiki und Wiktionary in diesem Fall nützliche Wissensquellen für zusätzliche Übersetzungen darstellen. In diesem Zusammenhang besprechen wir auch das Alignment zwischen beiden Ressourcen und geben Beispiele für das zusätzliche Wissen, welches durch die kombinierte Nutzung zugänglich wird.

Zuletzt beschreiben wir zahlreiche Ideen für weitere Arbeiten in der Zukunft, nicht nur in Bezug auf WSA, sondern auch für die Konstruktion von verlinkten Ressourcen wie UBY und die Anwendungen, die davon profitieren.



## Acknowledgments

Writing this dissertation was an effort to which many people contributed. I try to give credit where credit is due, but I apologize if anyone has been omitted here – rest assured that all contributions are warmly appreciated nevertheless.

That being said, I would first like to thank Prof. Dr. Iryna Gurevych for giving me the opportunity to join her group, despite my apparent lack of linguistic knowledge, and to conduct the research which I present in this thesis. Her excellent feedback and ongoing encouragement helped me become the computational linguist she saw in me before anyone else did. I would also like to thank Prof. Dr. Karsten Weihe and Prof. Roberto Navigli, not only for finding the time to review my thesis, but also for providing many inspirations and ideas which influenced my work in various ways. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant No. I/82806 and by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE)” as part of the research center “Digital Humanities”, for which I am also grateful.

My time at UKP was not always easy, but my colleagues made this chapter of my life more than worthwhile. First of all, I have to mention the co-authors of my publications, who I thank for numerous fruitful discussions and invaluable inspiration: Dr. Christian M. Meyer, Dr. Judith Eckle-Kohler, Silvana Hartmann, Tri Duc Nghiem, Christian Wirth, Tristan Miller, and the Open Linguistics Working Group. Other people who were in one way or the other involved in UBY, but who I did not have the honour to publish with, are Elisabeth Niemann, Yevgen Chebotar, Dr. Kostadin Cholakov, Zijad Maksuti and Than-Le Ha. Contributions to other aspects of my work (algorithmical, technical or otherwise) were made by Dr. Richard Eckart de Castilho, Dr. Wolfgang Stille, Dr. Daniel Bär and Prof. Dr. Chris Biemann – their help is greatly appreciated. On behalf of many others who helped, I thank Mladen Turković and Ilia Kuznetsov for extensive proofreading. For ongoing motivation, kind words, a helping hand now and then, sharing their wisdom, occasional mischief, and many other little things that made me smile, I especially have to thank Petra Stegmann, Nicolai Erbs, Emily Jamison, Dr. Niklas Jakob, Christof Müller, Benjamin Herbert, Dr. György Szarvas and Prof. Dr. Torsten Zesch.

Special mention deserves the small group of people who I joined for the occasional football match – apart from some colleagues already mentioned above and a some nice guys from other groups, I want to thank Johannes Daxenberger and Pedro Santos for this welcome opportunity to take my mind off work.

Two colleagues outside of UKP who I have to thank are Verena Henrich, for sharing her data and for some insightful discussions, and Prof. Dr. Stefan Conrad, who sparked my interest in science.

I would of course also like to thank my parents Christine and Waldemar Matuschek and my brother Dr. Dominik Matuschek, who always supported me and never doubted that I could achieve great things.

Above all others, I have to thank Charlotte. You were always by my side during these years, encouraging and supporting me in any imaginable way, although you went through difficult times yourself. I will be forever grateful for this, and I will always love you. My job is done, now it’s finally your turn.

In loving memory of my grandfather, Helmut Matuschek.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Overview . . . . .	1
1.2	Typographic and Terminological Conventions . . . . .	3
1.2.1	Terminology . . . . .	3
1.2.2	Typography . . . . .	5
1.3	Contributions . . . . .	5
1.4	Publication Record . . . . .	7
<b>2</b>	<b>WSA: Overview and Background</b>	<b>9</b>
2.1	Introduction . . . . .	9
2.2	Problem Definition: Word Sense Alignment . . . . .	10
2.2.1	Evaluation Metrics for WSA . . . . .	10
2.2.2	Baselines . . . . .	12
2.3	Similar Tasks from other Fields . . . . .	12
2.3.1	Ontology Matching . . . . .	12
2.3.2	Database Schema Matching . . . . .	14
2.3.3	Graph Matching . . . . .	14
2.4	Related NLP Tasks . . . . .	15
2.4.1	Word Sense Disambiguation . . . . .	15
2.4.2	Text Similarity . . . . .	16
2.4.3	Paraphrase Detection and Textual Entailment . . . . .	17
2.4.4	Semantic Relatedness . . . . .	18
2.5	Conclusions and Summary . . . . .	18
<b>3</b>	<b>Resources and Datasets</b>	<b>21</b>
3.1	Introduction . . . . .	21
3.2	Expert-built Resources . . . . .	23
3.3	Collaboratively Constructed Resources . . . . .	25
3.4	Analysis of LSRs . . . . .	34
3.4.1	Analysis of Glosses . . . . .	35
3.4.2	Analysis of the Graph Structure . . . . .	37
3.5	Selection of Evaluation Datasets . . . . .	40
3.5.1	Datasets Reported in Previous Work . . . . .	46
3.5.2	Datasets Created in Our Work . . . . .	48
3.6	Chapter Summary and Contributions . . . . .	50

<b>4</b>	<b>Similarity-based Word Sense Alignment</b>	<b>53</b>
4.1	Introduction . . . . .	53
4.2	Previous Work . . . . .	54
4.3	Wiktionary-OmegaWiki Alignment . . . . .	57
4.3.1	Alignment Procedure . . . . .	58
4.3.2	Evaluation . . . . .	58
4.3.3	Error Analysis . . . . .	60
4.4	WordNet-OmegaWiki Alignment . . . . .	60
4.4.1	Alignment Procedure . . . . .	61
4.4.2	Evaluation and Error Analysis . . . . .	62
4.5	Chapter Summary and Contributions . . . . .	64
<b>5</b>	<b>Graph-based Word Sense Alignment</b>	<b>67</b>
5.1	Introduction . . . . .	67
5.2	Previous Work . . . . .	68
5.3	Dijkstra-WSA . . . . .	70
5.3.1	Graph Construction . . . . .	71
5.3.2	Computing Sense Alignments . . . . .	75
5.3.3	Evaluation . . . . .	79
5.3.4	Error Analysis . . . . .	87
5.3.5	Issues with VerbNet Alignments . . . . .	89
5.4	Chapter Summary and Contributions . . . . .	89
<b>6</b>	<b>Joint Approaches to Word Sense Alignment</b>	<b>91</b>
6.1	Introduction . . . . .	91
6.2	Previous Work . . . . .	92
6.3	Joint Modeling of Features . . . . .	94
6.3.1	Feature Engineering . . . . .	94
6.3.2	Machine Learning Classifiers . . . . .	97
6.3.3	Experimental Results and Analysis . . . . .	98
6.4	Experiments on N-way Alignment . . . . .	107
6.4.1	Using Existing Alignments . . . . .	107
6.4.2	Considering Multiple Graphs at once . . . . .	107
6.4.3	Clustering Word Senses for a Lemma . . . . .	107
6.5	Chapter Summary and Contributions . . . . .	111
<b>7</b>	<b>UBY</b>	<b>113</b>
7.1	Introduction . . . . .	113
7.2	Related Work . . . . .	114
7.3	The Lexical Markup Framework and UBY-LMF . . . . .	119
7.4	UBY – The Final Resource . . . . .	124
7.5	Community Issues . . . . .	126
7.6	Chapter Summary and Contributions . . . . .	129



<b>8 Applications of Sense Alignments</b>	<b>131</b>
8.1 Using Alignments for Sense Clustering . . . . .	132
8.1.1 Introduction . . . . .	132
8.1.2 Related Work . . . . .	133
8.1.3 Task Definition . . . . .	134
8.1.4 Evaluation . . . . .	134
8.2 Computer-Aided Translation . . . . .	147
8.2.1 Introduction . . . . .	147
8.2.2 Motivation . . . . .	148
8.2.3 Related Work . . . . .	153
8.2.4 Discussion of Alignment Results . . . . .	154
8.3 Chapter Summary and Contributions . . . . .	156
<b>9 Conclusions</b>	<b>159</b>
9.1 Summary of the Thesis . . . . .	159
9.2 Outlook . . . . .	161
<b>A Implemented Software</b>	<b>185</b>
A.1 Java OmegaWiki Library (JOWKL) . . . . .	185
A.2 Access to UBY . . . . .	187
A.2.1 UBY-API . . . . .	187
A.2.2 UBY Web Interface . . . . .	189
A.2.3 Building a WSA Framework with UBY . . . . .	193
A.3 Chapter Summary and Contributions . . . . .	194



# List of Figures

1.1	Visual outline of the thesis . . . . .	4
2.1	Visual outline: WSA overview . . . . .	9
3.1	Visual outline: resources . . . . .	21
3.2	OmegaWiki’s defined meaning 5555 on <i>bass</i> . . . . .	27
3.3	Gloss overlap for WordNet, Wiktionary and OmegaWiki . . . . .	38
3.4	Overview of UBY . . . . .	43
4.1	Visual outline: similarity-based approaches . . . . .	53
5.1	Visual outline: graph-based approaches . . . . .	67
5.2	Monosemous linking example . . . . .	72
5.3	Dijkstra-WSA example . . . . .	77
6.1	Visual outline: joint approaches . . . . .	91
7.1	Visual outline: UBY . . . . .	113
7.2	Classes used for modeling OmegaWiki . . . . .	122
8.1	Visual outline: applications . . . . .	131
8.2	The translation alternatives for <i>bass</i> in Google Translate. . . . .	149
8.3	Wiktionary entry on <i>bass</i> . . . . .	150
8.4	OmegaWiki’s defined meaning 5555 on <i>bass</i> . . . . .	151
8.5	Illustration of the sense alignment between WKT and OW . . . . .	155
A.1	Establishing a connection to the OmegaWiki DB. . . . .	185
A.2	Accessing WordNet in the UBY-API. . . . .	188
A.3	Accessing knowledge from multiple LSRs in the UBY-API. . . . .	189
A.4	Web interface: Visual view . . . . .	190
A.5	Web interface: Standard search result . . . . .	191
A.6	Web interface: Sense details view . . . . .	191
A.7	Web interface: Sense comparison view . . . . .	192



# List of Tables

2.1	Tasks related to WSA . . . . .	19
3.1	OmegaWiki language editions . . . . .	29
3.2	Translations in OmegaWiki . . . . .	29
3.3	OmegaWiki relations . . . . .	30
3.4	OmegaWiki themes . . . . .	31
3.5	OmegaWiki annotations . . . . .	32
3.6	OmegaWiki parts of speech . . . . .	32
3.7	Growth of OmegaWiki . . . . .	33
3.8	OmegaWiki coverage for English . . . . .	33
3.9	OmegaWiki coverage for German . . . . .	33
3.10	OmegaWiki polysemy . . . . .	34
3.11	Gloss analysis statistics . . . . .	35
3.12	Lexical overlap of glosses (en) . . . . .	36
3.13	Lexical overlap of glosses (de) . . . . .	37
3.14	Graph statistics . . . . .	38
3.15	List of all alignments . . . . .	44
3.16	UBY alignment matrix . . . . .	45
3.17	Gold standard statistics . . . . .	46
3.18	Gold standard characteristics . . . . .	46
4.1	Previous work on similarity-based WSA . . . . .	55
4.2	Alignment results for the Wiktionary-OmegaWiki alignment. . . . .	59
4.3	Cross-lingual alignment results for WordNet-OmegaWiki German . . . . .	63
5.1	Previous work on WSA using the structure of LSRs. . . . .	70
5.2	Influence of the frequency limit $\phi$ . . . . .	73
5.3	Overview of isolated notes . . . . .	74
5.4	Pseudocode of the Dijkstra-WSA algorithm. . . . .	76
5.5	Influence of the allowed path length $\lambda$ . . . . .	78
5.6	Influence of allowing 1:1 or 1:n alignments . . . . .	79
5.7	Dijkstra-WSA results for WN-WKT and WN-WP . . . . .	80
5.8	Dijkstra-WSA results for GN-WKT and FN-WKT . . . . .	83
5.9	Dijkstra-WSA results for WN-OW and WKT-OW . . . . .	83
5.10	Dijkstra-WSA results for WKT-WP in German and English . . . . .	85
5.11	Dijkstra-WSA confusion matrices for datasets from previous work. . . . .	86
5.12	Dijkstra-WSA confusion matrices for newly created datasets. . . . .	87

6.1	Previous work on WSA using combined features. . . . .	92
6.2	Available features in LSRs . . . . .	95
6.3	Machine learning results for WN-WKT and WN-WP . . . . .	100
6.4	Machine learning results for GN-WKT and FN-WKT . . . . .	101
6.5	Machine learning results for WN-OW and WKT-OW . . . . .	101
6.6	Machine learning results for WKT-WP in English and German . . . .	103
6.7	Machine learning confusion matrices for datasets from previous work.	103
6.8	Machine learning confusion matrices for newly created datasets. . . .	104
6.9	Comparison to (Pilehvar and Navigli, 2014) . . . . .	106
7.1	The content of UBY . . . . .	126
7.2	List of all alignments . . . . .	127
7.3	UBY alignment matrix . . . . .	128
8.1	Statistics about WebCAGe . . . . .	135
8.2	German affected lexical items . . . . .	136
8.3	Coverage of lexical items (German) . . . . .	136
8.4	WSD results for German OmegaWiki . . . . .	137
8.5	WSD results for German Wiktionary . . . . .	138
8.6	WSD results for German Wikipedia . . . . .	140
8.7	German WSD results for combined resourecs . . . . .	140
8.8	Polysemy reduction statistics for GermaNet . . . . .	141
8.9	English affected lexical items . . . . .	141
8.10	Coverage of lexical items (English) . . . . .	141
8.11	WSD results for English OmegaWiki . . . . .	142
8.12	WSD results for English Wiktionary . . . . .	144
8.13	WSD results for English Wiktionary . . . . .	145
8.14	English WSD results for combined results . . . . .	146
8.15	Polysemy reduction statistics for WordNet . . . . .	146
8.16	Aggregate WSD accuracy for WordNet . . . . .	147
8.17	Comparison of different resource types . . . . .	154
8.18	Information gain through the alignment for one sense of <i>bass</i> . . . . .	156
8.19	Statistics about the combination of WKT and OW . . . . .	157
8.20	Alignment statistics for Wiktionary and OmegaWiki. . . . .	157
A.1	Some equivalent operations in the WordNet-API and the UBY-API. . .	188

# Chapter 1

## Introduction

### 1.1 Overview

*Lexical-semantic resources* (LSRs) are indispensable in many areas of Natural Language Processing (NLP). Plainly speaking, they encode the human knowledge about language in machine-readable form, and as such they are always needed as a reference when machines are asked to interpret natural language in accordance with the human perception. Examples for such tasks are word sense disambiguation (WSD) and information retrieval (IR). The aim of WSD is to discover the correct meaning of ambiguous words in context, and in order to formalize this discovery a so-called sense inventory is required. This is an LSR encoding the different meanings a word can express. In IR, the goal is to retrieve, given a user query formulating a specific information need, the documents from a collection which fulfill this need best. Here, knowledge is also necessary to correctly interpret short and often ambiguous queries, and to relate them to the set of documents.

Nowadays, LSRs exist in many variations. For instance, the META-SHARE repository<sup>1</sup> lists over 1,000 different lexical resources, and the LRE Map<sup>2</sup> contains more than 3,900 resources which have been proposed as a knowledge source for natural language processing systems.

A main distinction, which is also of utmost importance for this thesis, is between expert-built and collaboratively constructed resources. While the distinction is not always totally clear, the former are generally resources which are created by a limited set of expert editors or professionals using their personal introspection, corpus evidence or other means to encode the knowledge. Collaboratively constructed resources, on the other hand, are open for every volunteer to edit, with no or only few restrictions such as registration for a web site. Intuitively, the quality of the entries should be lower when laypeople are involved in the creation of a resource, but it has been shown that the collaborative process of correcting errors and extending articles (also known as the “wisdom of the crowds” (Surowiecki, 2005)) can lead to results of remarkable quality (Giles, 2005). The most prominent example is the Wikipedia, the largest encyclopedia and one of the largest knowledge sources known. Although originally not meant for that purpose, it has also become a major

---

<sup>1</sup><http://www.meta-share.eu>

<sup>2</sup><http://www.resourcebook.eu>

source of knowledge for all kinds of NLP applications (Medelyan et al., 2009).

Apart from the basic distinction according to the production process, LSRs exist in many varieties. Some are focusing mostly on encyclopedic knowledge (Wikipedia), others resemble language dictionaries (Wiktionary) or aim to describe the concepts used in human language and the relationships between them from a psycholinguistic (Princeton WordNet, (Fellbaum, 1998)) or a semantic (FrameNet, (Ruppenhofer et al., 2010)) perspective. Another important distinction is between monolingual resources, i.e. those covering only one language, and multilingual ones, which not only feature entries in different languages but usually also translations. However, despite the large number of different LSRs, the growing demand for large-scale resources in different languages is still not easily met. While the Princeton WordNet has emerged as a de facto standard for English NLP, for most languages corresponding resources are either considerably smaller or missing altogether. For instance, the *Open Multilingual Wordnet* project lists only 25 wordnets in languages other than English, and none of these match or surpass the Princeton WordNet’s size (Bond and Foster, 2013). Multilingual efforts such as Wiktionary or OmegaWiki provide a viable option for such cases and seem especially suitable for smaller languages due to their open construction paradigm and low entry requirements (Matuschek et al., 2013), but there are still considerable gaps in coverage which the corresponding language communities are struggling to fill. A closely related problem is that, even if comprehensive resources are available for a specific language, there usually does not exist a single resource which works best for all application scenarios or purposes, as different LSRs cover not only different words and senses, but sometimes even completely different information types. E.g., the knowledge about verb classes (i.e. groups of verbs which share certain properties) contained in VerbNet is not covered by WordNet, although it might be useful depending on the task.

These considerations have led to the insight that, to make the best possible use of the available knowledge, the orchestrated exploitation of different LSRs is necessary. This lets us not only extend the range of covered words and senses, but more importantly, gives us the opportunity to obtain a richer knowledge representation when a particular meaning of a word is covered in more than one resource. Examples where such a joint usage of LSRs proved beneficial include WSD using aligned WordNet and Wikipedia in BabelNet (Navigli and Ponzetto, 2012a), semantic role labeling using PropBank, VerbNet and FrameNet (Palmer, 2009) and the construction of a semantic parser using a combination of FrameNet, WordNet, and VerbNet (Shi and Mihalcea, 2005).

Cholakov et al. (2014b) address the special task of verb sense disambiguation. They use the large-scale resource UBY (Gurevych et al., 2012) which contains nine resources in two languages and which will be discussed in greater detail later on (Chapter 7).

However, while the notion of similar or even equivalent word senses in different resources is intuitively understandable and often (but now always) quite easily made by humans, it poses a complex challenge for automatic processing due to word ambiguities, different sense granularities and information types (Navigli, 2006). This task, known as *Word Sense Alignment* (WSA) is the main focus of our work, and we report various contributions to this area.

First, we provide a brief introduction to the terminological and typographic



conventions which are applied throughout this thesis in Section 1.2. Then, in Section 2.2, we give a more formal definition of the issue of WSA, which is perennial in this thesis. In Chapter 2, we also describe some related tasks in NLP and other fields and outline how WSA relates to them.

In preparation for the main part of the thesis, we describe and comparatively analyze a selection of resources from different angles with the intention of assessing their suitability for various WSA approaches. We especially focus on OmegaWiki, as this is a resource with several properties interesting for NLP applications which has not been comprehensively covered in the literature before. The results of these analyses are discussed in Chapter 3. Based on this, we present a selection of WSA datasets which are the foundation of our experimental work. This work is presented in Chapters 4, 5 and 6, which are dedicated to different approaches to WSA we investigated.

The greater context of this work lies of course not in the mere alignment of resources for its own sake, but in the potential it holds for NLP applications. Thus, it is necessary to make the resources and the alignments between them interoperable and accessible – this is the purpose of the aforementioned integrated resource UBY. Apart from the alignments, we make various contributions to its construction, from the basic theoretical concept to the final database instantiation. These contributions are discussed in Chapter 7. Again, we put a special focus on the integration of OmegaWiki into UBY.

Finally, in Chapter 8, we present some applications which actually benefit from the sense alignments, and also the standardization effort, in order to make the case that our work is indeed beneficial to NLP research and applications. In Chapter 9, we summarize our findings and contributions, and also point out directions for future work. A visual outline of the content of this thesis is given in Figure 1.1. We repeat this outline in the respective chapters to improve the orientation for the reader. Simply put, the first two chapters of the thesis represent the introductory part, the following three chapters deal with the computation of alignments (which is our main focus) and the last two chapters discuss where and how alignments are actually used.

## 1.2 Typographic and Terminological Conventions

In this section, we introduce and define some of the terminology used throughout this thesis, as well as a few typographic conventions, to avoid confusion and inconsistencies.

### 1.2.1 Terminology

- A *word* in a text is a sequence of letters or characters considered as a discrete entity, which, in itself, does not carry any meaning.
- A *lexeme* is a word in combination with a part of speech such as noun, verb or adjective.

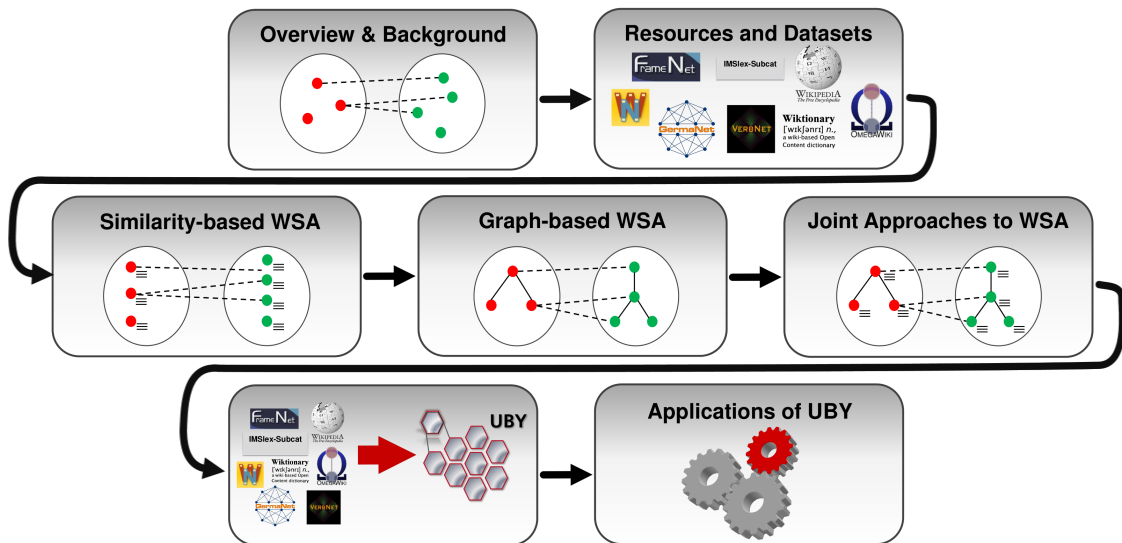


Figure 1.1: Visual outline of the thesis. The upper part represents the introduction, the middle part contains the main contributions, and the lower part presents further work which uses or is based on alignments.

- A *lexical item* is a fixed combination of words (i.e. a multiword expression) in combination with a part of speech; we usually use this term interchangeably with *lexeme*. For instance, the multiword noun *table cloth* is such a lexical item.
- A *sense* is one of the possible meanings or interpretations of a lexeme in a specific context. Note that the term *word sense* is commonly used, although senses are usually attached to lexemes. Lexemes can have more than one sense, and a sense is generally interpreted as representing a distinct concept of human perception.
- A *gloss* is a textual description of a sense’s meaning meant for human interpretation; it is also vital for many WSA and WSD approaches. In case of a missing gloss, senses can also be described by related words such as hypernyms and hyponyms (see below). This has been referred to as *artificial glosses* or *lexical fields* (Henrich et al., 2011).
- Semantic relations express a certain relationship between two senses. We list the most salient ones:
  - *Synonymy* connects senses which are lexically different but share the same meaning. Some resources such as WordNet subsume synonymous senses into *synsets*. However, for the sake of brevity we usually not distinguish between *sense* and *synset* as for most discussions and experiments we present they can be used interchangeably. Synonymy is reflexive, symmetrical and transitive.
  - *Antonymy* is a relation in which the source and target sense have opposite meanings (e.g. *tall* and *small*).

- *Hyponymy* denotes a semantic relation where the target sense has a more specific meaning than the source sense (e.g. from *limb* to *arm*).
  - *Hypernymy* is the inverse relation of hyponymy and thus denotes a semantic relation in which the target sense has a more general meaning than the source sense.
- A *Lexical-Semantic Resource*, *LSR* or simply *resource* generally consists of a description of lexemes and their possible senses (often, but not necessarily, by glosses), in a format which can be processed and accessed by machines. Additional information such as relations between senses is frequently given, but not strictly required. Different LSRs vary greatly with regard to their content (see Chapter 3). Other terms which are used in different contexts are *dictionary* and *lexicon*.

As an example, the noun *car* has (among others) two senses encoded in the LSR WordNet, with the glosses “a motor vehicle with four wheels” and “a wheeled vehicle adapted to the rails of railroad”, as well as the synonym *automobile* for the first sense. The central term *alignment* will be defined in the Section 2.2.

## 1.2.2 Typography

- Newly introduced terms and example lemmas are typed in *italics*
- Synsets are enclosed by curly brackets, e.g. {*car*, *automobile*}
- Concepts are typed in small caps, e.g. STREET VEHICLE WITH FOUR WHEELS
- Relations between senses are written as pairs in parentheses, e.g. (*car*, *vehicle*)
- Classes of the *Lexical Markup Framework* (LMF) standard are printed in a monospace font starting with an upper case letter (e.g., `LexicalEntry`).
- LMF data categories are printed in a monospace font starting with a lower case letter (e.g., `partOfSpeech`).

## 1.3 Contributions

We now give an overview of the main contributions of this thesis:

- In the previous work, there exists no comprehensive description of the multilingual, collaboratively constructed LSR OmegaWiki. We fill this gap by outlining its structure and content, putting it in relation to other LSRs (especially Wiktionary, which has been built according to a comparable paradigm) and thus motivate its usefulness for NLP applications (Section 3.3). In the context of UBY, we present how OmegaWiki can be modeled in terms of the Lexical Markup Framework, or more precisely, in UBY-LMF. We elaborate on the necessary steps to bring the OmegaWiki data into this unified format to achieve interoperability with other LSRs, and we compare the mapping process to the other resources which are contained in UBY (Section 7.3). For all

resources contained in UBY, we also perform an analysis of their suitability for different WSA approaches by comparatively examining their glosses as well as their inherent structures with regard to various parameters. To our knowledge, this is the first time an analysis is attempted from such an angle, and this analysis motivates the selection of resource pairs for our WSA experiments.

- For the first time, we present a full alignment between OmegaWiki and Wiktionary based on the similarity of glosses. To this end, we also present a manually annotated gold standard dataset covering Wiktionary and the English part of OmegaWiki. To our knowledge, this is the first time two collaboratively constructed LSRs have been used for such a task (Sections 3.5.2 and 4.3). We also present an algorithm based on gloss similarity which covers the cross-lingual case by introducing machine translation as an intermediate component. To demonstrate the validity of this approach, we align the German part of OmegaWiki to WordNet, and due to the multilingual nature of OmegaWiki (see Section 3.3), the dataset we created for this purpose can be used for a monolingual alignment as well. We also calculate an alignment for this case and provide a comparison between the mono- and cross-lingual scenarios (Sections 3.5.2 and 4.4).
- Going beyond similarity-based WSA, we present Dijkstra-WSA, a graph-based algorithm which exploits the structure of LSRs, i.e. the graphs induced by the relationships between senses by means of semantic relations or hyperlinks. This approach complements the previous approaches by covering a different aspect of sense similarity, and we show that our approach achieves a significant improvement in alignment precision on a variety of datasets, covering a wide range of resources with different properties. We also present two novel alignment datasets between Wiktionary and Wikipedia in English and German, where the latter is not only directly derived from Wiktionary (and hence is the first “crowd-sourced” WSA dataset), but also of unprecedented size (Sections 3.5.2 and 5.3). Finally, we also combine similarity-based WSA and structure-based WSA in two ways: First, by using a two-step backoff approach which first finds alignments based on the graph structures and falls back to gloss similarity if no alignment can be found for a sense and second, by jointly modeling similarity and structural features in a machine learning approach. We show that either approach outperforms the isolated usage of similarity and distance features, while the machine learning approach yields the best overall results and, to our knowledge, represents the current state of the art in WSA (Sections 5.3.2 and 6.3).
- For the construction of the unified resource UBY, which was a joint work with multiple colleagues and the bigger context for our WSA work, we present its underlying representational model UBY-LMF in detail and showcase how it is used to uniformly represent heterogeneous resources and the alignments between them (Chapter 7). We not only integrate our alignments into UBY, but using our Dijkstra-WSA algorithm as a foundation, we also present a generic approach for clustering fine-grained sense inventories to allow a more reliable sense distinction in applications. Our approach exploits the fact that

different sense granularities in LSRs lead to 1:n alignments (i.e. one sense is aligned to several ones in another resource), which can in turn be used to identify clusters of similar senses. We show that this approach yields significant improvement in WSD performance on GermaNet and WordNet when evaluated on standard datasets, while at the same time being far less complex and resource-specific than previous approaches (Section 8.1). As another application, we discuss how a combination of Wiktionary and OmegaWiki could be used in a computer-assisted translation environment – in this case, especially the multilingual properties of these collaboratively constructed resources can be exploited to allow faster and better creation of translations (Section 8.2).

- In the appendix, we summarize some noteworthy contributions from a software development perspective. First of all, we describe JOWKL, the Java OmegaWiki Library. Thus far, OmegaWiki had only been accessible via the OmegaWiki web site, or by exploring the raw SQL database it is based on. We present a Java-based API which makes all content within OmegaWiki easily accessible within applications, and which also forms the foundation of its integration into the unified resource UBY (Appendix A.1). For UBY itself, we present the Java-based API as well as the web interface which enable easy usage of the resource for application developers and researchers (Appendix A.2). The UBY-API was also used as a foundation to create the generic WSA framework employed in the experiments performed in the course of this thesis, and which is scheduled for a public release (Appendix A.2.3).

## 1.4 Publication Record

Large parts of this thesis' content have been previously published in peer-reviewed journals or conference proceedings. We list these below, and also indicate the respective sections which build upon them. The first batch of publications is concerned with algorithmic approaches to WSA and applications of sense alignments in NLP tasks; these constitute the main contributions of this thesis.

- A Language-independent Sense Clustering Approach for Enhanced WSD (with Tristan Miller and Iryna Gurevych). In: Proceedings of the 12th “Konferenz zur Verarbeitung natürlicher Sprache” (KONVENS 2014), p. 11–21, October 2014 (Section 8.1).
- High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity (with Iryna Gurevych). In: Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014), p. 245–256, August 2014 (Sections 3.5.2, 6.3).
- Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications (with Christian M. Meyer and Iryna Gurevych). In: Translation: Corpora, Computation, Cognition (TC3), vol. 3, no. 1, p. 87–118, July 2013 (Sections 3.5.2, 3.3, 4.3, 7.3).

- Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment (with Iryna Gurevych). In: Transactions of the Association for Computational Linguistics (TACL), vol. 1, p. 151–164, May 2013 (Section 5.3).
- Where the Journey is Headed: Collaboratively Constructed Multilingual Wiki-based Resources (with Iryna Gurevych). In: SFB 538: Mehrsprachigkeit: Hamburger Arbeiten zur Mehrsprachigkeit, 2011 (Section 3.3).
- Beyond the Synset: Synonyms in Collaboratively Constructed Semantic Resources (with Iryna Gurevych). In: Antti Arppe: Workshop on Computational Approaches to Synonymy at the Symposium on Re-Thinking Synonymy, p. 58–59, October 2010 (Section 3.3).

The second batch of publications deals with the construction of the unified LSR UBY. While the author has contributed to this on several levels, it was mostly a team effort with the co-authors of the respective papers. The main contributions concerning UBY are concentrated in Chapter 7, while several minor aspects are discussed in other chapters when appropriate.

- UBY-LMF – Exploring the Boundaries of Language-Independent Lexicon Models (with Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann and Christian M. Meyer). In: Gil Francopoulo: LMF Lexical Markup Framework, chap. 10, p. 145–156, ISTE - HERMES - Wiley, 2013 (Chapter 7).
- Navigating Sense-Aligned Lexical-Semantic Resources: The Web Interface to UBY (with Iryna Gurevych, Tri Duc Nghiem, Judith Eckle-Kohler, Silvana Hartmann and Christian M. Meyer). In: Proceedings of the 11th “Konferenz zur Verarbeitung natürlicher Sprache” (KONVENS 2012), p. 194–198, September 2012 (Appendix A.2.2).
- UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF (with Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann and Christian M. Meyer). In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), p. 275–282, May 2012 (Chapter 7).
- The Open Linguistics Working Group (with Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann and Christian M. Meyer). In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), p. 3603–3610, May 2012 (Chapter 7).
- UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF (with Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Christian M. Meyer and Christian Wirth). In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), p. 580–590, April 2012 (Sections 3.5.2, 4.4, A.2, Chapter 7).

# Chapter 2

## Word Sense Alignment: Overview and Background

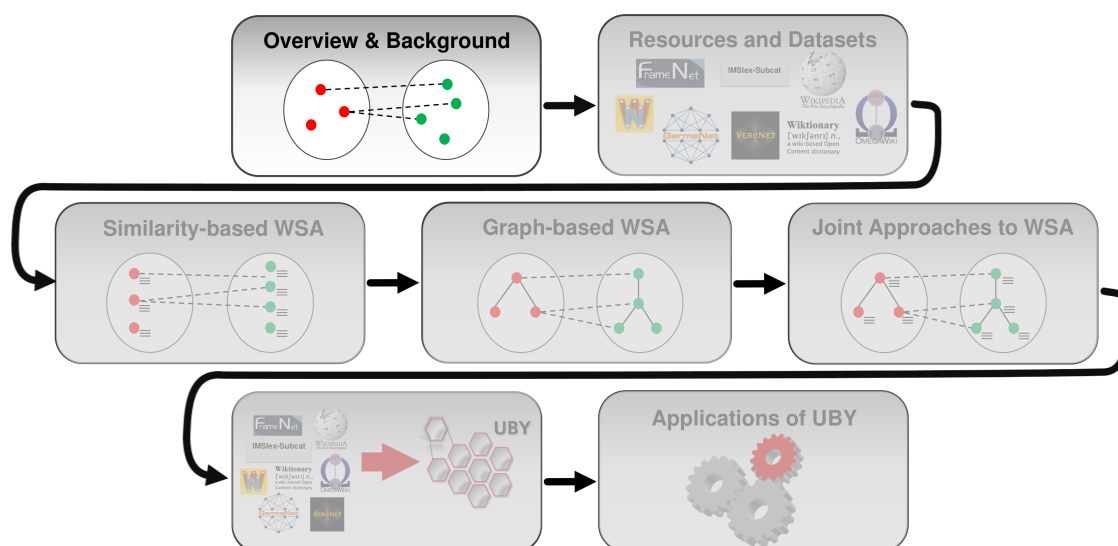


Figure 2.1: Visual outline of the thesis.

### 2.1 Introduction

As we already pointed out in the introductory section, this thesis is mainly concerned with the alignment of senses (or concepts) from different lexical-semantic resources. However, to fully understand the scope of this task and its relationship to other problems, we first want to take a step back and examine the big picture that our work is embedded in. This will also help in making it more clear what our contributions to this field are and in what ways our algorithmic approaches can be discriminated from related efforts.

To this end, we will first provide an exact problem definition, reflecting the way we interpret the challenge of WSA (Section 2.2), and also give an overview of evaluation metrics and common baselines for this task. After this, we move on

to describing tasks from other fields of research which are, in some respects, quite similar to WSA, and we point out differences and inspirations for our own work where applicable (Section 2.3). Moreover, we will discuss other tasks from NLP, and especially from semantic processing, which are related to WSA and sketch the ways in which these different issues are intertwined with our own work (Section 2.4). At the end of the chapter, we will provide a summary of our reflections on WSA.

Note that in this chapter we will not consider the previous efforts in WSA as an understanding of the resources participating in WSA is indispensable for that. We will discuss the corresponding previous work after the resources have been presented in detail in Chapter 3.

## 2.2 Problem Definition: Word Sense Alignment

We define a *Word Sense Alignment* (WSA), or *alignment*<sup>1</sup> for short, as a list of pairs of senses (or, more generally, concepts) from two LSRs, where the members of each pair represent an equivalent meaning. As an example, the two senses of the noun *letter* “The conventional characters of the alphabet used to represent speech” and “A symbol in an alphabet, bookstave” (taken from WordNet and Wiktionary, respectively) are clearly equivalent and should thus be aligned. *Alignment candidates*, or simply *candidates* for a particular sense  $s$  in one LSR  $A$  are all those senses  $t_1, \dots, t_n$  in another LSR  $B$  which are attached to the same lexeme, i.e. all senses which could potentially participate in a pair with  $s$ . For instance, for the “programming” sense of *Java* in one resource, their might exist senses for “programming”, “island” or “coffee” in the other one which are all possible alignment targets.

Creating an alignment is then, essentially, the task of deciding which pairs of senses and candidates would constitute a valid equivalence relation. While such a decision can be made by human annotators (which is the usual way the WSA evaluation datasets we report in Section 3.5 are created), we are interested in the process of *automatically* creating an alignment of two LSRs. For this task, we also use the term Word Sense Alignment (or WSA), and as such it is ambiguous, describing both the process and its result. In this work, it should usually be clear from the context which meaning we refer to, but we will explicitly state it in case we see potential for misinterpretation.

Note that our definition is not necessarily restricted to 1:1 alignments, i.e. a sense may participate in more than one pair, so it is possible that  $s$  is assigned to several of the candidate senses  $t_1, \dots, t_n$ , in case  $B$  has more subtle or fine-grained sense distinctions. In some configurations, however, it proves helpful to restrict ourselves to 1:1 alignments. We will explicitly point out occasions where this is done.

### 2.2.1 Evaluation Metrics for WSA

The performance of an alignment algorithm is usually assessed with a variety of different metrics, measured against gold standard datasets which were created by

---

<sup>1</sup>Note that in related work the terms *sense mapping* and *sense matching* are also used. Sense alignment should, however, not be confused with *word alignment*, which takes place at the lexical level and is a preprocessing step in machine translation.



human annotators and are known to be correct. To calculate them, it is necessary to count the number of all possible decisions made: i) true positives (TP), i.e. correct detection of positive examples, ii) true negatives (TN), i.e. correct detection of negative examples (non-alignments), iii) false positives (FP), i.e. examples which are aligned but should not be and iv) false negatives (FN), i.e. examples which are not aligned but should be.

**Precision** reports how many of our decisions to align two senses are correct, i.e. the higher the precision of our algorithm the more confident we can be that the senses we align are equivalent. It is formally defined as:

$$P = \frac{TP}{TP+FP}$$

**Recall** reports how many of the positive examples in the gold standard are found by our algorithm, i.e. the higher the recall of our algorithm the more confident we can be that we detect all valid alignments between senses. It is formally defined as:

$$R = \frac{TP}{TP+FN}$$

**F-measure** (or F-score) is the harmonic mean of precision and recall. It is usually considered as the crucial alignment metric, as neither precision nor recall are useful in isolation. They are also antagonistic: perfect precision can be achieved by not aligning at all (no incorrect decision is made), while perfect recall is achieved by aligning everything (no alignment is missed). F-measure is defined as:

$$F_1 = \frac{2 \cdot P \cdot R}{P+R}$$

**Accuracy** reports how many of the decisions made by the algorithm are correct in total, i.e. considering both positive and negative examples. While this is also an indicator of alignment quality, it should be carefully judged depending on the dataset. If the data is heavily skewed, good accuracy can easily be achieved by always assigning the majority class, e.g. if 90% of the gold standard examples are non-alignments, a baseline aligning nothing would reach an accuracy of 0.90. Thus, F-measure is usually the more meaningful number. Accuracy is defined as:

$$A = \frac{TP+TN}{TP+TN + FP+FN}$$

**Annotator agreement** is not a quality measure of the automatic alignment result, but of the gold standard that is used for evaluation. Specifically, we report the F-measure between annotators (as defined by Hripcsak and Rothschild (2005)) as well as the *observed annotator agreement*, which is defined as the sum of all cases in which the annotators agree. Formally, for items  $i \in I$  which annotators have to judge,

$$agr_i = \begin{cases} 1 & \text{if the annotators assign } i \text{ to the same category} \\ 0 & \text{if the annotators assign } i \text{ to different categories} \end{cases}$$

and

$$A_0 = \frac{1}{i} \sum_{i \in I} agr_i$$

Intuitively, as  $A_0$  also makes a statement about all alignment decisions and judges the reliability of human performance on this task, it can be considered a plausible upper bound for the accuracy  $A$ .

### 2.2.2 Baselines

As a point of reference for arbitrary alignment setups, it is common to report the aforementioned evaluation metrics for a set of naive baselines, which are trivial to compute and should thus be outperformed by any given algorithm. The three baselines we report for WSA throughout this thesis are:

- *Random*: A random sense from the set of candidates is chosen in each case.
- *1:1*: An alignment is always made if and only if there is exactly one candidate. This baseline is expected to be strong when the degree of polysemy (i.e. the average number of senses per lexeme) is rather low.
- *1st*: The first of the candidate senses is always selected. While this corresponds to the most frequent sense baseline in some cases, note that no explicit frequency information is available for the resources other than WordNet so that the first sense baseline is only a rough approximation.

Other baselines will be reported as well where appropriate. For instance, for graph-based and combined WSA approaches (reported in Chapters 5 and 6, respectively) we will also repeat the results for the similarity-based setup as reference.

## 2.3 Similar Tasks from other Fields

There are several tasks from related fields which are comparable to WSA in the sense that an algorithmic matching of particular entities (possibly carrying a certain meaning) is performed. The circumstances such as the available information differ significantly, however, posing unique challenges for each task.

### 2.3.1 Ontology Matching

An ontology is formally defined as a specification of a conceptualization (Euzenat and Shvaiko, 2013). Plainly said, it provides the vocabulary for describing a domain of interest, and specifies the meaning of the terms used in this vocabulary. Usually, well defined relations between concepts such as *subclass* exist which provide structure to the ontology and comprehensively define the properties of its instantiations. For example, *car* is a subclass of *vehicle*, so if a vehicle can be used for transportation a car can trivially be used for this purpose as well; such reasoning over ontologies is usually highly desired. Many LSRs can also be considered ontologies as they contain corresponding relations between concepts (Veale et al., 2004). However,

LSRs usually are not limited to a particular domain as they most often aim to encompass the entirety of real-world concepts and perceptions which are expressible and conceivable via written language. Examples for such “language ontologies” are OmegaWiki (Section 3.3), OntoWiktionary (Meyer, 2013) or the resources contained in the *Linguistic Linked Open Data Cloud*.<sup>2</sup>

The matching of ontologies is important in case different conceptualizations need to be used in conjunction or merged. For instance if a company is acquired by another one and the internally used ontologies for the goods they produce must be harmonized. Hence, many different approaches have been developed, which are ideologically comparable to the ones suggested for WSA (which we will present in detail later on) and can for most part be sorted into one of two broad categories: i) terminological approaches are based on the lexical comparison of ontological entities and their descriptions (Cohen et al., 2003; Yatskevich and Giunchiglia, 2004), ii) structural approaches exploit the relationships between entities and basically try to find well-matching substructures in both ontologies (Maedche and Staab, 2002; Giunchiglia et al., 2004). Hybrid approaches combining both directions seem to show the best results (Le et al., 2004).

The so-called extensional approaches present a major difference with regard to WSA (Dhamankar et al., 2004; Doan et al., 2003). These compare the actual instantiations of an ontology to identify entities which correspond. For example, it would be possible to look up the instances which are categorized as CAR by one ontology and see if their attributes (like number of seats, size, price etc.) match the instances in the other ontologies. This is useful in case of ontologies with very different descriptions or structures. Such an approach is not possible for WSA, because there is no way to determine which real-world entities are covered by a specific word sense – e.g., it is not possible to look at all existing (or even hypothetical) horses and see if they are covered by a sense definition in WordNet.

Closely related to this, another fundamental distinction to WSA is that in an ontology different entities usually have very different attributes (such as the aforementioned attributes for a car). This is indispensable for ontologies as they are tailored for specific domains and need to reflect, in a well-structured manner, the discriminating properties of heterogeneous objects. As such, the consideration of the number and specification of attributes is as important for ontology matching as the examination of their content. For LSRs, on the other hand, the description of the concepts is usually homogeneous in the sense that the same set of descriptive features (such as gloss, example sentences etc.) is used for each concept (cf. Section 4.2). These features need thus be general enough to be applicable to all conceivable kinds of concepts, while this is not a requirement for ontologies.

Lastly, the well-defined semantics of relations, which are often exploited by ontology matching algorithms, are not always given in WSA. While there are approaches which exploit particular relations in WordNet (cf. Section 5.2), these ideas are not applicable to all LSRs. In Wikipedia, for instance, links between articles usually represent only a general notion of relatedness without specifying its exact nature, and for FrameNet, the participation of senses in the same frame is a sign of relatedness which is hard to more specifically reason about. Moreover, many resources, espe-

---

<sup>2</sup><http://linguistics.okfn.org/resources/llod/>

cially collaboratively constructed ones, suffer from disconnected or sparse graphs (cf. Section 3.4.2) which renders the sole usage of structural approaches ineffective.

In summary, WSA is a harder task in comparison to ontology matching, as the participating resources are usually only lightly structured, not as strictly specified semantically, and instantiations of concepts can usually not be obtained for examination. Thus, WSA algorithms (at least those which aim to be universally applicable) can only rely on the few generally available semantic information types such as glosses and example sentences, and the exploitation of structural information in terms of paths and distances is only possible if no overly strict assumptions about the semantics of the relations are made.

### 2.3.2 Database Schema Matching

Database schema matching is, in many respects, comparable to ontology matching as the participants in an alignment and the relations between them are strictly and soundly defined from the formal point of view, for instance, via foreign key relations which connect certain database tables; thus, many approaches from ontology matching are also applicable in this case (Berlin and Motro, 2002). The fundamental difference, however, is that in many cases a semantic interpretation of the database content is not made explicit. While database schemata usually also model real-world concepts and relations, there is often no other access to the interpretation of the information than the tables' and attributes' names, and even this possibility is sometimes tedious due to cryptic, "non-speaking" denominations. These are, for instance, chosen for reasons of storage efficiency. Moreover, database relations expressed via keys also express no more than a "relationship" between two tables in a generic, technical sense of the term. Their actual semantic interpretation is usually even harder than for table attributes as most database system implementations such as SQL do not allow explicitly labelling such relations.

Thus, even more than for ontology matching, algorithmic approaches have to rely on technical specifications such as number and data types of the attributes and instantiations of entities, i.e. content-based matching (Kang and Naughton, 2003). Graph-based approaches are further impaired by the fact that different database design paradigms allow expressing the exact same information with a different segmentation and allocation of data across tables. Thus, in summary, this task is even further removed from WSA as semantic interpretation of the data for alignment purposes is of minor importance and usually eclipsed by metadata- or instantiation-based algorithms.

### 2.3.3 Graph Matching

Another closely related problem, which by definition solely relies on structural properties, is graph matching, or more precisely, the computation of graph isomorphisms. Here, the task is to calculate pairs of nodes from two distinct graphs which have the same position in the respective graph topologies. The problem here is that, akin to database schema matching, usually no additional information is given which allows the semantic interpretation of the data. In general, the only information available is whether two nodes in a graph are linked or not. Hence, without further

constraints, an effort exponentially increasing with the number of nodes involved is required, which means that the task is NP-hard (Arvind et al., 2012), but probably not NP-complete (Schöning, 1988).

For the transfer of graph isomorphism algorithms to WSA, we could rely on additional constraints such as limiting the set of candidates which are applicable for a particular node (cf. Section 2.2). We would, however, still be presented with the problem that LSR topologies are very different, as the interpretation and manifestation of edges between nodes (for instance, semantic relations or mere links) varies greatly and exact matches of subgraphs are thus only likely for very small groups of nodes. Thus, for alignments of sufficient coverage and precision, less restrictive, distance-based matching seems necessary, in combination with complementary gloss-based approaches which provide the necessary background knowledge.

## 2.4 Related NLP Tasks

Apart from the tasks in related fields which are comparable to WSA, there are also quite a few challenges in NLP which are directly related to WSA. Either because they have a similar, but slightly different definition or scope, or because they are a sub-task of WSA which needs to be solved to compute an alignment. We will discuss these tasks in this section.

### 2.4.1 Word Sense Disambiguation

Word sense disambiguation (WSD), as already briefly stated in the introduction, is the task of assigning the correct sense of a lexeme in the context of a document (see the seminal work by Navigli (2009b) for a comprehensive overview). As such, WSA can be considered as a special case of WSD, as we also strive to assign a meaning to a lexeme relative to a “target” sense inventory – however, the context of the lexeme is not given by the document that contains it, but rather by the sense in the “source” sense inventory it is attached to, more specifically, the description of the sense and the other senses in the vicinity which are related to it.

Due to this similarity of the task definitions, many methods used in WSD can be adapted or straightforwardly used for WSA. WSD using the overlap of a context of a lexeme (i.e. the “window” around it) with a description of a sense in an LSR was first introduced by Lesk (1986) and then refined and extended in many efforts afterwards (e.g. (Banerjee and Pedersen, 2002)). The WSA approach we introduce in Chapter 4 which calculates the similarity between sense descriptions in two different LSRs is directly based on this idea. Graph-based approaches exploiting the structure of the target sense inventory have also been widely adopted. Two of the most prominent examples are the SSI algorithm (Navigli and Velardi, 2005) and the SSI-Dijkstra+ algorithm (Laparra et al., 2010) which are based on finding appropriate paths for polysemous lexemes to WordNet synsets, starting from unambiguous (i.e monosemous) words in a text which can be trivially disambiguated. These approaches inspired the Dijkstra-WSA algorithm we present in Chapter 5. As we already pointed out in the introduction, linked resources are also successfully used for knowledge-based WSD. For instance, Navigli and Ponzetto (2012d) use *Ba-*

*belNet* to combine evidence from different resources and languages for a more precise disambiguation.

However, despite their close relatedness, there are a few notable differences between the two tasks. First of all, one of the commonly accepted assumptions in WSD is that, within a document or even document collection, several occurrences of the same lexeme can safely be assigned the same sense as humans typically avoid to use the same term with different meanings in the same context to avoid misunderstandings (“one sense per discourse”, cf. Navigli (2009b)). This assumption is, however, useless in WSA as per definition there are no two senses for the same lexeme in an LSR which express the same meaning. Closely related is the notion of *all-words WSD*. This is a setup for WSD which aims to disambiguate all lexemes in a document at once, as opposed to *lexical sample* or *targeted WSD*, which only requires, for instance, a single word within a sentence to be disambiguated. This “global” solution to a WSD task impairs results for supervised systems as training samples for every non-stopword in a document would be required (Màrquez et al., 2006). However, knowledge-based systems (i.e. ones relying on knowledge from LSRs) can potentially benefit from the intuition that senses of lexemes which occur in the same document are likely to be related, so that finding disambiguations for a subset of target lexemes with high confidence can considerably facilitate disambiguating the remaining ones (Agirre et al., 2009). Nevertheless, this assumption of “discourse coherence” within a single document is not directly applicable to WSA. While the Dijkstra-WSA algorithm presented in Chapter 5 relies on the intuition that linked senses are semantically related and thus, in a certain way, form a coherent sub-graph within an LSR, the boundaries are not as clear-cut as with documents here. Single links to only distantly related senses (as it is, for instance, common in Wikipedia) can impair the results substantially – we will discuss this in more detail in Section 5.3.3. Thus, a global solution to WSA, i.e. one which is plausible when considering all senses within an LSR, seems not easily achievable as per definition a general purpose language resource contains lexemes and concepts from very different, and for the most part unrelated topics.

## 2.4.2 Text Similarity

Text similarity is intimately related to both WSD and WSA, as for many approaches the similarity between sense descriptions and/or context forms the foundation for making the disambiguation or alignment decision (see Section 4.2). As such, it can be considered a sub-task which needs to be solved. However, it is vital to understand that text similarity, as the name says, operates on texts and not concepts as WSA. The semantics of a text is, of course, the key for determining the similarity of texts regarding their content (which is usually the most interesting aspect for WSA), but for this and other dimensions (such as style) also surface level features such as n-gram overlap have proven useful (Bär, 2013).

If the “texts” considered consist only of a single term, the challenge faced in text similarity is closest to WSA as in this case an assessment of the term is not possible by means of the context surrounding it, but only implicitly by considering it in combination with the term it is compared to. Consider, for instance the pairs *jaguar, porsche* and *jaguar, tiger*: For each pair, an implicit assumption about the

sense of *jaguar* is made (i.e. a certain sense is activated by the term it is compared to, cf. Cruse (1986)) and the task of comparing the texts amounts to comparing the single concepts which they represent. For longer texts, it is common to either calculate composite measures based on a combination of word similarities or non-compositional measures which aim to capture the semantics of a document in a global fashion, e.g. Latent Semantic Analysis (Deerwester et al., 1990).

A fundamental difference to WSA, however, is that text similarity is usually given in grades, not binary – a simple notion of “similar” or “not similar” would be too coarse-grained for most applications such as text reuse detection (Bär et al., 2012). Thus, to make this feature applicable to WSA, a threshold needs to be set or learned to discretize the decision (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011), unless text similarity is used in combination with other features; we will discuss this later on in Chapter 6. Combination with other features is also recommendable because text similarity, although without doubt being a useful feature for WSA if sense descriptions are available, is not a magic bullet – equivalent senses can have very different (or even non-overlapping) descriptions. This motivates our work on graph-based and combined approaches in Chapters 5 and 6.

### 2.4.3 Paraphrase Detection and Textual Entailment

Textual entailment concerns the logical deduction of truth values about statements (Androutsopoulos and Malakasiotis, 2010) – for instance, the statement “all men must die” entails that the author will one day die as well. Accordingly, two statements are paraphrases if they entail each other, like “all men must die” and “no man lives forever”. As this is strictly defined as a binary decision, it is in this respect more similar to WSA than text similarity calculation. Also, if two sense descriptions are paraphrases (i.e. they have the exact same meaning) this automatically entails that the concept they describe must be equivalent.

As such, paraphrase detection could intuitively be a useful feature for WSA, but closer observation reveals that the definition of paraphrase detection is actually too strict for our purposes. First of all, paraphrase detection is concerned with exactly discernible truth values of statements and thus needs to take care of things like negation (Zaenen et al., 2005), which is crucial for identifying paraphrases in general texts, but presumably not for WSA. Negative statements (“A bird is *not* a ...”) are, from our experience, rarely observed in sense descriptions found in LSRs<sup>3</sup>. Moreover, equivalence of concepts does not entail that their descriptions are paraphrases. Often, the truth value of glosses is not discernible as they are not full-fledged sentences, but even if this is solved by reformulating the glosses (e.g. attaching the prefix “A house is ...” to the corresponding gloss “a dwelling that serves as living quarters for one or more families” in WordNet), the definition of paraphrasing might still be too strict. In this case, the Wiktionary sense with the (prefixed) gloss “A house is a structure serving as an abode of human beings” would intuitively be chosen as the correct alignment target, but the two glosses would still not qualify as paraphrases since human beings are not necessarily members of a family.

---

<sup>3</sup>Note that negation does not usually influence common text similarity measures as, for instance, “not” would be filtered out as a stopword

In summary, while paraphrase detection seems promising as a sub-task of WSA at a first glance, it is too strictly defined to be generally useful. For the sense descriptions usually found in LSRs, textual similarity is preferable as it allows for a more fuzzy notion of similarity which better reflects the human perception of sense equivalence.

#### 2.4.4 Semantic Relatedness

Unlike text similarity and paraphrase detection, semantic relatedness is most commonly considered for single words, and thus for concepts implied in context (as in the *jaguar* example mentioned above). In this respect, it is similar to WSA, and trivially, if two concepts are equivalent or aligned, they are also related. Also, the methodologies used for calculation of semantic relatedness are quite similar to WSA, ranging from gloss-based methods (Banerjee and Pedersen, 2002) over path-based methods on the graph representation of an LSR (Rada et al., 1989) to hybrid (Patwardhan and Pedersen, 2006) and multilingual approaches such as *BabelRelate* (Navigli and Ponzetto, 2012c). Semantic relatedness is, like text similarity, an important sub-task for tasks like information retrieval or determining textual coherence.

There are however, a few substantial differences: First of all, semantic relatedness is usually calculated within a single LSR. The intuition behind this is that semantic relatedness tries to express to what extent two real world concepts encoded in a reference sense inventory are related, so that the question if concepts from different sense inventories are related is usually irrelevant. This is also in line with the fact that semantic relatedness, like text similarity, is given in grades. A binary decision would not make sense in the context of a single LSR since the fact that two concepts are related at all is usually already expressed via semantic relations or links. Another very important difference is that semantic relatedness is much more loosely defined than the other related tasks we discuss in this section. It is not only applicable across different parts of speech (for instance, *car* and *drive* can be considered closely related), it also covers cases like antonymy: *love* and *hate* would be considered neither equivalent nor similar, they are however related as they are often used in the same context and also frequently linked in LSRs accordingly.

## 2.5 Conclusions and Summary

In this chapter, we discussed the task of word sense alignment, by first giving a precise definition of the problem at hand and explaining common evaluation measures and baselines which are applied for this task. After that, we located WSA in the broad field of semantic analysis by comparing it to similar tasks from other areas of computer and information science, and related tasks from NLP. We have seen that matching problems are perennial and concern, for instance, ontologies, databases and graphs, but that WSA faces unique challenges due to the properties of LSRs, and especially the descriptions and relations of the concepts they encompass. We also explained that WSA is most closely related to WSD (and can be considered a special case of it), but still has some important peculiarities which make the development of specialized algorithms necessary. Finally, we have shown that related



Task	Works on	Result	Information Sources			
			Desc.	Struct.	Meta	Instances
Word Sense Alignment	Concepts	Binary	✓	✓	✗	✗
Ontology Matching	Concepts	Binary	✓	✓	✓	✓
Schema Matching	Concepts	Binary	✗	✓	✓	✓
Graph Matching	Nodes	Binary	✗	✓	✗	✗
Word Sense Disambiguation	Documents	Binary	✓	✓	✗	✗
Text Similarity	Text	Graded	✓	✓	✗	✗
Paraphrase Detection	Statements	Binary	✓	✓	✗	✗
Semantic Relatedness	Concepts	Graded	✓	✓	✗	✗

Table 2.1: Overview of the tasks related to WSA. We list on what data the algorithms work on, how the outcome of an algorithm is usually expressed and what information sources can be exploited: textual descriptions (Desc.), structure of a document or resource (Struct.), meta information (Meta) or instantiations of concepts (Instances).

tasks such as semantic relatedness calculation, text similarity calculation and paraphrase detection all share some common traits with WSA, but only text similarity is a useful sub-task for WSA as the scope of the others is either too broad or too narrow. An overview of our findings is given in Table 2.1



# Chapter 3

## Resources and Datasets for Word Sense Alignment

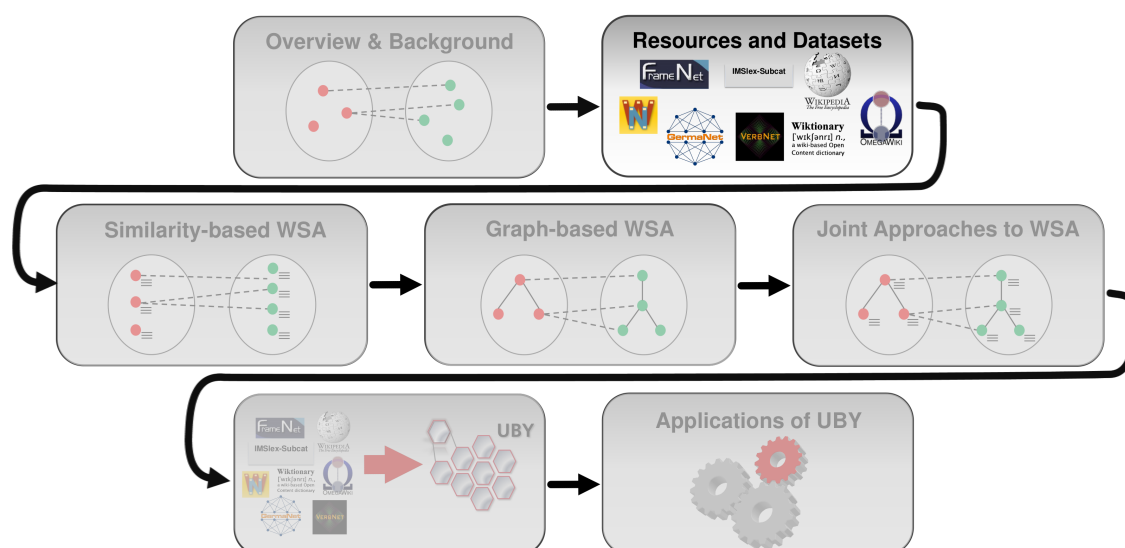


Figure 3.1: Visual outline of the thesis.

### 3.1 Introduction

The purpose of this chapter is to give an overview of our “working material” for WSA – the lexical-semantic resources which are to be aligned to each other, and the evaluation and development datasets in which they participate. As we have already outlined in the introductory chapter, there is a plethora of different resources with very different properties, and while a discussion of them could easily fill book volumes, we limit ourselves to a particular selection of resources which we describe. This selectivity is motivated by the greater context of our work, which is the construction of the unified resource UBY.

Simply put, the goal of UBY is to combine existing LSRs which are used for NLP into one “big” resource which covers all of their content in a unified representation

format and interface, and which also contains connections between them – the sense alignments. These enable a richer representation of concepts and thus better performance in NLP applications by a joint usage of resources. While this brief definition of UBY will suffice for the purpose of this chapter, we will discuss its construction in detail in Chapter 7. The resources that are contained in UBY and the rationale for their selection will be described here.

From the beginning of the UBY project, it was immediately clear that only a limited number of resources can be covered in such an effort. This is because each LSR to be integrated requires an extensive analysis of its content and properties, as well as the creation of transformation routines to the unified format, which both require considerable effort. After numerous discussions and considerations on how to limit our selection, we decided that LSRs to be contained in UBY should fulfill several properties:

- Coverage of at least two languages, with the purpose of investigating cross-lingual issues. German and English were chosen in this case.
- Subsequently, coverage of the expert-built resources for these languages most widely used in NLP, to ensure practical applicability and compatibility to previous work.
- Coverage of the largest collaboratively constructed resources, with the explicit goal to further the investigation and integration of these resources in NLP research.
- Furthermore, coverage of resources which have been constructed according to different paradigms, e.g. encyclopedic and lexicographic resources, resources with a focus on syntax, etc. in order to investigate how they can be harmonized in a unified format and, most important for this thesis, aligned at the level of word senses.

It goes without saying that these decisions are tailored to the specific needs we had in mind for this project – their rationale is open to discussion, and there are numerous other integration projects with equally valid considerations in the context of their creation. Prominent examples are *YAGO* (Suchanek et al., 2007) and *DBpedia* (Bizer et al., 2009), which focus on encyclopedic knowledge, as well as *NULEX* (McFate and Forbus, 2011), *SemLink* (Palmer, 2009) and the *Meaning Multilingual Central Repository* (MCR) (Atserias et al., 2004) which are more concerned with lexical-semantic knowledge. *BabelNet* (Navigli and Ponzetto, 2012a), like the MCR, has a strong focus on supporting multilingual applications, but also integrates Wikipedia and thus combines lexicographic and encyclopedic knowledge.

That being said, we will now present the resources that were selected for the integration into UBY, and thus were also the subject of the WSA work which we will present later on. We try to limit ourselves, as far as possible, to brief descriptions which only present the most salient features and statistics about the respective LSRs which are relevant in the context of our work. Where possible, we cite works which discuss the resources in more detail. One exception to this is the discussion of OmegaWiki – this resource has not been comprehensively covered in previous work,

thus we will provide a detailed description which might itself serve as a reference in the future. The other exception is the discussion of UBY, to which we dedicate a whole chapter (Chapter 7) so that the greater scope of our work becomes more evident.

An important aspect to keep in mind is that, in the course of the UBY project, it was decided to stick to fixed versions of the respective resources, i.e., for each resource we decided on a particular version or date (in case of continually updated LSRs) which was covered, deliberately disregarding previous or future versions. The motivation for this was, on the one hand, to avoid having multiple versions of the same resource in the same framework<sup>1</sup>, and on the other hand, to maintain re-usability of computed word sense alignments. An alignment can, per definition, only be computed between two particular versions of resources, as covered words and senses are bound to change (along with their identifiers) if a resource is extended or edited. Re-establishing an alignment would either require a complete recalculation, which can be computationally expensive, or a transfer of the existing alignments to the new versions, which can be tedious and error-prone.<sup>2</sup> Thus, for each resource we specify the version which was used for UBY, and unless otherwise stated, these versions of the resources were also used in the experiments throughout this thesis.

The rest of this chapter is structured as follows: in Section 3.2 we present the expert-built LSRs we use, while in Section 3.3 we discuss the collaboratively constructed ones, with a special focus on OmegaWiki (Section 3.3). In Section 3.4, we perform a thorough analysis of the covered LSRs with regard to their suitability for WSA, and thereby motivate the selection of resource pairs we consider for our WSA experiments. Following this, in Section 3.5, we present the WSA evaluation datasets we use for evaluation throughout this thesis: The ones created in previous work (Section 3.5.1) as well as the ones created by us (Section 3.5.2). In Section 3.6, we summarize this chapter as well as our contributions.

## 3.2 Expert-built Resources

Expert-built resources, in our definition of this term, are resources which are designed, created and edited only by a closed group of people, e.g. a research group at a university, the editorial board of a dictionary publisher, or the employees of a company. While it is possible that there is influence on the editorial process from the outside (e.g. via suggestions provided by users or readers), there is usually no direct means of participation. This form of resource creation has been predominant since the earliest days of lexicography (or, more broadly, creation of language

---

<sup>1</sup>Arguably, it can be desirable to have multiple (especially older) versions of the same resource available, e.g. when results from previous work need to be reproduced or when a particular version of a sense inventory is needed for evaluation in a shared task. However, to avoid certain versioning issues in the conceptualization of UBY, we decided to disregard this aspect in our current work. Nevertheless, this feature might still be integrated in future work.

<sup>2</sup>There are, of course, cases where a transfer of the alignment is possible without extensive effort, e.g. if identifiers for senses stay stable across versions. However, especially when senses are merged, split, deleted etc., this usually leads to information loss as the affected alignments should be considered incorrect and disregarded. Recovering these lost alignments between altered sense inventories would essentially require to repeat the original word sense alignment effort.

resources), and while this dependence on expert knowledge is believed to produce quality results, an obvious disadvantage are the slow production cycles – for all of the resources discussed in this section, it usually takes months (if not years) until a new version is published, while at the same time most of the information remains unchanged. This is due to the extensive effort needed for the creation of a resource of considerable size, in most cases provided by a very small group of people. Nevertheless, these resources play a major role in NLP. One reason is that up until recent years there were no real alternatives available, and some of these LSRs also cover aspects of language which are rather specific and not easily accessible for layman editors.

**WordNet (WN)** (Fellbaum, 1998) is a computational lexicon for English created at Princeton University, and probably the most popular NLP resource to date. It encodes concepts as synsets (i.e. sets of synonymous words) which are represented by textual definitions (glosses). A hierarchical organization is induced via semantic relations such as hyponymy, meronymy etc. This arrangement is psycholinguistically motivated, i.e. WordNet aims to represent real-world concepts and relations between them as they are commonly perceived. Version 3.0, which we use in our experiments, contains 117,659 synsets.

**GermaNet (GN)** (Hamp and Feldweg, 1997) can be considered the German counterpart to WordNet and is maintained at the University of Tübingen. It is also organized in synsets (around 70,000 in version 7.0) which are connected via semantic relations. Unlike WordNet, GermaNet originally contained very few glosses. In a recent effort, via aligning GermaNet with the German Wiktionary (cf. Section 3.5.1), this situation was rectified to make the resource more useful for NLP applications.

**FrameNet (FN)** (Baker et al., 1998) is an expert-built lexical-semantic resource (created at the ICSI in Berkeley) based on the theory of frame semantics (Fillmore, 1982), grouping word senses into frames which represent different situations. For instance, the verb *complete* and the noun *completion* belong to the “Activity finish” frame. The participants of these situations, typically realized as syntactic arguments, are the semantic roles of the frame, for instance the *Agent* performing an activity, or the *Activity* itself. Version 1.5 of FrameNet, which is used in UBY and thus in our experiments, contains 1,015 such frames and 11,942 word senses.

**VerbNet (VN)** (Kipper et al., 2006) is one of the largest verb lexicons available for English. It is organized into verb classes based on Levin’s classes (Levin, 1993). Each verb class is described by thematic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function. However, VerbNet senses do not have glosses and are implicitly defined by their usage patterns, which makes their understanding as well as their usage in NLP applications challenging. It is maintained by a research group at the University of Colorado at Boulder and contains approximately 4,600 verb senses.

**IMSLex-Subcat (IMSLex)** (Fitschen, 2004) is a dictionary which covers information on inflection, word formation, and valence for about 50,000 German base verb forms. From the IMSLex database, lexicon data for applications such as information retrieval can be extracted. UBY contains the current version from May 2003, but no reports on alignment efforts to other resources have been published so far.

### 3.3 Collaboratively Constructed Resources

In recent years, the emergence of the Web 2.0 has opened new possibilities for handling the aforementioned effort of constructing large scale lexical-semantic resources. The basic idea is that, instead of a small group of experts, a community of users (“crowd”) collaboratively edits and refines the lexical information. This information is in turn free to use, adapt and extend for everyone under permissive licenses. This is a major advantage of collaboratively constructed resources over efforts like GermaNet (Hamp and Feldweg, 1997), where the expert-built resource is subject to restrictive licenses. This crowd-based construction approach has turned out to be very promising to cope with the enormous effort of building such resources, since the large body of collaborators can quickly adapt to new language phenomena like neologisms while at the same time ensuring a remarkable quality – a phenomenon known as the “wisdom of crowds” (Surowiecki, 2005). The approach seems especially appropriate for multilingual resources as users speaking any language and from any culture can easily contribute through a wiki (Leuf and Cunningham, 2001) or another easily accessible interface. This is crucial for minor, usually resource-poor languages where expert-built resources are small or not available at all. Meyer and Gurevych (2012b) found, for instance, that the collaborative construction approach of Wiktionary yields language versions covering the majority of language families and regions of the world, and that it covers a vast amount of domain-specific descriptions not found in wordnets. Another issue is that expert-built resources usually have a narrow scope of information types. WordNet focuses, for example, on synsets and their taxonomy, but mostly disregards syntactic information. Finally, many expert-built resources utilize proprietary or even completely non-machine readable formats, which makes the integration into applications difficult.

However, despite their advantages, not all of these resources have been systematically investigated until now. This makes it hard to fully understand their characteristics and in turn exploit them for text analysis purposes. Therefore, following previous work on Wikipedia and Wiktionary (which we will also briefly describe here), the main goal of this section is to analyze OmegaWiki. To this end, we describe the way it emerged and characterize the resulting content. This is meant as a first step to integrating it into applications and the unified resource UBY, and it is particularly important in light of the alignment experiments we perform, as OmegaWiki shows some unique characteristics which in turn pose interesting challenges for WSA.

**Wikipedia (WP)** (Medelyan et al., 2009) is a collaboratively constructed online encyclopedia<sup>3</sup> and one of the largest freely available knowledge sources. The current English version contains around 4,400,000 articles and the German one around 1,700,000 articles, each of which usually describes a particular concept which (for our purposes) can be considered as word sense. Different senses of the same lexeme are usually marked by “bracketed disambiguations” in the article title such as *Java (island)* and *Java (coffee)*, and, due to its focus on encyclopedic knowledge, Wikipedia almost exclusively contains nouns. Related articles are connected via hyperlinks within the article text (implying a graph structure), and the first paragraph of an article usually gives a concise summary of the topic, which serves as a gloss for our experiments. UBY contains the Wikipedia dump version from 2009/08/22 with around 2,921,000 articles for English, and the dump from 2009/08/16 for German which contains around 845,000 articles. Although it was not created with this purpose in mind, Wikipedia is also commonly used as a resource in computational linguistics (Zesch et al., 2007).

**Wiktionary (WKT)** is a dictionary “side project” of Wikipedia which was created in order to better cater to the need to represent lexicographic knowledge, which is not well-suited for an encyclopedia. Wiktionary<sup>4</sup> is available in over 500 languages, and currently the English edition of Wiktionary contains over 500,000 lexical entry pages, while the German one contains around 350,000 ones. For each lexeme, multiple senses can be encoded, and these are usually represented by glosses. Wiktionary also contains hyperlinks which lead to synonyms, hypernyms, meronyms, etc. and a variety of other information types such as etymology or translations to other languages. The link targets are not disambiguated in all language editions however, e.g. in the English edition the links merely lead to pages for the lexical entries, which is problematic for WSA as we will see in Section 3.4.2. We use a snapshot from 2010/02/01 which contains around 421,000 senses for the English version and a snapshot from 2011/04/06 for the German one with around 72,000 senses. For interested readers, we refer to Meyer (2013) who analyzes Wiktionary from a lexicographic perspective and as a resource for NLP.

## OmegaWiki (OW)

### Overview

OmegaWiki<sup>5</sup> is a lexical-semantic resource which is freely editable via its web frontend (see Figure 3.2 for an example), and as such it is similar to Wiktionary. The current version of OmegaWiki contains over 46,000 concepts and lexicalizations in almost 500 languages (see Table 3.1). We use the database dump from 3 January 2010, which is about equal in size to the current version (see Table 3.7). Until now, it has only passingly been discussed in NLP as well as lexicography (Bergenholtz et al., 2009), which is why we aim at analyzing and describing it in more detail.

---

<sup>3</sup><http://www.wikipedia.org>

<sup>4</sup><http://www.wiktionary.org>

<sup>5</sup><http://www.omegawiki.org>



### ▼ Definition

Language ▾	Text
Castilian	Pez marino (Percicthyidae o Centrarchidae) popular para la pesca.
Dutch	Een zeevis (Percicthyidae of Centrarchidae) die populair is als sportvis.
English	A marine fish (Percicthyidae or Centrarchidae) that is popular as game.
French	Poisson d'eau de mer (Percicthyidae or Centrarchidae) populaire pour la pêche.
Slovak	Morská ryba (Percicthyidae alebo Centrarchidae) populárna ako lovná zver.

### ▼ Synonyms and translations

Expression	
Language ▾	Spelling
Castilian	<a href="#">lubina</a>
Castilian	<a href="#">róbalo</a>
Castilian	<a href="#">robalo</a>
Dutch	<a href="#">zeebaars</a>
English	<a href="#">bass</a>
French	<a href="#">basse</a>
German	<a href="#">Seebarsch</a>
Italian	<a href="#">spigola</a>
Japanese	<a href="#">バス</a>
Portuguese	<a href="#">robalo</a>
Swedish	<a href="#">bass</a>

### ▼ Annotation

Property ▾	Value
<a href="#">is part of theme</a>	<a href="#">fish</a>

### ▼ Class membership

Class ▾
<a href="#">animal</a>

Figure 3.2: An excerpt of OmegaWiki's defined meaning 5555 on *bass*. [http://www.omegawiki.org/DefinedMeaning:bass\\_\(5555\)](http://www.omegawiki.org/DefinedMeaning:bass_(5555))

One of OmegaWiki’s discriminating features in comparison to other such resources is that it is based on a fixed database structure which users have to comply to. It was initiated in 2006 and explicitly designed with the goal of offering structured and consistent access to lexical information, or as the creators put it: “The idea of OmegaWiki was born out of frustration with Wiktionary.”<sup>6</sup> The statement must be seen in context of Wiktionary’s creation paradigm; Wiktionary has been primarily designed to be used by humans rather than machines. The entries are thus formatted for easy perception using appropriate font sizes and bold, italic, or colored text styles, but for machines, data needs to be available in a structured manner in order to become able to obtain, for instance, a list of all translations or enumerating all English pronouns. This kind of structure is not explicitly encoded in Wiktionary, but needs to be inferred from the wiki markup.<sup>7</sup> Although there are guidelines on how to properly structure a Wiktionary entry, it is permitted to choose from multiple variants or deviate from the standards if this can enhance the entry. This presents a major challenge for the automatic processing of Wiktionary data. Another hurdle is the openness of Wiktionary – that is, the possibility to perform structural changes at any time, which raises the need for constant revision of the extraction software.

To alleviate Wiktionary’s problem of inconsistent entries caused by the free editing, the creators of OmegaWiki decided to limit the degrees of freedom for contributors by providing a “skeleton” of elements which interact in well-defined ways. The central elements of OmegaWiki’s organizational structure are language-independent concepts (so-called *defined meanings*) to which lexicalizations of the concepts are attached. These can be considered as multilingual synsets, comparable to resources such as WordNet. This way, no language editions exist for OmegaWiki as they do for Wiktionary. Rather, all multilingual information is encoded in a single resource. As such, OmegaWiki can be considered a lexicalized ontology (cf. Section 2.3.1). As an example, defined meaning no. 5616 (representing the concept HAND) carries the lexicalizations *hand*, *main*, *mano*, etc. and also definitions in different languages which describe this concept, for example, “That part of the fore limb below the forearm or wrist”. This method of encoding the multilingual information in a synset-like structure directly yields correct translations as these are merely lexicalizations of the same concept in different languages. Consequently, a “language edition” akin to Wiktionary would be obtained by only considering the set of concepts which are lexicalized in a certain language. It is of course also valid to have multiple lexicalizations in the same language, which are nothing else than synonyms. More details about the structure of OmegaWiki will be discussed in Chapter 7, where we present UBY and its underlying data model, UBY-LMF. There, OmegaWiki will serve as an example how resources can be mapped and transformed to a common model.

Table 3.1 and Table 3.2 show some statistics about the different “language edi-

---

<sup>6</sup><http://www.omegawiki.org/Help:OmegaWiki>, accessed on June 20th, 2012

<sup>7</sup>Wiki markup is an annotation language consisting of a set of special characters and keywords that can be used to mark headlines, bold and italic text styles, tables, hyperlinks, etc. within the article. The four equality signs in “====Translations====” denote, for example, a small headline that usually precedes the list of a word’s translations. This markup can be used by a software tool to identify the beginning of the translation section, which supposedly looks similar on each article page.

Language	Size
English	45,368
Castilian	35,549
French	28,565
German	24,589
Dutch	22,779
Italian	20,853
Portuguese	15,271
Swedish	11,447
Finnish	11,376
Polish	10,883
Russian	9,567

Table 3.1: Size of the OmegaWiki language editions or the ten most common languages as of 2014. This is calculated as the number of concepts (i.e. defined meanings) for which at least one lexicalization in the given language is available.

Resource	OmegaWiki en	OmegaWiki de
Translations	335,173	304,590
<i>...into Chinese:</i>	4,377	4,248
<i>...into English:</i>	-	56,471
<i>...into Finnish:</i>	18,997	19,536
<i>...into French:</i>	54,068	46,931
<i>...into German:</i>	56,471	-
<i>...into Italian:</i>	27,499	25,288
<i>...into Japanese:</i>	10,879	11,088
<i>...into Spanish:</i>	67,622	47,554
Languages	279	265

Table 3.2: Number of translations for selected languages and the sum of languages for which translations are available for the German and English parts of OmegaWiki as of 2014. So to speak, we only consider concepts for which English or German lexicalizations are available (see Table 3.1), and add up the lexicalizations in other languages for these.

tions” as well as about the translations between different languages that we derived from these multilingual synsets.<sup>8</sup> Note that the number of languages into which translations are available should be taken with a grain of salt, as for many languages only very few translations exist. Another important thing to note here is that the number of translations from English to German is the same as for the opposite direction. The reason is that translations only exist if a concept is lexicalized in both languages. The number of possible translations for a concept is then the product of the number of lexicalizations in either language, which is symmetric.

A useful consequence of this concept-centered design, especially for multilingual applications such as cross-lingual semantic relatedness (cf. Section 2.4.4), is that semantic relations are unambiguously defined between concepts regardless of exist-

<sup>8</sup>For the sake of illustration, we focus on the English and German parts of OmegaWiki, but our results and insights can for the most part be directly applied to other languages.

Relation	English	German
is part of theme	30,266	29,730
hypernym	23,909	12,292
hyponym	21,341	9,162
related term	6,242	5,941
subject	3,096	1,277
antonym	915	1,224
holonym	108	216
meronym	75	212

Table 3.3: The semantic relations applicable to the English and German parts of OmegaWiki. While the relations are defined between the language-independent defined meanings, we only consider those with a lexicalization in the given language.

ing lexicalizations. Consider for example the Spanish noun *dedo*: it is marked as hypernym of *finger* and *toe*, although there exists no corresponding lexicalization for the defined meaning FINGER OR TOE in English. This is for instance immediately helpful for translation tasks, since concepts for which no lexicalization in the target language exists can be described or replaced by closely related concepts. Exploiting this kind of information is not as easy in other multilingual resources like Wiktionary, where the links are not necessarily unambiguous (cf. Section 8.2.4).

Some statistics about the semantic relations are given in Table 3.3 for English and German. Note again that in OmegaWiki there is no explicit expression of synonymy as synonyms are just two lexicalizations of the same defined meaning. It is also noteworthy that category or domain labels which are common in many LSRs are also present in OmegaWiki, but expressed via relations and not mere labels. If a concept belongs to a particular theme or subject, the intention in OmegaWiki is to also include a concept for this and link accordingly, e.g. the concept FISH is linked to the concept BIOLOGY instead of representing the latter as a textual label. Table 3.4 gives an overview about the most frequent themes, and it is apparent that the focus of OmegaWiki labels is mostly on science. A more detailed discussion of the connectivity of the OmegaWiki relation graph will be given in Section 3.4.2 with regard to its suitability for graph-based WSA algorithms.

### Gaps and Criticism

OmegaWiki’s fixed structure is manifested in an SQL database, and it is proprietary in the sense that it does not conform to existing standards for encoding lexicographic information such as the Lexical Markup Framework (Francopoulo et al., 2006). Plainly spoken, it was designed and over time extended in a “grass-roots approach” by the community to cater for the needs identified for such a multilingual resource. While this approach to structuring the information is not easy to tackle in terms of interoperability, it still makes the use of this resource easier than for Wiktionary. The underlying database ensures straightforward structured extraction of the information and less error-prone results due to the consistency enforced by the definition of database tables and relations between them. However, the fixed structure also has the major drawback of limited expressiveness. As an example,

English theme	Size	German theme	Size
biology	1,556	Soziale Aspekte	1,026
economics	1,078	Biologie	746
biological science	778	Chemie	715
chemistry	729	Industrie	637
administration	708	Verunreinigung	599
material	662	Wasser	586
industry	660	Landwirtschaft	581
agriculture	606	Ökonomie	528
water	598	Bevölkerung	513
pollution	561	Forschung	483

Table 3.4: The 10 most frequent theme labels applicable to the English and German parts of OmegaWiki. While the themes are defined for the language-independent defined meanings, we only consider those with a lexicalization in the given language.

the coding of grammatical properties is only possible to a small extent. Complex properties such as verb-argument structures cannot be encoded at all, because they have not been catered for in the underlying database design. Moreover, an extension of this structure is not easy, as this would, in many cases, require a reorganization of the database schema by administrators to which present and future entries would have to conform. While it could be argued that such information is outside of the scope of the resource and thus does not need to be reflected, the possibility given in Wiktionary to encode (in theory) any kind of lexicographic information using the more flexible wiki markup makes it more attractive for future extension. In OmegaWiki, the users are not allowed to extend the structure and thus are tied to what has been already defined.

Another issue with the database-centric implementation is that, unlike for a genuine wiki implementation which is based on documents (like, for instance, MediaWiki), a full-fledged revision history is not available for OmegaWiki. For Wiktionary, every edit (down to single characters) can be tracked, and past states of the article pages can be reconstructed. While this would theoretically also be possible for OmegaWiki, the interface only offers an overview of the database commit operations, and no option to review or revert them – this is only possible for database administrators, and, in general, reverting to a previous database version is still problematic as consistent database states must always be ensured. Like Wiktionary, OmegaWiki has an attached discussion page for each entry so that users can collaboratively sketch and edit entries, but with the missing means to transparently track and discuss past changes this possibility is hardly ever used.

While the database structure has its limitations, it allows expressing many types of lexicographic information – however, the interface is also a limiting factor in this respect. While the most basic information types like definitions and lexicalizations are plainly visible and editable (see Figure 3.2), all other possibly useful pieces of information are subsumed as *annotations* and initially hidden in the interface, which is apparently an obstacle to editing them. This includes essentials like part of speech information, but also example sentences, hyphenation and phonetic pronunciations. Moreover, adding this information is rather tedious, as the user is overburdened

Annotation	Size
hyphenation	22,847
example sentence	3,915
International Phonetic Alphabet	1,125
pinyin	806
usage	306

Table 3.5: The most common lexicographic annotations available for OmegaWiki entries, not including part of speech. Pinyin is the most common transcription system for Chinese characters into the Latin alphabet, see ISO7098 (1991).

Part of speech	Size
Noun	10,769
Verb	1,332
Adjective	1,864
Adverb	359
Other	37
None	37,327

Table 3.6: The distribution of different parts of speech in OmegaWiki. For many entries, no such information is available.

with selecting the correct database fields without any further explanations about their meaning. Thus, for only a small number of entries such additional information is available (where hyphenation is by far the most common one, see Table 3.5), and part of speech information is also often missing (Table 3.6). For these cases, the database design explicitly allows `null` values. For the WSA experiments on OmegaWiki we discuss in the following chapters, this meant that we had to relax the condition that only senses with matching parts of speech should be aligned (cf. Section 2.2) in order to ensure satisfactory coverage.

Consequently, OmegaWiki’s lack of flexibility and extensibility, in combination with the rather unintuitive interface and the fact that Wiktionary was already quite popular at its creation time, has caused the OmegaWiki community to remain rather small. While OmegaWiki had 6,746 users at the time of writing, only 19 of them had actively been editing in the past month, i.e. the community is considerably smaller than for Wikipedia or Wiktionary (Meyer, 2013). Because of this, OmegaWiki has grown rather slowly (see Table 3.7) and remained small in comparison to other LSRs. It also only has moderate lexeme overlap with the most resources in UBY (Tables 3.8 and 3.9), with the exception being the verb-focused resources FrameNet and VerbNet. However, it should be kept in mind that these resources are also rather small, so that the absolute overlap is modest. OmegaWiki also has a comparatively low degree of polysemy (Table 3.10), which is evidence that many words and senses covered by other resources are missing, a phenomenon which we will further discuss with regard to the gold standard datasets for WSA (Section 3.5) and the exploitation of OmegaWiki for sense clustering (Section 8.1). Calculating the sense overlap of resources amounts to calculating full sense alignments between them as described in the following chapters. We will provide statistics about this in Section 7.4.

	WKT 2010	WKT 2014	OW 2011	OW 2014
Entries (Total)	14,021,155	20,401,055	442,723	557,763
Entries (English)	2,457,506	3,737,251	55,182	60,347
Entries (German)	177,124	364,117	34,559	34,889
Languages covered	>400	>1400	290	482
Languages with >10.000 entries	54	72	12	12

Table 3.7: Descriptive statistics about OmegaWiki (OW), in comparison to Wiktionary (WKT), considering the versions from 2010 contained in UBY as well as the current versions. It is clearly visible that Wiktionary experienced a significant growth, while the size of OmegaWiki mostly remained stable. “Entries” refers to lexical entry pages in case of Wiktionary and lexicalizations in a particular language for OmegaWiki. This number can be larger than the total number of concepts as multiple lexicalizations for a concept may exist.

Coverage of...	lexeme	lemma only
WordNet	6.7%	16.6%
Wiktionary	3.2%	9.5%
Wikipedia	0.3%	0.7%
FrameNet	30.1%	60.3%
VerbNet	23.1%	58.9%

Table 3.8: The number of English lexical items from other LSRs in UBY covered by OmegaWiki, i.e. the lexical coverage. As many OmegaWiki entries do not contain part of speech (POS) information, we distinguish between matching only the lemma or the full lexeme (lemma + POS).

Coverage of...	lexeme	lemma only
GermaNet	3.8%	17.2%
Wiktionary	7.3%	27.5%
Wikipedia	0.3%	1.5%
IMSLex	3.4%	27.3%

Table 3.9: The number of German lexical items from other LSRs in UBY covered by OmegaWiki, i.e. the lexical coverage. As many OmegaWiki entries do not contain part of speech (POS) information, we distinguish between matching only the lemma or the full lexeme (lemma + POS).

Despite the above mentioned issues, we still believe that OmegaWiki is not only interesting for usage in NLP applications (and thereby for integration into UBY), but also as a case study, since it exemplifies how the process of collaboratively creating a large-scale lexical-semantic resource can be moderated by means of a structural “skeleton” in order to yield a machine readable result. Exploiting this property of OmegaWiki, we also developed a Java-based API which allows easy programmatic access to the resource – this is discussed in Appendix A.1. We also deem it especially interesting for investigation of WSA algorithms due to its remarkable properties, which we will discuss in the next section.

Part of speech	Polysemy
Noun	1.11
Verb	1.29
Adjective	1.15
None	1.12

Table 3.10: The polysemy in OmegaWiki, i.e. the ratio of senses per lexeme. In comparison to other resources, OmegaWiki is quite coarse-grained.

### 3.4 Analysis of LSRs

In the previous two sections, we have presented the resources we decided to include into UBY and pointed out that they have very different properties, as they have been constructed according to different paradigms and with different applications in mind. While the variety of resources already suggests that WSA between any pair of them presents interesting challenges, we deem it necessary to further substantiate the choice of resources for the actual WSA experiments – after all, with as much as nine resources contained in the first release of UBY,<sup>9</sup> we would have (theoretically) no less than 36 potential alignment pairs to investigate. As time and computational resources are limited, we thus have to make (and motivate) a selection which is reasonable not only considering the benefit for the overarching UBY project, but also with respect to the potentially interesting research questions for WSA.

Therefore, we will present in this section a comparative analysis of the LSRs which goes beyond previous work such as (Garoufi et al., 2008) or (Meyer and Gurevych, 2010): these works provided in-depth analyses of various resources and also discussed the potential of aligning them by pointing out differences in lexical coverage and covered information types. However, such a perspective is not sufficient for our particular task, as metrics concerning the suitability of an LSR for WSA are largely disregarded. For instance, if we know that a certain lexeme is covered by both WordNet and Wiktionary, we know that there is potential for an alignment – however, we do not know on what basis we can make a well-informed (and thus correct) alignment decision. As our discussion of existing resources and approaches showed, two types of information are available for the vast majority of LSRs which are immediately relevant in this case: i) glosses, or more general, textual descriptions of senses or concepts, and ii) relationships between concepts inducing a graph, given through semantic relations, links, or other means. Thus, in this section we will add another facet to the discussion of LSRs and investigate these two aspects regarding different parameters, and relate our results to the challenge at hand. This is, in spirit, related to the efforts made in the field of corpus analysis (see, for instance, (Biber et al., 1998) for an overview) which aims to comprehensively describe and analyze corpora and relate these observations to properties of systems that build upon them and, more generally, language phenomena which can be examined in such a corpus. In other words, we consider the set of glosses in an LSR as a set of (short) documents to analyze, and add the additional layer of structural analysis by

<sup>9</sup>IMSLex-Subcat was not contained in the first version of UBY and was also not subject of any previous work on WSA. We will thus mostly disregard it for the remainder of this thesis.



Resource	Senses	Empty	Tokens			Type/Token
	Total	Glosses	Max.	Avg.	Median	Ratio
WordNet	117,659	2	505	51.2	44	3.8%
FrameNet	11,942	18	316	50.2	44	9.2%
GermaNet	74,612	63,936	242	37.3	32	24.4%
Wiktionary en	421,848	122,541	1,455	60.0	46	3.5%
Wiktionary de	72,752	14	1,277	67.6	53	12.4%
Wikipedia en	2,921,455	19	17,524	273.4	210	1.2%
Wikipedia de	838,428	1	10,362	250.5	196	4.2%
OmegaWiki en	45,137	2,334	2668	64.0	37	6.0%
OmegaWiki de	24,509	14,573	897	72.2	59	19.8%

Table 3.11: Statistics about the glosses of the considered LSRS. The type/token ratio is defined as the number of different tokens in the text (in this case, the combination of all glosses) divided by the total number of tokens and expresses the lexical variety of a text. Note that VerbNet does not have glosses and was thus not considered for this analysis.

also briefly examining the relations between concepts described by these glosses.

### 3.4.1 Analysis of Glosses

The first observation about the glosses is that the expert-built resources WordNet and FrameNet only have very few senses with missing glosses – this is to be expected considering their aim to provide consistent and complete knowledge about the concepts covered. For the other two expert-built resources we cover, GermaNet and VerbNet, the situation is quite different. VerbNet has no glosses at all (and is thus omitted from Table 3.11) and relies on the implicit “definition” of senses by example sentences and verb classes they belong to. GermaNet has glosses for only a “core” subset of concepts, while the majority of them is also implicitly defined via their relations to other concepts.

For Wikipedia, there are also almost no gaps – keep in mind that the glosses in this case are the first paragraphs of the articles, and empty “stub” articles are usually either quickly deleted or extended by the community, which leads to the expected observation that empty articles are very rare.

For the other collaboratively constructed resources OmegaWiki and Wiktionary, the situation is quite different. We observe many missing glosses, and at this point the disadvantages of the missing quality control in the collaborative construction process become apparent. Because in a dictionary an entry without any definition can still provide some useful information such as pronunciation (while in contrast an encyclopedia article without definition would be utterly useless), these entries are usually not deleted right away. However, their extension might take some time depending on different factors such as the frequency of a word or the availability of domain experts for a certain topic; such incomplete entries are usually not accepted in an expert-built resource. Considering these factors, the high quality of the German Wiktionary is remarkable – there are almost no missing glosses, so that this LSR rivals expert-built resources in this respect. This speaks in favor of the

	FrameNet	WordNet	OmegaWiki en	Wiktionary en	Wikipedia en
FrameNet	-	76.8%	89.4%	96.0%	97.8%
WordNet	19.3%	-	41.1%	81.3%	95.4%
OmegaWiki en	26.5%	65.8%	-	85.1%	95.9%
Wiktionary en	8.8%	34.4%	22.5%	-	86.2%
Wikipedia en	0.5%	2.2%	1.4%	4.6%	-

Table 3.12: Lexical overlap of glosses in the English resources covered.

very thorough and quality-focused German Wiktionary community, which was also described by Meyer and Gurevych (2010).

Apart from the mere presence or absence of glosses, we also analyzed their lengths. First of all, we observe that the expert-built resources all show very similar properties, with WordNet and FrameNet being almost indistinguishable. This suggests that in expert lexicography, there is a common understanding on the appropriate verbosity of glosses – overly long glosses are rare, and even the longest ones are still far shorter than the corresponding glosses in collaboratively constructed dictionaries. There, we observe not only higher maximum lengths, mostly due to overly long, technical definitions, but also a higher average. Although the length of a gloss is not necessarily correlated with its quality, this at least suggests that contributors to Wiktionary and OmegaWiki also aim at providing complete and informative sense description. Interestingly though, the difference between the median and the average is substantially higher in these resources than for the expert-built ones, i.e. we observe a higher variance of gloss lengths, which is understandable in light of the large number of different authors with different takes on the verbosity of glosses. For Wikipedia, all numbers are naturally much larger than for the other LSRs, since (as mentioned above) we consider the first paragraph as a gloss.

As a last means of analyzing the glosses of the individual resources, we calculated the type/token ratio (TTR) which is defined as the number of different tokens in the text divided by the total number of tokens. It is usually considered as an indicator of the lexical variety of a document, and in this respect (as for the length of the glosses) we observe similar values for expert-built and collaboratively constructed resources. This indicates that the richness of vocabulary is not necessarily worse in the latter group – the TTR difference between different resources can mostly be attributed to the differences in the number of glosses and hence the total number of considered tokens. Generally speaking, in case of relatively few glosses (e.g. for GermaNet or OmegaWiki), the TTR is bound to be high as a certain “baseline” of lexical variation is required for different definitions, while for Wikipedia the TTR is very low simply because of the abundance of glosses which inevitably contain lots of tokens which are repeated in other articles. As a general observation, it is remarkable that the TTR for German is much higher than for English. For instance, for each collaboratively constructed LSR we get approximately three times the TTR for German as compared to the English version. This is probably due to the higher lexical variety, stronger inflection and different formation of compounds in German which allows for the more frequent formation of rarely observed tokens.

As an additional layer of analysis, we also investigated the pairwise lexical overlap between the glosses for LSRs in the same language (see Tables 3.12 and 3.13),

	GermaNet	OmegaWiki de	Wiktionary de	Wikipedia de
GermaNet	-	44.8%	78.6%	90.9%
OmegaWiki de	23.5%	-	68.4%	90.3%
Wiktionary de	10.5%	17.3%	-	81.4%
Wikipedia de	0.8%	1.4%	5.1%	-

Table 3.13: Lexical overlap of glosses in the German resources covered.

i.e. the number of individual tokens used in glosses of one resource which are also contained in the other. Intuitively, this is an indicator to what extent the “vocabularies” of the two resources match, and the higher this value is the more likely it is that a meaningful gloss similarity value for two senses from these resources can be computed.

For FrameNet, we observe that it has a high overlap with all resources, which makes sense considering the relatively few glosses and the resulting small “pool” of tokens, which are mostly also found in the other, considerably larger LSRS. Interestingly though, it has more overlap with collaboratively constructed resources than with WordNet – this suggests it would be especially suitable for similarity-based alignment with these. Wiktionary and Wikipedia also have a high overlap, which comes at no surprise due to the parallel development of both LSRS and their at least partially overlapping set of contributors.

A very interesting observation for English is that WordNet and OmegaWiki have a rather low lexical overlap, which indicates a substantially different vocabulary, although both resources have a high overlap with Wiktionary (and also Wikipedia). The plausible explanation from a set-theoretic perspective is that both OmegaWiki and WordNet share large portions of their vocabulary with Wiktionary, but *different* portions. This is also in line with the different thematic foci of the three resources (expressed, for instance, by domain labels) which we discussed in Section 3.3 and which was also observed by Meyer (2013). A visualization of this intuition is given in Figure 3.3. A very similar observation can be made for GermaNet and the corresponding German LSRS, so that it seems plausible that similarity-based WSA approaches would work substantially worse when aligning WordNet and GermaNet to OmegaWiki in comparison to the other collaboratively constructed resources.

### 3.4.2 Analysis of the Graph Structure

For the purpose of the structural analysis (and the graph algorithms we present later on), we consider the set of senses (or synsets, if applicable) of an LSR  $L$  as a set of nodes  $V$  where the set of edges  $E \subseteq V \times V$  between these nodes represents semantic relatedness between them. A Wikipedia article is considered a sense, as it represents a distinct concept.

There are multiple options for deriving the edges from the resources. The most straightforward approach is to directly use the existing semantic relations (such as hyponymy), as it has been reported in previous work (Navigli, 2009a; Laparra et al., 2010) – these are present in OmegaWiki, WordNet and GermaNet. For FrameNet, there are no semantic relations between senses, but between frames that contain them, and senses in the same frame can also be linked. For instance, two different

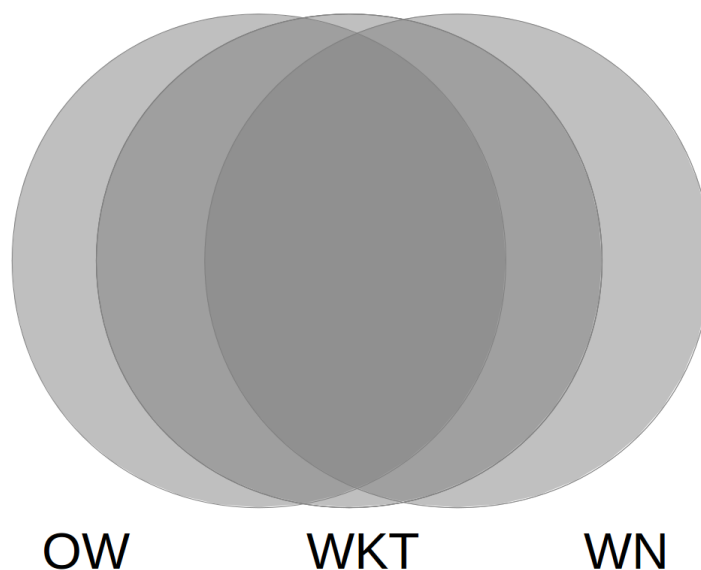


Figure 3.3: Visualization of the set-theoretic intuition that, while OmegaWiki (OW) as well as WordNet (WN) glosses have a high lexical overlap with Wiktionary (WKT) glosses, their mutual overlap is considerably lower.

Resource	Senses	Relations	Relations/Sense	Isolated Senses
WordNet	117,659	570,696	4.85	25%
FrameNet	11,942	76,315	6.39	2%
VerbNet	31,891	197,824	6.20	0%
GermaNet	74,612	193,669	2.60	0%
Wiktionary en	421,848	5,132	0.01	98%
Wiktionary de	72,752	44,523	0.61	69%
Wikipedia en	2,921,455	83,220,212	28.49	4%
Wikipedia de	838,428	12,965,148	15.46	4%
OmegaWiki en	45,137	62,104	1.38	41%
OmegaWiki de	24,509	32,705	1.33	45%

Table 3.14: This table describes, among other statistics, what percentage of nodes remains isolated (i.e. with no attached edges) using semantic relations given in the LSRs as edges. Note that this number is highest for the English Wiktionary as the few unambiguous semantic relations do not offer many possibilities for connecting nodes, while the German Wiktionary and OmegaWiki do not suffer from this problem as much. GermaNet is fully linked via relations, and also WordNet, FrameNet and Wikipedia are relatively well-linked.

senses of *to see* are contained in the frames “Perception\_experience” (along with *taste* and *smell*) and “Categorization” (along with *perceive*). The same rationale is applicable to the VerbNet verb classes which group senses together. Using these relations, we get intuitively plausible graphs, i.e. graphs which connect groups of semantically related concepts. For this group of LSRs, we observe that in the expert-built resources, the large majority ( $> 70\%$ , see Table 3.14) of nodes are connected by sense relations, while this is not the case for OmegaWiki. OmegaWiki also contains far fewer relations per sense than the expert-built resources, which suggests that adding relations between senses has a lower priority than editing the description of the actual concepts. This might, at least partly, be also due to the issues with the OmegaWiki interface we discussed in Section 3.3 and might impair the performance of graph-based WSA algorithms.

Not for all LSRs such straightforwardly usable relations exist. This is especially an issue for the English Wiktionary as its relations are not sense disambiguated (Meyer and Gurevych, 2012b), i.e. they only lead to lexical entries. We thus cannot determine the correct target sense if a relation is pointing to an ambiguous word. While this is no problem for human users, it severely interferes with Wiktionary’s usability for NLP applications. As a workaround to still obtain a graph for analysis we create an edge  $(s, t)$  for each sense  $s$  for those semantic relations which have a monosemous target  $t$ , as in this case the target sense is unambiguous. It might still be wrong, but from our observation the target is correctly disambiguated in over 90% of the cases in this way. This approach obviously only recovers a subset of all encoded relations, however, which results in a sparse graph with many isolated nodes.<sup>10</sup> For the German Wiktionary, the relations are disambiguated so that this issue does not apply, and this is directly reflected in fewer isolated nodes and a higher number of relations per sense. Another factor contributing to this difference is that, as we reported in (Matuschek et al., 2013), the English Wiktionary is almost 6 times as large as the German one for the versions we used in our experiments (421,000 senses vs. 72,000 senses), while it contains not even twice as many relations (720,000 vs. 430,000). This means that even if we disregard the issue of ambiguous relation targets the German Wiktionary is still considerably denser – nevertheless, it is still no match for expert-built resources which contain an order of magnitude more relations.

For Wikipedia, we can directly use the given hyperlinks between articles as they also express a certain degree of relatedness (Milne and Witten, 2008); these links are also unambiguous as they always lead to a distinct article. For this resource, we observe an extremely high number of relations per sense in comparison to the other LSR. This is not surprising considering typical Wikipedia articles which are usually well-linked – this is also in line with the very small fraction of isolated nodes, which should intuitively make the resource suitable for graph-based alignment, just like the expert-built resources discussed above.

---

<sup>10</sup>An effort to alleviate this problem was undertaken in the OntoWiktionary project (Meyer and Gurevych, 2012a), which aimed at disambiguating Wiktionary relations and inferring new ones. While we briefly experimented with these enriched versions of Wiktionary, we could not gain any notable improvement over the approaches to graph construction we present here. The recall for our graph-based WSA algorithms usually got better (as should be expected), but at the same time the precision dropped sharply. A thorough analysis of this behavior is beyond the scope of this thesis and will thus not be discussed here.

### 3.5 Selection of Evaluation Datasets

After presenting the LSRs which we decided to consider for WSA, we now want to go one step further and discuss which *pairs* of LSRs are covered in our experiments. Especially, we lay out the motivation for this selection, regarding the properties of the resources discussed in the previous sections. As mentioned above, limitations of time and computational resources forced us to carefully constrain our research efforts.

The first, and simplest, inclusion criterion is to use pairs for which previous work exists, as this allows us to use previously created gold standard datasets and spare us the effort of manually creating new ones. Apart from the obvious benefit of reduced work load, these pairs are also interesting from a research perspective in their own right – for the most part, the previous efforts had motivations for considering these resources similar to our own. For instance, WordNet is the resource which is covered most often as it is the most comprehensive and widely used LSR for English, and the decision to align it to the most popular collaboratively constructed resources Wiktionary (Meyer and Gurevych, 2011) and Wikipedia (Niemann and Gurevych, 2011; Navigli and Ponzetto, 2012a) seems immediately plausible. A similar argumentation holds for the most popular German resources, GermaNet and the German Wiktionary, which were considered by Henrich et al. (2011). Considering the similar properties of the resources involved with regard to the glosses (and especially the large lexical overlap), the good results for the similarity-based approaches presented in the respective papers seem logical (see Table 3.15). The same holds for the similarity-based alignment between FrameNet and Wiktionary discussed by Hartmann and Gurevych (2013). Nevertheless, no graph-based alignments have previously been investigated for these datasets, which is the main reason we include them in our analyses – especially the sparsity of the Wiktionary graph as compared to the other resources should present an interesting challenge.

The VerbNet-FrameNet and VerbNet-WordNet alignments, presented by Palmer (2009) and Kipper et al. (2006), respectively, are particularly interesting for the inclusion into UBY, as they are manually created (or at least validated) and thus of high quality, while at the same time representing full alignments between the LSRs. This, however, means that an automatic alignment for these particular LSR pairs would not be strictly necessary for all practical purposes – there would be no conceivable way to outperform a manual full alignment in precision or size. Nevertheless, we experimented with these datasets as a “test case” in order to see to what extent the human annotation could be reproduced by means of an automatic alignment algorithm, and more importantly, to see in what way VerbNet with its unique focus on syntactic usage of verbs and lack of glosses could be integrated into our WSA framework. We will present this and all other datasets we used and which were created in previous work in detail in Section 3.5.1.

It is clear that there are also many resource pairs which were not considered in previous work, but still seem worth investigating. For instance, OmegaWiki has not been covered at all in previous work, and in order to understand how its properties (relatively small/few glosses and few semantic relations) affect different approaches to WSA, we decided to investigate its alignment to the two following LSRs:

- **WordNet** Not only is WordNet the most popular LSR for NLP applications and thus an obvious choice, its glosses also have a relatively low lexical overlap with OmegaWiki, so that an investigation of similarity-based measures for this case is reasonable. As OmegaWiki is inherently multilingual (see Section 3.3), we also want to use this resource pair as a testbed for cross-lingual LSR alignment, a task which would be substantially more challenging with the also multilingual Wiktionary, as links to translations are ambiguous in this case. Moreover, the much sparser resource graph of OmegaWiki in comparison to WordNet is also an interesting factor in the discussion of graph-based WSA approaches.
- **Wiktionary** No alignment between two collaboratively constructed resources has been suggested so far, and aligning the two most important dictionaries in this area is the first logical step to take. In addition, while this setup includes two resources with relatively high lexical overlap of glosses, both resource graphs are relatively sparse, so it will be interesting to see to what extent graph-based approaches can be effective in this case. Another motivation for considering this pair is the multilinguality of both involved LSRs – we also investigate the usage of aligned resources for translation applications in the course of this thesis (Section 8.2), for which this pair seems ideally suited.

Finally, we investigate an alignment between Wiktionary and Wikipedia. First of all, this pair provides another set of circumstances for WSA algorithms which is not covered by the pairs mentioned above (high vocabulary overlap, different resource graph density), and it also seems imperative to investigate this case as one of Wiktionary’s explicit purposes is to complement the knowledge in Wikipedia (cf. Section 3.3). This is the core idea of all WSA efforts, which is why an alignment between these widely used resources, which are also the two largest collaboratively constructed resources to date, seems a natural and important extension to the body of work in this field. Also, as both resources are multilingual, we can investigate both the English and the German case, which adds a second German dataset to our portfolio. This allows us to make more general statements about the influence the language has on the alignment performance.

In every case, we created novel datasets for the resource pairs not covered thus far, filling the gaps in the construction of UBY and broadening the foundation for WSA research in general. The datasets created in the course of our work will naturally be discussed in greater detail than the others (see Section 3.5.2).

We would also like to briefly discuss the pairs that we decided to omit from our analysis (cf. Table 3.16). Obviously, there are a few combinations which do not make sense, for instance Wikipedia-VerbNet, which both exclusively cover different parts of speech (nouns and verbs), but apart from that, there exist a couple of plausible combinations which were still disregarded, for various reasons:

- In the case where manually created or validated full alignments already exist, especially for the cross-lingual links between the German and English editions of OmegaWiki and Wikipedia, we refrain from reproducing these alignments as they promise little additional insight and no immediate practical benefit.

The alignments involving VerbNet are exceptions to this, as already mentioned above.

- FrameNet-WordNet alignments have been investigated several times in the past (Ferrandez et al., 2010; Laparra et al., 2010). However, the different previous works also used different versions of the resources, which would be an obstacle for the integration into UBY. Moreover, the corresponding evaluation datasets are only partially available, which makes investigating this case with reasonable effort infeasible. Nevertheless, we still want to cover this pair of LSRs in future work.
- A very interesting case for cross-lingual WSA would be an alignment between WordNet and GermaNet – a possible gold standard for this is provided by the interlingual index of EuroWordNet (Vossen, 1998). However, due to unclear licensing issues and outdated versions of the LSRs used for this index we refrain from using this dataset.
- In any case, the integration of further cross-lingual alignments (on top of WordNet-German OmegaWiki) seems reasonable, but our preliminary experiments show that the necessary translation step has little influence on the alignment results, at least for English and German (see Section 4.4). Therefore, we postpone further work on this issue for now as little additional insight is to be expected.
- We created monolingual GermaNet-Wikipedia and GermaNet-Wiktionary alignments for the WSA-based word sense clustering method presented in Section 8.1. As the WSA performance was not the focus of this work, however, no gold standards were created for these cases. One reason for this is the fact that the alignment task is quite similar to the corresponding English WordNet-Wikipedia and WordNet-Wiktionary alignment scenarios, so that the additional insight was expected to be modest. Nevertheless, the good sense clustering results (evaluated on coarse-grained WSD) suggest that our alignment is also effective in this case.

An overview of all resource pairs is given in Table 3.15. We list, for the original work where the dataset was introduced as well as for our own efforts, the results using the respective algorithms. Moreover, we state the sizes of the obtained full alignments as well as the information whether an alignment was included in the first (0.1.0) or current (0.6.0) version of UBY. For completeness, we also list the existing datasets we did not use for our experiments, marked with parentheses.

As a supplement, the matrix in Table 3.16 displays in a different way which resources contained in UBY have been considered for alignment, and also which ones are subject to future work. Finally, Figure 3.4 gives a graphical overview of the existing alignments.

Note that at this point these tables do not need to be fully understood, they are merely meant as a visual “outline” of what has been done by us and others, i.e. what will be covered in the remainder of this thesis. In particular, the algorithmic approaches from the previous work as well as our own efforts will be discussed in



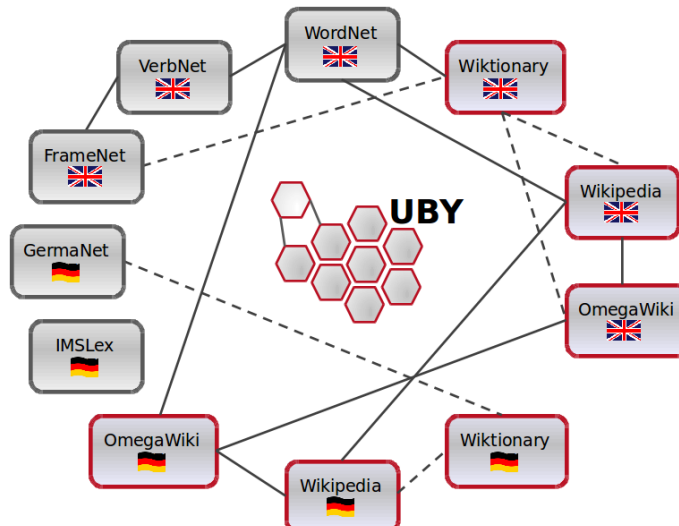


Figure 3.4: An overview of the resources and alignments contained in UBY in versions 0.1.0 (solid lines) and 0.6.0 (dashed lines).

detail in Chapters 4, 5 and 6, while more details on the unified resource UBY, its different versions and the alignments it contains will be given in Chapter 7.

The characteristics of the gold standard dataset will be discussed in the remainder of this chapter, and in all cases, these gold standards encode pairwise sense alignments (according to the definition in Section 2.2) between two of the resources described above. They are created by humans, either by a small group of annotators, or by a crowd of contributors to a collaboratively constructed resource. Due to creation and validation by humans, these datasets are known to be correct,<sup>11</sup> and are thus the reference (or gold standard) for the performance of automatic algorithms. As such, they are essential for developing and evaluating WSA techniques.

Notably, due to the size of many of the resources and the complexity of the task, it is uncommon to manually annotate a full alignment, i.e. between all applicable senses from two LSRs. While there are notable exceptions such as (Palmer, 2009) and (Henrich et al., 2011) which we also describe later on, the usual case is that only a subset of senses is selected as a representative sample for the full alignment setup. Detailed information about all datasets is available in Tables 3.17 and 3.18, including the observed inter-rater agreement  $A_0$  (where available) which can be considered as an upper bound for automatic alignment accuracy, the annotator F-measure (i.e. the F-measure if one annotation is regarded as gold standard and the others are evaluated against it) and the degree of ambiguity (i.e. the number of possible alignment targets per sense) which is a hint towards the difficulty of the alignment task.

<sup>11</sup>It goes without saying that humans do not agree in every case, so that it is debatable for each instance whether it has been correctly annotated or not; the degree of agreement is usually measured though (as explained in Section 2.2.1). The gold standards are correct insofar as disagreements are resolved by majority voting or validation by an expert annotator.

Resource Pair	Source	Abbrev.	Approach	P	R	$F_1$	Size	UBY 0.1.0	UBY 0.6.0
WN-WKT	(Meyer and Gurevych, 2011) (Matuschek and Gurevych, 2014)	MG11 <b>MG14</b>	Gloss similarity Machine learning	0.67 0.70	0.65 0.84	0.66 0.77	99,662 138,282	✓ ✗	✗ ✓
WN-WP	(Niemann and Gurevych, 2011) (Matuschek and Gurevych, 2013) (Navigli and Ponzetto, 2012a)	NG11 <b>MG13</b> NP12	Gloss similarity Shortest paths Gloss/Structure	0.78 0.75 0.81	0.78 0.87 0.75	0.78 0.81 0.78	50,351 83,192 47,956	✓ ✗ ✗	✗ ✓ ✗
WN-OW	(Gurevych et al., 2012) (Matuschek and Gurevych, 2014)	<b>Gu12</b> <b>MG14</b>	Gloss similarity Machine learning	0.55 0.75	0.53 0.72	0.54 0.74	23,024 27,529	✓ ✗	✗ ✓
FN-WKT	(Hartmann and Gurevych, 2013) (Matuschek and Gurevych, 2013)	HG13 <b>MG13</b>	Gloss similarity Shortest paths	0.73 0.77	0.75 0.79	0.74 0.78	12,326 12,340	✗ ✗	✗ ✓
GN-WKT	(Heinrich et al., 2011) (Matuschek and Gurevych, 2013)	He11 <b>MG13</b>	Gloss overlap Shortest paths	0.84 0.83	0.84 0.86	0.84 0.85	27,127 32,850	✗ ✗	✗ ✓
WKT-OW	(Matuschek et al., 2013) (Matuschek and Gurevych, 2014)	<b>Ma13</b> <b>MG14</b>	Gloss similarity Machine learning	0.67 0.79	0.65 0.64	0.66 0.71	25,742 25,727	✗ ✗	✗ ✓
WKT-WP en	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.82	0.70	0.76	66,050	✗	✓
WKT-WP de	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.89	0.81	0.85	21,872	✗	✓
VN-WN	(Kipper et al., 2006)	Ki06	Manual	N/A	N/A	N/A	40,716	✓	✓
VN-FN	(Palmer, 2009)	Pa09	Manual	N/A	N/A	N/A	17,529	✓	✓
(WN-GN)	(Vossen, 1998)	Vo98	Manual	N/A	N/A	N/A	12,079	✗	✗
(WN-FN)	(Ferrandez et al., 2010)	Fe10	Machine Learning	N/A	N/A	0.77	30,744	✗	✗
(WN-FN)	(Laparra et al., 2010)	La10	Structure	0.76	0.74	0.75	28,352	✗	✗
(OW-WP)	OmegaWiki crowd	OW	Manual	N/A	N/A	N/A	5,057	✓	✓
(OW en-OW de)	OmegaWiki crowd	OW	Manual	N/A	N/A	N/A	58,785	✓	✓
(WP en-WP de)	Wikipedia crowd	WP	Manual	N/A	N/A	N/A	463,311	✓	✓

Table 3.15: List of all available full alignments between resources in UBY. We report papers where these alignments were covered, best results achieved, the sizes of the full alignments as well as the availability in the first (0.1.0) and current (0.6.0) version of UBY. For works to which we contributed, the abbreviation is given in bold face. Pairs we did not consider for our experiments are marked with parentheses. “N/A” stands for “not available”.

	WN	VN	FN	WP en	WKT en	OW en	GN	WP de	WKT de	OW de
WN		manual Ki06	automatic (Fe10) (La10)	automatic NG11 MG13 NP12	automatic MG11 MG14	automatic Gu12 MG14	manual (Vo98)	✗	✗	automatic Gu12
VN			manual Pa09	✗	✗	✗	✗	✗	✗	✗
FN				✗	automatic HG13 MG13	✗	✗	✗	✗	✗
WP en					automatic MG14	manual (OW)	✗	manual (WP)	✗	✗
WKT en						automatic Ma13 MG14	✗	✗	✗	✗
OW en							✗	✗	✗	manual (OW)
GN								✗	automatic He11 MG13	✗
WP de									automatic MG14	manual (OW)
WKT de										✗
OW de										

Table 3.16: This matrix visualizes what alignments for resources in UBY are available and supplements Table 3.15.

Pair	Source	Pos.	Neg.	1 cand.	1:n	Ambiguity
WordNet-Wikipedia	NG11	227	1,588	54.3%	5.2%	5.7
WordNet-Wiktionary	MG11	313	2,110	18.6%	2.7	4.8
GermaNet-Wiktionary	He11	27,127	18,509	64.9%	5.6%	1.8
FrameNet-Wiktionary	HG13	775	2,014	24.9%	9.3%	3.7
VerbNet-WordNet	Ki06	39,603	8,642	4.1%	99.4%	11.1
VerbNet-FrameNet	Pa09	17,529	5,351	3.6%	99.3%	9.5
WordNet-OmegaWiki en	Gu12	210	473	75.3%	9.5%	1.4
WordNet-OmegaWiki de	Gu12	212	491	75.6%	9.6%	1.5
Wiktionary-OmegaWiki	Ma13	190	396	66.6%	3.2%	1.7
Wiktionary-Wikipedia en	MG14	75	292	87.6%	1.3%	1.3
Wiktionary-Wikipedia de	MG14	21,855	9,953	77.6%	1.1%	1.5

Table 3.17: Statistics of the gold standards from others (top) and from our own work (bottom) used in the evaluation. The degree of ambiguity (i.e. the number of possible alignment targets per sense) hints towards the difficulty of the alignment task, as do the percentages of 1:n alignments and senses with only one candidate.

Pair	Source	Sampling	Annot.	$A_0$	$F_1$
WordNet-Wikipedia	NG11	Balanced	3	0.97	0.87
WordNet-Wiktionary	MG11	Balanced	10	0.93	0.78
GermaNet-Wiktionary	He11	All	2	N/A	N/A
FrameNet-Wiktionary	HG13	Balanced	2	N/A	0.80
VerbNet-WordNet	Ki06	All	N/A	N/A	N/A
VerbNet-FrameNet	Pa09	All	N/A	N/A	N/A
WordNet-OmegaWiki	Gu12	Random	3	0.85	0.84
Wiktionary-OmegaWiki	Ma13	Random	2	0.85	0.80
Wiktionary-Wikipedia en	MG14	Automatic	2	0.79	0.95
Wiktionary-Wikipedia de	MG14	Crowdsourced	2	0.85	0.89

Table 3.18: Details about the gold standards from others (top) and from our own work (bottom) used in the evaluation, considering their creation.  $A_0$  and  $F_1$  describe the inter-rater agreements which can be considered as upper bounds for alignment accuracy and F-measure. N/A stands for “Not Available”. The “Sampling” column describes in what way the instances for the gold standard were selected. For both Wiktionary-Wikipedia datasets, two annotators manually validated a subset of 100 random alignment pairs. The agreement values between them apply accordingly.

### 3.5.1 Datasets Reported in Previous Work

**WordNet-English Wikipedia** This gold standard dataset was first reported in (Niemann and Gurevych, 2011) and was one of the first datasets created in the context of WSA within UBY. For the specific versions of the resources involved in this and the other datasets, please refer to Sections 3.2 and 3.3. This particular dataset, which only contains nouns because of Wikipedia’s restriction to this part of speech, stands out with respect to its characteristics such as ambiguity as it was

manually composed to fulfill certain restrictions. One intention, for instance, was to equally cover synsets from all levels of the WordNet hierarchy to investigate how well WSA works on synsets with different properties, although by definitions synsets from the lower levels of the hierarchy would be more common in a random sample. Thus, the distribution of synsets is not representative of the full alignment. An alternative would have been the data set presented by Navigli and Ponzetto (2012a) in the context of the creation of *BabelNet*. However, since we could not access their gold standard, we unfortunately could not apply it in our work. Nevertheless, we will consider the full alignment embedded in *BabelNet* for comparison purposes later on.

**WordNet-English Wiktionary** We use the gold standard created by Meyer and Gurevych (2011). This gold standard was created with the same restrictions in mind as the WordNet-Wikipedia gold standard mentioned above, and as both datasets were created by our group at the same time and in a coordinated effort, the set of WordNet noun synsets contained in both gold standards is identical. This does not apply to other parts of speech, obviously, as Wikipedia almost exclusively contains noun concepts (cf. Section 3.3). Here as well, an alternative would have been the WordNet-Wiktionary alignment contained in *BabelNet*. However, the evaluation data set is also not publicly available, and (unlike for the WordNet-Wikipedia case) no full alignment could be extracted from *BabelNet* due to a different internal representation which does not allow to reconstruct the original identifiers.

**GermaNet-German Wiktionary** Henrich et al. (2011) reported the only existing alignment between these two resources so far, and in our work we use their freely available dataset.<sup>12</sup> This alignment dataset is one of the largest, if not the largest, reported in previous work. It is notable though that Henrich et al. (2011) aim at enriching GermaNet with Wiktionary glosses rather than aligning the two resources, i.e. the goal was to directly integrate knowledge from Wiktionary into GermaNet. The resulting alignment can be considered a by-product of this process, which included automatically matching GermaNet lexical fields (i.e. synsets and related words) to Wiktionary glosses and then validating each alignment in a time-consuming manual effort. As they aimed at the highest possible quality of the glosses, their original dataset also contains manually corrected entries; these could not be directly used for our experimental purposes (and were filtered out) as these corrected glosses cannot be found in the the actual Wiktionary. Unfortunately, Henrich et al. (2011) do not report inter-annotator agreement, although two annotators were involved in the annotation.

**FrameNet-Wiktionary** Hartmann and Gurevych (2013) constructed this dataset, which is also the only one so far for this resource pair. The properties of this gold standard mirror the properties of the full FrameNet, e.g. the sampling preserves the distribution of POS (around 40% verbs and nouns, 12% adjectives, and the rest mostly adverbs and prepositions). Another goal was to ensure that highly polyse-

---

<sup>12</sup><http://www.sfs.uni-tuebingen.de/GermaNet/wiktionary.shtml>

mous words as well as words with few senses are selected. The final gold standard contains 2,789 sense pairs, 28% of which were annotated as aligned.

**VerbNet-WordNet / VerbNet-FrameNet** These alignment datasets were constructed by Kipper et al. (2006) and Palmer (2009), respectively, and the full alignments between VerbNet and the other resources were integrated into the first version of UBY (see Section 7.4).

### 3.5.2 Datasets Created in Our Work

**WordNet-OmegaWiki** This dataset, first reported by us in (Gurevych et al., 2012), is unique as, unlike any others that we report here, it covers two cases at once: monolingual English alignment as well as crosslingual English-German alignment. This is possible due to the multilingual nature of OmegaWiki (cf. Section 3.3), so that an alignment to a sense in one language automatically entails an alignment to the entire defined meaning, and thus alignments to all senses available in other languages. As mentioned previously, the original intention for this dataset was to cover the cross-lingual case, so that the candidate extraction included a machine translation step (explained in more detail in Section 4.4) which yielded 11,806 unique pairs with German candidates from OmegaWiki for English WordNet synsets. It should be noted that in OmegaWiki, it is quite common to have a gloss in only a few, widely spoken languages, while lexicalizations of the concept exist for many further languages; thus, we filtered out the instances lacking a German gloss. Based on the resulting candidates, we randomly selected 500 WordNet synsets for inclusion into our dataset, yielding 703 alignment candidate pairs. These were manually annotated as being (non-)alignments by three annotators fluent in both languages.

Due to the fact that English is the predominant language in OmegaWiki (cf. Section 3.3), for each sense contained in the German part of OmegaWiki there also exists a lexicalization in the English part. Thus, the gold standard is directly usable for the monolingual case. The only postprocessing step we applied was to filter out senses without an English gloss, which rendered the number of available candidate pairs slightly smaller (683 instead of 703). Note that a WordNet-OmegaWiki alignment is also available in *BabelNet*, but like for the WordNet-Wiktionary alignment a reconstruction of the original identifiers was not possible, so that this data set was disregarded.

**Wiktionary-OmegaWiki** For creating this (monolingual) gold standard, we extracted candidates for all English Wiktionary senses from OmegaWiki, which yielded 98,272 unique candidate sense pairs overall, covering 56,111 Wiktionary senses and 20,674 OmegaWiki defined meanings (that is, synsets containing one or more senses). Considering the over 400,000 word senses in Wiktionary and the over 50,000 senses in OmegaWiki, this is in line with the lexeme overlap we calculated for the two resources (Table 3.8). Note, however, that we filtered out senses which contained an empty or invalid description (less than 5 characters). This was necessary as, unlike for expert-built resources, there is no guarantee that encoded senses carry a description in the same language, especially for OmegaWiki (see Section 3.4).

We randomly selected 500 Wiktionary senses, yielding 586 alignment candidate pairs. These were manually annotated by a computational linguist as representing the same meaning (190 cases) or not (396 cases). Note that the created gold standard could probably be re-used to evaluate alignments of Wiktionary to other languages in OmegaWiki, similar to the WordNet-OmegaWiki dataset.

**Wiktionary–Wikipedia (English)** As the datasets for WordNet-Wiktionary (Meyer and Gurevych, 2011) and WordNet-Wikipedia (Niemann and Gurevych, 2011) are lexically overlapping, we were able to automatically create a gold standard for Wiktionary-Wikipedia by exploiting the transitivity of the alignment relation, i.e. by using WordNet as a pivot. This idea was first suggested by Kirschner (2012). Note that, unlike Wiktionary, WordNet synsets have multiple lexicalizations for a concept, introducing alignment candidates from Wikipedia which might not be applicable to a particular Wiktionary sense. For instance, for the WordNet synset {*car*, *automobile*} we would consider both corresponding articles in Wikipedia as candidates, while for the Wiktionary entry on *car* only the first one would be immediately applicable. Although we would still obtain some valid alignments by retaining all candidates, we decided to filter the examples where the lexeme of the Wiktionary sense and the Wikipedia article title did not match to ensure the greatest possible precision of the dataset. An effect of this process was that words not contained in all three resources were filtered out, and many examples were left with few or only one candidate, leading to a low ambiguity. Two annotators also manually checked the derived gold standard and corrected a small number of wrong annotations introduced through the automatic process. In course of this, we also calculated the inter-annotator agreement on 100 randomly selected sample pairs which we report in Table 3.18. The resulting dataset is considerably smaller than the others, but it still turned out to be sufficient for the experiments we conducted.

**Wiktionary–Wikipedia (German)** Same as for the English editions, neither a gold standard nor an alignment was previously reported for this pair. We were able to create a gold standard in a novel way by exploiting the fact that many German Wiktionary senses contain a link to the corresponding Wikipedia articles, inducing a sense alignment between the two LSRs manually validated by the Wiktionary community. However, we were unable to extract such an alignment for English (although there is also extensive linking between the two LSRs by the crowd), as Wikipedia articles are attached to the lexical entry page in this version and not to a specific sense.

In the German Wiktionary, a large portion of the senses is linked in this way, and even after aggressively filtering out possibly invalid link targets (e.g. disambiguation pages or pages with a non-matching title), we remained with over 20,000 alignments between Wiktionary senses and Wikipedia pages. A sample of 100 pairs was validated manually by two annotators, and for this task we also report the inter-annotator agreement (see Table 3.18). Of course, this only yields positive examples; to also include cases of non-alignment into the dataset, we extracted the other candidate (i.e. lexically matching) Wikipedia articles for each aligned Wiktionary sense, assuming that Wiktionary editors also considered and discarded them before eventually creating a link. Interestingly, the number of negative examples derived in this

way is relatively low in comparison to most other datasets. An analysis revealed that a considerable fraction of the linked Wiktionary senses are either scientific terms (e.g. from biology) or named entities such as cities. Both types of senses tend to have relatively few alternative candidates in Wikipedia due to their specificity, and it seems logical that Wiktionary users predominantly link these senses to the explanatory Wikipedia articles which are not familiar to the majority of users. The bias towards positive examples is also in line with the only other dataset of comparable size, namely GermaNet-Wiktionary. This suggests that, in a full alignment task, it is common to have only few (and most often, only one) alignment candidates. This being said, it appears justified to cover alignment cases from different levels of the LSR hierarchy like Meyer and Gurevych (2011) and Niemann and Gurevych (2011) to investigate and improve WSA algorithms, but this approach apparently overestimates the true difficulty of computing a full alignment as the number of candidates is higher in this case.

In the end, this process yielded a WSA dataset with unprecedented characteristics: it is an order of magnitude larger than most of the previously reported datasets (which only cover a tiny fraction of all senses), with the GermaNet-Wiktionary dataset being the sole exception (Table 3.17). While both of these large datasets enable us to assess the performance of WSA algorithms in a scenario which is close in size to a full alignment task, allowing a more well-grounded statement about their effectiveness, the Wiktionary-Wikipedia alignment was created and validated by the crowd of Wiktionary editors rather than a handful of expert annotators, making it the first crowd-sourced WSA dataset. The process of its creation is thus an interesting subject of study in itself, and largely different from the extensive manual effort that was required for GermaNet-Wiktionary.

### 3.6 Chapter Summary and Contributions

In this chapter, we discuss the guiding principles when selecting resources to be included in UBY, and we present the most important properties of all of them. We specifically discuss the collaboratively constructed, multilingual resource Omega-Wiki, which is a special focus of our own work. Moreover, we also perform a thorough comparative analysis of all LSRs with regard to their suitability for WSA. More specifically, we analyze their glosses and structures, the two most salient information sources for alignment algorithms, and find that there exist substantial differences between different LSRs.

Building on this analysis, we motivate our selection of LSR pairs we consider for WSA and subsequently present the WSA datasets we will use in the remainder of the thesis, including the ones we created ourselves.

In summary, our contributions presented in this chapter are:

**Contribution 1** We present a detailed discussion of the properties of OmegaWiki, relating it to other resources and especially to Wiktionary, which was created according to similar principles.

**Contribution 2** We perform a comparative analysis of the discussed LSRs with regard to their suitability for WSA. In particular, we analyze statistical proper-



ties of the glosses and also the graph structures induced by semantic relations, links etc.

**Contribution 3** For the first time, we present alignment datasets for the resource pairs WordNet-OmegaWiki, Wiktionary-OmegaWiki and Wiktionary-Wikipedia (for both English and German). The German Wiktionary-Wikipedia dataset is especially interesting since it was not created by expert annotators, but rather derived from the Wiktionary contributors.





reliable binary (yes-no) decisions poses unique challenges for WSA as opposed to, for instance, ontology matching or semantic relatedness calculation.

In this chapter, we would like to start by examining similarity-based approaches, as glosses are a prerequisite for humans to recognize the meaning of an encoded sense, and thus also an intuitive way of judging the similarity of senses. For tasks such as WSD, the gloss is also indispensable, as one common approach is to compare this gloss to the context of a word to be disambiguated and use this information to select the correct disambiguation (Lesk, 1986). Thus, it seems natural that the predominant approaches for WSA are also based on gloss similarity.

It should be noted, however, that we do not investigate similarity-based approaches for all of the previously discussed evaluation datasets, as these mostly have been covered in the previous work. We rather discuss selected cases which are of particular interest to us according to their properties discussed in Section 3.4. Also note that VerbNet is not covered in this chapter, as it does not have glosses. This means that the approaches presented here are not applicable.

In Section 4.2, we first discuss related work in general, and then we put a special emphasis on the work we directly build upon, before presenting our own contributions. In particular, we discuss the monolingual alignment between Wiktionary and OmegaWiki (Section 4.3) as well as the cross-lingual alignment between WordNet and OmegaWiki (Section 4.4). At the end of the chapter, in Section 4.5, we provide a summary of the contributions we made to similarity-based WSA.

## 4.2 Previous Work

In recent years, there have been many works which aimed at aligning LSRs, most of which were centered around WordNet as this is the resource predominantly used in the field. While there are a few alignments of (parts of) WordNet which have been produced manually to assure a certain level of quality, e.g. to Cyc (Reed and Lenat, 2002), Wikipedia (Mihalcea, 2007), VerbNet and FrameNet (Shi and Mihalcea, 2005), most approaches tried to automatically identify equivalent senses. Apart from a few naive approaches like the WordNet-Wikipedia alignment by Suchanek et al. (2007) which uses the most frequent sense (MFS), the majority of works used some notion of similarity between senses, mostly gloss overlap or semantic relatedness based on glosses, expressed by semantic vectors or Personalized PageRank scores (Agirre and Soroa, 2009) (see also next section).

Knight and Luk (1994) align WordNet to the *Longman Dictionary of Contemporary English* (LDOCE), as does Kwong (1998), who also considers *Roget's Thesaurus*. Burgun and Bodenreider (2001) align WordNet to the *Unified Medical Language System*, and Navigli (2006) to the *Oxford Dictionary of English* (ODE). There are numerous approaches which automatically align WordNet to Wikipedia based on similarity or gloss overlap, either to Wikipedia categories (Toral et al., 2009) or articles (Ruiz-Casado et al., 2005; De Melo and Weikum, 2010). The resulting integrated resources *WordNet++* and *Universal WordNet* are used for a considerable number of NLP applications. Also (few) resource pairs not including WordNet have been considered. For instance, Henrich et al. (2011) use a similarity measure based on word overlap for aligning GermaNet and Wiktionary, with the eventual goal of

Work	Method	Resource pair
(Reed and Lenat, 2002)	manual	WordNet-Cyc
(Shi and Mihalcea, 2005)	manual/structure	WordNet-VerbNet/FrameNet
(Mihalcea, 2007)	manual	WordNet-Wikipedia
(Suchanek et al., 2007)	MFS	WordNet-Wikipedia
(Knight and Luk, 1994)	overlap	WordNet-LDOCE
(Kwong, 1998)	overlap	WordNet-LDOCE/Roget
(Burgun and Bodenreider, 2001)	overlap	WordNet-UMLS
(Ruiz-Casado et al., 2005)	overlap	WordNet-Wikipedia
(De Melo and Weikum, 2010)	overlap	WordNet-Wikipedia
(Henrich et al., 2011)	overlap	GermaNet-Wiktionary
(Navigli, 2006)	relatedness	WordNet-ODE
(Toral et al., 2009)	relatedness	WordNet-Wikipedia
(Meyer and Gurevych, 2011)	relatedness	WordNet-Wiktionary
(Niemann and Gurevych, 2011)	relatedness	WordNet-Wikipedia
(Hartmann and Gurevych, 2013)	relatedness	FrameNet-Wiktionary

Table 4.1: Previous work on aligning LSRs manually (top), using gloss overlap (middle) or some other notion of semantic relatedness (bottom).

enriching GermaNet with more glosses (cf. Section 3.5.1).

Many efforts for aligning WordNet with other LSRs based on text similarity have been undertaken at our research group – we will discuss them in more detail in the following, as these are the works we base our own contributions on. Table 4.1 gives an overview of the related work we discuss in this chapter. In general, all of these approaches give reasonable results (with precision in the range of 0.67-0.84), but as each word sense alignment approach has been evaluated on a separate, manually annotated dataset with different characteristics (cf. Section 3.5), these numbers cannot be directly compared to each other.

As already mentioned in Section 3.5.1, Niemann and Gurevych (2011) and Meyer and Gurevych (2011) created WordNet-Wikipedia and WordNet-Wiktionary alignments. For this, they designed and implemented an alignment algorithm which forms the foundation of our work and which we will now outline in more detail. Their approach supports WSA for a large number of resources across languages and allows alignments between different representations of senses as found in different resources, for example WordNet synsets, Wikipedia articles or FrameNet frames, as demonstrated by Hartmann and Gurevych (2013) who align FrameNet and Wiktionary. The only requirement is that the individual sense representations are distinguishable by a unique identifier in each resource.

The basic idea of the algorithm is, in a nutshell:

1. For each sense  $s$  in one resource, all possible candidates  $t_1 \dots t_n$  in the other resource are retrieved, i.e. all senses which have the same attached lexeme  $l$  as  $s$  (cf. our definition of candidates in Section 2.2).
2. One or several similarity scores between the sense descriptions of the candidate pairs are calculated, i.e. for each sense pair  $(s, t)$  we calculate similarity values

$sim_1(s, t) \dots sim_n(s, t)$ . Note that there are different options for constructing the description of a sense, such as the gloss, example sentences etc. The exact choice is dependent on the information available in the LSRs considered for alignment. While, intuitively, it seems advisable to include as much knowledge as possible, we refrain from including information types which are present in one resource but not in the other, in order to avoid having descriptions of substantially different length and content. This would make the calculation of similarity values less informative.

3. For a subset of the candidate pairs, the decision “alignment” or “non-alignment” is manually annotated, creating a gold standard (cf. Sections 3.5.1 and 3.5.2).
4. Based on the annotations and similarity scores of the gold standard, a machine learning classifier learns optimal similarity thresholds  $\theta_{sim_1} \dots \theta_{sim_n}$ , i.e. similarity values which a candidate pair needs to exceed to be considered correct. This value is optimized with regard to F-measure in a 10-fold cross validation setup. Note that this alignment framework explicitly allows training on all measures in combination, which means that a pair needs to exceed all learned thresholds in order to be aligned. Formally, a pair  $(s, t)$  is aligned if  $sim_1(s, t) > \theta_{sim_1} \wedge sim_2(s, t) > \theta_{sim_2} \dots \wedge sim_n(s, t) > \theta_{sim_n}$ .
5. Using these learned thresholds, the alignment decision is made for all candidates to produce a complete alignment of the resources.

In our case, the two similarity measures described in the following paragraphs are used, since they were reported to produce meaningful results in previous work. The approach is, however, open to extension, so that in future work additional measures for text similarity (as, for instance, reported by Bär et al. (2013)) can be integrated.

**Cosine similarity** (COS) calculates the cosine of the angle between vector representations of the two senses  $s_1$  and  $s_2$ :

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

To represent a sense as a vector, we use a bag-of-words approach – that is, a vector  $\text{BoW}(s)$  contains the term frequencies of all words in the description of  $s$ .

**Personalized PageRank** (PPR) (Agirre and Soroa, 2009) estimates the semantic relatedness between two word senses  $s_1$  and  $s_2$  by representing them in a semantic graph (derived from a reference LSR such as WordNet) and comparing the semantic vectors  $\mathbf{Pr}_{s_1}$  and  $\mathbf{Pr}_{s_2}$  by computing

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

which is a  $\chi^2$  variant introduced by Niemann and Gurevych (2011). The main idea of choosing  $\mathbf{Pr}$  is to use the personalized PageRank algorithm for identifying those nodes in the graph that are central for describing a sense’s meaning. These nodes

should have a high centrality (that is, a high PageRank score), which is calculated as

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

with the damping factor  $c$  controlling the random walk, the transition matrix  $M$  of the underlying semantic graph, and the probabilistic vector  $\mathbf{v}$ , whose  $i^{\text{th}}$  component  $\mathbf{v}_i$  denotes the probability of randomly jumping to node  $i$  in the next iteration step.<sup>1</sup> Unlike in the traditional PageRank algorithm, the components of the jump vector  $\mathbf{v}$  are not uniformly distributed, but personalized to the sense  $s$  by choosing  $\mathbf{v}_i = \frac{1}{m}$  if at least one lexicalization of node  $i$  occurs in the definition of sense  $s$ , and  $\mathbf{v}_i = 0$  otherwise. The normalization factor  $m$  is set to the total number of nodes that share a word with the sense descriptions, which is required for obtaining a probabilistic vector.

### 4.3 Monolingual Alignment between Wiktionary and OmegaWiki

One of our first contributions to the field of WSA is the alignment of the two collaboratively constructed resources Wiktionary and OmegaWiki. As mentioned previously, apart from the perennial goal of establishing as many alignments as possible between the resources contained in UBY, we had two special goals in mind when creating this alignment:

- For the first time, an alignment between two collaboratively constructed resources is proposed. While in past efforts at least one expert-built resource with high quality content was involved, we want to investigate how the lower quality of information (in this case, glosses of different length and lexical variety, cf. Section 3.4.1) affects the results of the alignment process. To our knowledge, this is also the first study on aligning OmegaWiki with another LSR.
- As both resources are inherently multilingual, we wanted to align them with the intention of creating a resource which could be exploited in translation applications. While Wiktionary offers a greater coverage and a richer variety of encoded information (see Section 3.3), OmegaWiki provides the advantage of unambiguous translations and relations which are potentially useful in such translation applications. We will give more background information and motivation for the exploitation of these aligned resources for translation applications in Section 8.2.

Here, we focus on aligning the English Wiktionary with the English part of OmegaWiki. As English is the language with the most entries in both resources,

---

<sup>1</sup>The publicly available UKB software (Agirre and Soroa, 2009) is used for calculating the PageRank scores with the WordNet 3.0 graph augmented with the Princeton WordNet Gloss Corpus (<http://wordnet.princeton.edu/glosstag.shtml>) as basis for the transition matrix  $M$ . The damping factor  $c$  is set to 0.85.

such an alignment is bound to yield the largest number of links between the two LSRs and thus the greatest benefit.

Moreover, as OmegaWiki defined meanings are multilingual by design, if Wiktionary (or any other resource) is aligned to a defined meaning  $d$  in OmegaWiki, we trivially obtain an alignment to all senses in all languages which are contained in  $d$ , e.g. we obtain an alignment between the English Wiktionary and the German part of OmegaWiki for all defined meanings for which an English as well as a German lexicalization exists. Apart from the advantage of directly available, unambiguous translations of Wiktionary senses (see Section 8.2 for more details), we also benefit from the fact that the manually annotated gold standard created for this task is directly usable for cross-lingual alignment experiments in the future (cf. Section 4.4).

### 4.3.1 Alignment Procedure

Using the alignment framework described above, we first extract OmegaWiki defined meaning candidates for each entry in the English Wiktionary. This is solely based on the combination of lemma and part-of-speech as explained in Section 2.2, and we manually annotate a subset of candidate pairs as “alignment” or “non-alignment”. Then, we extract the sense descriptions to compute the similarity of word senses with the two similarity measures COS and PPR. Note that we only focus on glosses, as other information such as sense examples is only present for few OmegaWiki senses (see Section 3.3). Including these for Wiktionary would likely lead to a length bias in the data.

As described in Section 4.2, a 10-fold cross-validation setup was used for training and evaluating the threshold-based machine learning classifier. Note that, as suggested in the earlier work, the threshold was optimized for F-measure; optimizing for precision would have led to a higher threshold and thus fewer alignments.

### 4.3.2 Evaluation

Table 4.2 summarizes the results for the two different similarity measures and their combination, i.e. the learning of separate similarity thresholds for both measures which both have to be exceeded for an alignment to be considered correct. The results of the baselines (cf. Section 2.2.2) are given for comparison. As there is no explicit sense frequency information encoded in either resource, the application of a most frequent sense baseline was not possible.

We observe that the more elaborate similarity measure PPR yields worse results than the cosine similarity (COS), while the best result is achieved by combining both. However, this difference between COS and the combination of COS and PPR is not statistically significant<sup>2</sup>. All measures outperform the baselines by a large margin.

These results differ from those reported in earlier work (Niemann and Gurevych, 2011; Meyer and Gurevych, 2011) which state that the more semantically oriented

---

<sup>2</sup>All significance statements in this thesis are based on McNemar’s test at a confidence level of 5%, unless otherwise stated.



Similarity measure	$P$	$R$	$F_1$	$A$
<i>Random</i>	0.35	0.40	0.37	0.57
<i>1st</i>	0.41	0.77	0.54	0.57
<i>1:1</i>	0.49	0.61	0.54	0.67
COS	0.64	0.67	<b>0.66</b>	0.77
PPR	0.43	<b>0.90</b>	0.58	0.58
PPR and COS	<b>0.67</b>	0.65	<b>0.66</b>	<b>0.78</b>
Agreement	-	-	0.80	0.85

Table 4.2: Alignment results for the Wiktionary-OmegaWiki alignment. The best result per evaluation measure is marked in bold.

PPR usually gives better results than the rather naive COS. We hypothesize, however, that this can (at least partly) be attributed to the fact that the PPR distance as it is defined and implemented in the UKB package (cf. Section 4.2) is based on the WordNet graph. As the previous alignments reported by Meyer and Gurevych (2011) and Niemann and Gurevych (2011) both focused on WordNet, it is reasonable to assume that the WordNet synsets and their similarities to other senses are accurately represented by the PPR measure, while Wiktionary and OmegaWiki suffer from the fact that their structure and coverage of senses are not appropriately reflected in the WordNet graph. Another reason for the relatively strong performance of the COS measure is that, according to our observation, a substantial number of sense definitions in OmegaWiki have been copied or adapted from Wiktionary. This seems natural, as both resources allow free editing and copying of content, and OmegaWiki was created at a later time by a group reasonably familiar with the content of Wiktionary (cf. Section 3.3). Due to the resulting sense descriptions which are very similar in their wording, cosine similarity alone already gives a strong hint towards the correct sense.

The F-measure of 0.66 in the best configuration is in line with the result that was reported in (Meyer and Gurevych, 2011) (0.66) for the alignment between Wiktionary and WordNet. While, due to the different resources and datasets involved, these results are never fully comparable, this suggests that the similarity-based approach works comparably well for the two collaboratively constructed resources considered here. Other than expected, the quality of the glosses does not affect the results in a negative way. The fact that the resources are (at least to some extent) overlapping in their sense descriptions (see also Section 3.8) helps to achieve state-of-the-art alignment results.

The application of the trained classifier to all candidate pairs led to a final alignment of 25,727 senses between Wiktionary and OmegaWiki. This alignment, like the other alignments we created in the course of this work, is freely available from our website.<sup>3</sup>

<sup>3</sup><https://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary-omegawiki-alignment/>

### 4.3.3 Error Analysis

We carried out an error analysis to identify the main errors made by the similarity-based alignment algorithm. Of the 586 sense pairs in the gold standard, the classifier yields 61 false positives (i.e. incorrectly aligned senses) and 66 false negatives (senses which should be aligned but are not).

For the false positives, the main error source is the same as already identified in the previous work (Niemann and Gurevych, 2011; Meyer and Gurevych, 2011). Different senses are aligned because of very similar sense descriptions expressing only a slight difference which is hard to distinguish for such a gloss-based approach. An example for this are two senses of *to carry*: (1) “To lift (something) and take it to another place; to transport (something) by lifting” (2) “To transport with the flow” which are undoubtedly highly related, but not equivalent.

For the false negatives, we could identify two major categories of errors, which are also in line with the observations made in previous work:

1. Different sense descriptions for the same concept. This phenomenon, known as the “lexical gap”, is not easy to tackle as a certain degree of inference and world knowledge would be required. An example for this are two senses of the adjective *aware* which are not aligned because of insufficient overlap: (1) “conscious or having knowledge of something” (2) “noticing something; aware of something”. The COS similarity is obviously affected most by this, as it does not consider semantics so that low lexical overlap inevitably leads to low similarity. PPR is supposed to alleviate this problem at least to some extent. However, apart from the issues with the implementation of PPR based on WordNet which impair its effectiveness (see previous section), it stands to reason that more sophisticated (semantic) similarity measures or lexical expansion approaches would be required to recognize such equivalent senses using gloss similarities as the only information source.
2. Short definitions making references to other, closely related or derived words. An example are these two definitions of *alluvial*: (1) “Pertaining to the soil deposited by a stream” (2) “Of or relating to alluvium”. Without making the connection between *alluvium* and the derived word *alluvial*, a disambiguation is nearly impossible. Character-level or derivational similarity measures might be considered as an additional source of information here to discover the similarity of the two lexemes, however, these are prone to introduce errors for words which are lexically similar but unrelated. Another angle to tackle this kind of error is the exploitation of the underlying structure of the resources, making use of the fact that the two terms are closely related. This motivates the work on graph-based approaches we report later on.

## 4.4 Cross-lingual and Monolingual alignment of WordNet and OmegaWiki

After introducing OmegaWiki into the field of WSA by aligning it to Wiktionary, the next natural step for the integration into UBY is its alignment to WordNet.

WordNet is the most densely linked resource within UBY due to its importance for NLP and the large number of previously created alignments, so that aligning OmegaWiki to it is an important cornerstone for a large-scale linked resource. Moreover, WordNet and OmegaWiki show (according to our analysis in Section 3.4) the greatest differences regarding their content and structure between any two resources we investigate, which makes this case especially interesting for the evaluation of WSA approaches.

Primarily, we address the issue of cross-lingual alignment, i.e. alignment between resources in different languages. While monolingual alignment has been thoroughly investigated in previous work, this is (to our knowledge) the first work to investigate cross-lingual alignment. It is especially important since UBY is designed as a multilingual resource, subsuming different monolingual LSRs. Thus, it is vital to investigate alignments between them as well. On the one hand, this enables the greatest possible improvement by enhanced sense representations and better coverage, not only for monolingual, but also for cross-lingual application scenarios such as cross-lingual semantic relatedness. On the other hand, it showcases what difficulties can arise for future cross-lingual alignment efforts and how these can be addressed.

#### 4.4.1 Alignment Procedure

First, we explored the most straightforward idea of using the existing WordNet-Wiktionary and Wiktionary-OmegaWiki alignments to directly infer an alignment between WordNet and OmegaWiki exploiting the transitivity of the equivalence relations, i.e. using Wiktionary as a pivot. Due to the multilingual nature of OmegaWiki, this would have yielded an alignment between WordNet and other languages contained in OmegaWiki as well. However, the different sense granularities in combination with small lexical overlap of all three resources (see Section 3.3) rendered this approach very ineffective – after all, the intention was to create a full alignment between WordNet and OmegaWiki without the additional constraint that the lexemes should also be contained in Wiktionary. While the resulting full alignment would have been of at least acceptable size, the set of gold standard examples for evaluation was rendered very small, containing only a few dozen pairs, and there was also a considerable number of errors introduced through this automatic process due to error propagation. Roughly speaking, if the two original alignments have a precision of 0.67 and 0.78, respectively, we can reasonably assume that only around half of the derived alignments are correct ( $0.67 \times 0.78 = 0.52$ ). We made similar observations for the automatic creation of a gold standard for the alignment of Wiktionary to Wikipedia (see Section 3.5.2), but in the latter case the gold standard was still large enough after filtering and manual correction to efficiently use it for experimentation.

In conclusion, it seemed insufficient for our purposes to use such a small dataset with questionable quality. Moreover, the intention of our experiments is also to cover the general case of two LSRs in different languages which have no existing alignments to a common third resource. Hence, in order to align word senses across languages, we extend the monolingual sense alignment described above to the cross-lingual set-

ting by using a machine translation component. For this, we utilize Moses<sup>4</sup>, a freely available machine translation framework, for our purposes trained on the Europarl corpus (Koehn, 2005). While the usage of commercial translation services such as Google Translate or Bing Translator would have been possible, we refrained from this option to ensure the reproducibility of our results as well as the sustainability of our alignment framework, making it independent of components we cannot control.

Furthermore, the motivation for our approach is to rely on the established strengths of the existing alignment approach and to minimize the introduction of additional errors. Thus, for a cross-lingual setting, the lemma (or lemmas, in case of a synset representation) of a sense to be aligned as well as its gloss are translated into the language of the other resource, again yielding a monolingual setting which can be handled by the existing algorithm. For instance, the WordNet synset  $\{vessel, watercraft\}$  with its gloss “a craft designed for water transportation” can be translated into the German  $\{Schiff, Wasserfahrzeug\}$  and “Ein Fahrzeug für Wassertransport”, and then the candidate extraction and all downstream steps can take place in German.

For evaluating our approach, we create a cross-lingual alignment between WordNet and the German part of OmegaWiki, i.e. the concepts in OmegaWiki with a German lexicalization. This was motivated by the fact that it would not only allow an easy manual error analysis, but also a direct comparison to the monolingual setup by using the English sense descriptions contained in OmegaWiki in an otherwise identical configuration. The resulting gold standard dataset is described in Section 3.5.2.

Utilizing the established alignment framework in conjunction with the machine translation component enabled us to cover both translation directions. We used the COS similarity for comparing the German OmegaWiki sense descriptions with the German translations of WordNet sense descriptions, and COS and PPR similarity for comparison of the German OmegaWiki sense descriptions translated into English with the original English WordNet sense descriptions. In order to provide the machine translation component with as much context as possible (a prerequisite for finding an appropriate translation), we also included example sentences and synonyms in the same language into the sense descriptions along with the glosses, while for the actual similarity calculation only the glosses were used as described above. Note that PPR similarity is not available for German in the original implementation as it is based on WordNet; adapting PPR to a German resource such as GermaNet would be a promising direction for future work. A similar idea for other English resources was recently presented by Pilehvar and Navigli (2014).

#### 4.4.2 Evaluation and Error Analysis

For the machine learning task, we again used the threshold-based classifier and 10-fold cross validation. The results for different translation directions and similarity measures are given in Table 4.3. Note that we also report the results for the monolingual alignment, i.e. for directly using the English sense description in OmegaWiki instead of using the translations of the German ones.

First of all, we observe that using cosine similarity alone, neither the translation

---

<sup>4</sup><http://www.statmt.org/moses/>

Translation direction	Similarity measure	$P$	$R$	$F_1$	$A$
<i>Random</i>		0.46	0.35	0.40	0.51
<i>1:1</i>		0.36	0.64	0.46	0.55
<i>1st</i>		0.34	<b>0.80</b>	0.48	0.47
en > de	COS	0.37	0.65	0.47	0.58
de > en	COS	0.39	0.68	0.49	0.58
de > en	PPR	0.51	0.56	0.54	0.71
de > en	PPR and COS	0.52	0.56	<b>0.55</b>	0.72
en > en	PPR and COS	<b>0.55</b>	0.53	0.54	<b>0.73</b>
Agreement				0.84	0.85

Table 4.3: Cross-lingual alignment results for WordNet and the German part of OmegaWiki. The best monolingual results, naive baselines and the annotator agreements are given for reference. The best result per evaluation measure is marked in bold.

from English to German nor vice versa works very well; while the recall is good, the precision is very low, even failing to beat the random baseline. Judging from a manual analysis, this seems to stem from the machine translation component. In many cases, it provides a generally acceptable translation (if we disregard grammar) which, on the other hand, does not lexically match the WordNet sense description. This is an especially grave issue for OmegaWiki. In comparison to WordNet, only very few senses are accompanied by example sentences (see Table 3.5), and as dictionary glosses are generally shorter than documents which are usually handled by machine translation systems (with OmegaWiki glosses being among the shortest on average, see Table 3.11), we face two problems: i) the translation algorithm lacks the proper context for finding an exact translation, and ii) even if an appropriate translation is found, there are still only few non-stopwords (i.e. words contributing to the meaning of a text) for the cosine similarity to work with. If these are not spot-on, i.e. lexically matching, the resulting cosine similarity value is very low. A typical example are the two descriptions for *childless*, “without offspring” and “Keine Kinder habend”, where the latter is translated to “having no children”: the translation is correct, but there is no lexical overlap. While this issue should intuitively lead to lower recall as fewer positive examples are aligned, the strong tendency across positive examples to have very low similarity causes the machine learning classifier to set a low threshold, which in turn causes many negative examples to be incorrectly aligned; hence the low precision.

This problem is, at least to some extent, alleviated by using PPR similarity, as it more accurately captures the semantic similarity between words in the sense descriptions. For instance, in the example mentioned above, the similarity between *child* and *offspring* is recognized, which leads to an alignment. The best result is achieved by a combination of both measures, although the improvement is not statistically significant. In this scenario, it is seemingly beneficial to use PPR as it is based on WordNet, so that more meaningful similarity values can be computed. For the Wiktionary-OmegaWiki alignment discussed in Section 4.3, this was not the

case as both resources did not conform to the structure implicitly expected by the PPR algorithm.

As the resulting F-measure is substantially lower than the values reported in the previous work ((Meyer and Gurevych, 2011), 0.66 and (Niemann and Gurevych, 2011), 0.78), it seems crucial to investigate if this is really due to machine translation as we hypothesize above. Thus, we repeat the experiments, but this time using the English sense description originally contained in OmegaWiki, i.e. we eliminate the errors introduced through the translation. Surprisingly, the results are almost on par – the precision is slightly higher, at a small expense of recall. This parity of results would suggest that the machine translation approach is appropriate after all, in the sense that it achieves the same performance as the monolingual setup. Examination of the English OmegaWiki glosses indeed revealed that the translations from German are, in many cases, rather accurate: for the example above, “having no children” is the actual English gloss.

Thus, we can conclude that, while the machine translation gives acceptable results, the actual difficulty when aligning WordNet and OmegaWiki in this setup are the quite differently worded sense descriptions, i.e. the lexical gap. This is in line with the relatively low lexical overlap of glosses between the two resources which we discussed in Section 3.4.1 and which was one motivation for selecting this resource pair for experimentation. This was not the case for the earlier reported alignment between Wiktionary and OmegaWiki, as many glosses are quite similar (see Section 4.3). Although the datasets are not fully comparable, it seems noteworthy that the results for the WordNet-Wiktionary alignment by Meyer and Gurevych (2011) are also relatively low in comparison to the results on Wikipedia (Niemann and Gurevych, 2011), for which WordNet glosses have the highest lexical overlap. Meyer and Gurevych (2011) also report that their results are impaired by lexical mismatch, but unlike OmegaWiki, Wiktionary at least contains a significant number of example sentences which can be included into the sense description, helping to calculate meaningful similarity values. Moreover, they also report that the PPR similarity yields significantly better precision than the cosine similarity (0.659 vs. 0.646 F-measure). From our observations, we hypothesize that lexical mismatch of the glosses is a decisive factor in the calculation of similarity-based alignments, and that a consideration of this is crucial when choosing an alignment setup. Graph-based algorithms have the potential to alleviate this issue, and we will discuss this in the forthcoming chapters.

## 4.5 Chapter Summary and Contributions

In this chapter, we apply a text similarity-based approach to the problem of aligning Wiktionary and OmegaWiki, achieving results which are in line with the previous work. The resulting alignment is the first of this kind between two collaboratively constructed resources, thus proving that this approach is also applicable in cases where the sense descriptions are not curated by experts. However, the results are worse than for previously reported datasets since they depend on the quality of the textual descriptions, which is usually lower in collaboratively constructed LSRs. As mentioned in the introduction, one of the original motivations for this alignment

was the creation of an integrated, multilingual resource to support multilingual applications. We will discuss this in Section 8.2, along with explaining some of the obtained alignments in more detail.

We also enhance the similarity-based alignment approach by using a machine translation component, for the first time explicitly addressing the issue of cross-lingual WSA. While cross-lingual approaches have been investigated for related tasks such as WSD (Navigli and Ponzetto, 2012d) and semantic relatedness computation (Navigli and Ponzetto, 2012c) (see also Section 2.4), the introduction of machine translation into the alignment process opens up new possibilities for the investigation of new alignment pairs (such as WordNet-GermaNet) as well as algorithms in the future.

WordNet and the German part of OmegaWiki serve as our testbed, and while we fail to achieve results which are comparable to previous work, additional experiments with English OmegaWiki glosses reveal that the main issue are short and differently worded glosses, not errors introduced by the translation component. Thus, we believe that our approach is viable, but as for the monolingual setup, longer sense descriptions (e.g. including usage examples) are immensely helpful, and the usage of elaborate similarity measures such as PPR is advisable to bridge the lexical gap. In the context of cross-lingual alignment, this implies that, if possible, the translation should be made into a language where such a measure is available (in this case, English) or that the measure needs to be adapted for other languages. Generally, PPR operates on all languages, but an underlying resource graph corresponding to WordNet is required, as it has been suggested by Pilehvar and Navigli (2014).

The resulting alignment is, to our knowledge, the first automatically created cross-lingual alignment and we provide it as part of UBY<sup>5</sup> (Section 7.4). Since our goal is to generally investigate and enable an alignment between German and English resources in UBY, our future work will include the investigation of other resource pairs and improvements in the alignment setup.

In summary, the contributions presented in this chapter are:

**Contribution 1** We adapted an existing text similarity-based approach to create a word sense alignment between two collaboratively constructed resources, namely Wiktionary and OmegaWiki, with alignment results comparable to previous work. This proves that also for glosses which have not been curated by experts such an approach is viable.

**Contribution 2** We extend the aforementioned algorithm by including a machine translation component, and we create a cross-lingual alignment between WordNet and the German part of OmegaWiki. We also create an alignment between WordNet and the English part of OmegaWiki, and as both setups yield quite similar results, we consider this as evidence that the cross-lingual alignment approach performs sufficiently well. Nevertheless, the relatively low results in comparison to other resource pairs suggest that a low vocabulary overlap of glosses (as it is the case for WordNet and OmegaWiki) is a severe issue, which needs to be considered when investigating other alignments in the future.

---

<sup>5</sup><https://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/>





# Chapter 5

## Graph-based Word Sense Alignment

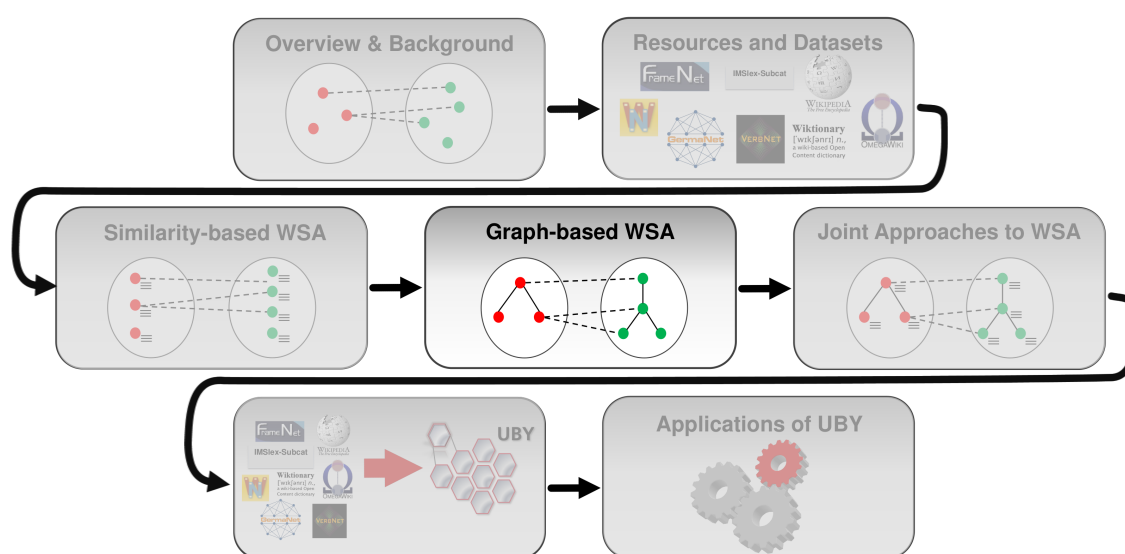


Figure 5.1: Visual outline of the thesis.

### 5.1 Introduction

As we have seen in the previous chapter, alignment based on gloss similarity is an intuitively valid approach which, in general, gives reasonable results, with an F-measure significantly outperforming naive baselines. Nevertheless, the error analyses have also shown that it suffers from the inherent problem of low recall if there is insufficient lexical overlap (also known as the “lexical gap”), i.e. if the glosses do not match lexically.

This apparent difference between human judgement and similarity-based approaches for judging the equivalence of senses motivates the investigation of similarity measures which do not rely on the glosses, but the structure of the resources. This development is intimately intertwined with the recent development of electronic and especially machine-readable language resources which allow automatic analysis and exploitation of this structure. In classic dictionaries, there are also references

to related words, synonyms etc., but these references are either not disambiguated at all, or additional look-up effort is required to make the connection between entries (cf. (Engelberg and Lemnitzer, 2001), Chapter 4.4).

The situation is different for more recent resources such as WordNet, which unambiguously connects synsets via semantic relations (e.g. hyponymy), or Wikipedia, which contains an abundant number of hyperlinks between distinct articles. In both cases, a graph structure (with concepts as nodes) can easily be derived, and it is intuitively clear that directly connected concepts must be somehow related, while in general concepts which are closer to each other have a higher likelihood of belonging to the same topic. This intuition, was for instance, confirmed in the context of the semantic relatedness task (Rada et al., 1989; Zesch et al., 2008; Navigli and Ponzetto, 2012c), where the length of paths in the graph representation of a resource is a good indicative feature. Following this idea, considering the structure of LSRs for WSA is a promising approach to alleviate the shortcomings of the similarity-based approaches. In this case, the approach does not depend on the properties of the glosses. Rather it depends on the underlying graph structures, which we already analyzed in Section 3.4.2. We found that, especially for expert-built resources and Wikipedia, a dense graph can be obtained which covers the majority of senses, while the graphs for the other collaboratively constructed resources have not only fewer edges in general, but also many isolated components in the graph. We investigate in this chapter what influence these differences have on WSA.

A general issue with these graphs is that, although relations can carry certain explicit semantics (such as hyponymy), this is not always the case (e.g. for Wikipedia links), and different resources also express different notions of relatedness (such as frame relations in FrameNet) which are not straightforwardly applicable to other resources. Thus, the common denominator for investigating algorithms based on the structure of multiple LSRs is to treat the edges as unlabeled, which is the rationale applied throughout this thesis. We deem this the most reasonable way to keep the developed approaches as flexible and generally usable as possible, although we are aware that this limits the potential benefit from exploiting the structure – reasoning over relations is, for instance, used for many ontology matching approaches (see Section 2.3), and also for WSA many approaches have been suggested which make explicit use of the semantics encoded in the relations (see next section). We refrain from this, though, in order to take a broader perspective on the issue of graph-based WSA, hopefully gaining insights which are useful not only for the resources and datasets presented in this chapter, but also for future WSA efforts on other LSRs. This aspect has been largely disregarded in previous work.

In section 5.2, we discuss related work on graph-based approaches, before presenting our novel alignment algorithm Dijkstra-WSA in Section 5.3. We discuss its different parameters, and present an evaluation and error analysis on the evaluation datasets described in Section 3.5. We summarize our findings and contributions in Section 5.4.

## 5.2 Previous Work

Toral et al. (2008) align WordNet synsets to Wikipedia categories in order to enrich

WordNet with more information about named entities present in Wikipedia. To this end, they compare and match the WordNet graph to the Wikipedia category graph. They reach very good results, with an F-measure of up to 0.82 depending on the configuration. Their approach is however not applicable to the general case of WSA for several reasons. First, they only focus on “instance of” relations in WordNet, disregarding concepts which do not have instantiations. This is valid in the scope of their paper, but not sufficient in general and especially not for other parts of speech since, for instance, verbs cannot be instantiated. Second, the Wikipedia category graph is easier to handle than the full Wikipedia link graph, as the latter is much larger and also less restricted because it can contain links between arbitrary articles.

Ponzetto and Navigli (2009) also propose a graph-based method to tackle the problem of aligning WordNet synsets and Wikipedia categories, with the purpose of restructuring the Wikipedia category graph in a subsequent step. Using semantic relations, they build WordNet subgraphs for each Wikipedia category and then align those synsets which best match according to these subgraphs, reaching an accuracy of 0.81. Like Toral et al. (2008), they only focus on a particular kind of semantic relations in WordNet (“is a” relations) in order to cover their specific application scenario, which is not applicable to parts of speech other than nouns. Moreover, they also focus only on the category graph, not the full Wikipedia graph, and even the (potentially useful) information in the category graph is disregarded in the alignment step as only the WordNet taxonomy is used as an information source.

Laparra et al. (2010) utilize the SSI-Dijkstra+ algorithm, which is based on calculating shortest paths, to align lexical units (LUs, the FrameNet equivalent to senses) with WordNet synsets and create the combined resource *WordFrameNet*. The basic idea is to align monosemous LUs first and, based on this, find the closest synset in the WordNet graph for the other LUs in the same frame. They reach a result of 0.79 (F-measure), however, they make some assumptions which apply only to their particular case. For instance, the algorithm not only relies on the semantic relations found in WordNet, but also from the enriched *eXtended WordNet* (Mihalcea and Moldovan, 2001b) in order to find a sufficient number of targets. Thus, it is not straightforwardly applicable to other resources which have no or only few relations such as Wiktionary and for which no such high-quality extensions exist. Moreover, for the case that no monosemous LU exists in a frame, they align to the most frequent sense. This information is not available in most other resources (cf. Chapter 3). The issue of missing monosemous “anchors” into WordNet could be tackled by also considering LUs from other frames connected via frame relations, i.e. exploiting the global graph structure for FrameNet as we do for the Dijkstra-WSA algorithm presented in this chapter. However, as SSI-Dijkstra+ is originally a word sense disambiguation (not alignment) algorithm, it disregards this structure and merely considers LUs as texts which are to be disambiguated in isolation. In other words, only the “local” information for each LU is used.

Navigli (2009a) aims at disambiguating WordNet glosses, i.e. assigning the correct senses to all non-stopwords of each WordNet gloss. His approach is to find the shortest possible circles in the WordNet relation graph to identify the correct disambiguation and reaches an F-measure of 0.64. This can be considered as “resource-internal” WSD and is a very useful idea for enriching WordNet and making the graph more dense. For Dijkstra-WSA, we build upon this idea of the finding short-

Work	Resource (pair)
(Toral et al., 2008)	WordNet-Wikipedia categories
(Ponzetto and Navigli, 2009)	WordNet-Wikipedia categories
(Laparra et al., 2010)	WordNet-FrameNet
(Navigli, 2009a)	Disambiguation of WordNet glosses
(Flati and Navigli, 2012)	RBEID (English part-Italian part)

Table 5.1: Previous work on WSA using the structure of LSRs.

est paths (circles are a special kind of path), but we extend it to the case of multiple resources and generalize it to edges other than semantic relations, such as Wikipedia links. This makes our task inherently more difficult, as not all LSRs have a densely connected graph like WordNet, so that paths or circles are not as easily exploitable. To alleviate this issue of sparse graphs, we also follow the idea of Navigli (2009a) by disambiguating senses in glosses as a preparatory step for our algorithm. We only link to monosemous senses though, making our approach (presented in Section 5.3.1) a streamlined version of the idea presented here. This makes it applicable to all kinds of LSRs having a gloss, regardless of their structure.

In later work, the idea of building resource-internal cycles was extended to the disambiguation of translations in the *Ragazzini-Biagi English-Italian bilingual dictionary* (RBEID) (Flati and Navigli, 2012), which is very close to the WSA of two LSRs which we are aiming at, as the English and Italian part of the dictionary could be considered separate LSRs. They reach an F-measure of 0.85, nevertheless, the algorithm again benefits from the circumstances of the task: i) the English and Italian parts of the RBEID have comparably dense graph structures, which is not always given in the general case, and ii) as the English and Italian entries were created in a coordinated effort, we can assume that most senses in one part are represented in the other (i.e. we have very high conceptual overlap), and also that the sense granularities are similar. Both of these properties make the task substantially easier than aligning two heterogeneous LSRs. Additionally, for this cross-lingual case, the identification of correct alignment candidates is usually an issue in itself, which we tackle by using a machine translation component (see Section 4.4). Here, the alignment candidates are already given by the list of translations for an entry.

A summary of the related work in this area is given in Table 5.1.

### 5.3 Dijkstra-WSA

As we have seen in the previous section, most existing approaches to graph-based WSA, while helping to address the shortcomings of similarity-based approaches discussed in Chapter 4, do not exploit all structural information available in both resources, and they also rely on certain resource-specific assumptions about the structure which make the alignment task less difficult.

To cover the general case of aligning arbitrary, heterogeneous LSRs, we propose Dijkstra-WSA, a novel, robust algorithm for word sense alignment which is designed to be applicable to a wide variety of resource pairs and languages. It is, to our knowledge, the first attempt to apply a graph-based algorithm to full graph

representations of two arbitrary resources, i.e. using all structural information contained in both participating LSRs rather than just sub-graphs. This allows taking a more abstract perspective and reducing the problem of identifying equivalent senses to the problem of matching nodes in these graphs. To investigate the effectiveness of this approach, we comparatively evaluate this WSA algorithm on the variety of different datasets with different characteristics we presented earlier.

The key properties of Dijkstra-WSA are:

**Robustness** The entities within the LSRs which are to be aligned (usually senses or synsets) are modeled as nodes in a graph. These nodes are connected by an edge if they are semantically related. While, as mentioned earlier, semantic relations lend themselves very well to deriving edges, different possibilities for graph construction are equally valid as the algorithm is agnostic to the origin of the edges. Hence, it can directly be applied to a wide variety of LSRs.

**Language-independence** No external resources such as corpora or other dictionaries are needed; the graph construction and alignment only rely on the information from the considered LSRs. Moreover, no external knowledge in the form of training data is needed, so that even for languages where such knowledge is rare or missing an alignment is possible.

**Flexibility** The graph construction as well as the actual alignment are highly parameterizable to accommodate different requirements regarding precision or recall and the peculiarities of the LSRs considered.

In the following sections, we discuss Dijkstra-WSA in more detail. This especially includes a description of its two basic steps: i) the initial construction of the graphs using appropriate parameters and ii) the alignment itself.

### 5.3.1 Graph Construction

As introduced in Section 3.4.2, we represent the set of senses (or synsets) of an LSR  $L$  as a set of nodes  $V$  and a set of edges  $E \subseteq V \times V$  between these nodes which represents semantic relatedness between them. We call the resulting graph a *resource graph*, and the properties of these resource graphs were described in Table 3.14.

#### Linking of Monosemous Lexemes

For resources such as Wiktionary where no or only few relations are present, we additionally propose to use the glosses of senses in the LSR to derive further edges in the following way, inspired by Navigli (2009a): for each monosemous, non-stopword lexeme  $\ell$  in the gloss of a sense  $s_1$  with a sense  $s_\ell$ , we introduce an edge  $(s_1, s_\ell)$ . Moreover, if there is another sense  $s_2$  with  $\ell$  in its gloss, we also introduce an edge  $(s_1, s_2)$ . This technique will be called *linking of monosemous lexemes* or *monosemous linking* throughout the rest of this thesis. The intuition behind this is that monosemous lexemes usually have a rather specific meaning, and thus it can be

expected that the senses in whose description they appear have at least a certain degree of semantic relationship. This relates to the notion of “information content” (Resnik, 1995), stating that senses in an LSR which are more specific (and hence more likely to be monosemous) are more useful for evaluating semantic similarity. Note that this step requires part of speech tagging of the glosses, which we perform as a preprocessing step. Thereby we filter out stopwords and words tagged as “unknown” by the POS tagger.<sup>1</sup>

As an example, consider the gloss of *Java*: “An object-oriented programming language”. Even in the absence of any semantic relations, we could unambiguously derive an edge between this sense of *Java* and the multiword noun *programming language* if the latter is monosemous, i.e. if there exists exactly one sense for this lexeme in the LSR. Also, if *programming language* appears in the gloss of one of the senses of *Python*, we can derive an edge between these senses of *Java* and *Python*, expressing that they are semantically related. An illustration of this is given in Figure 5.2.

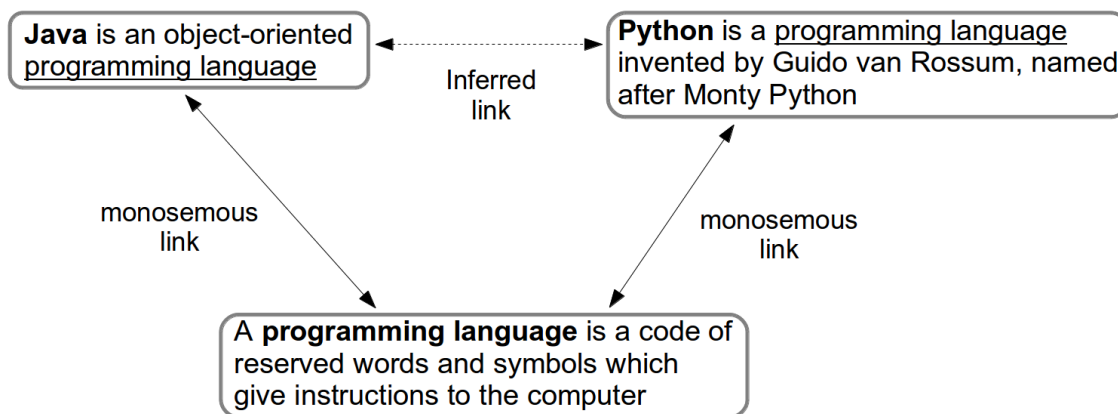


Figure 5.2: An example of monosemous linking: As *programming language* only has one meaning, we can create edges between this sense and the corresponding senses of *Java* and *Python* based on the glosses. We can also infer an edge between the latter two senses based on the assumption that they must also be related.

An important factor to keep in mind is the density of the resulting graph. In preliminary experiments, we discovered that linking every monosemous lexeme yields very dense graphs with short paths between most senses, which makes the distances less distinctive. In turn, we decided to exclude “common” lexemes and focus on more specific ones in order to increase the graph’s expressiveness. The indicator for this is the frequency of a lexeme in the LSR, i.e. how often it occurs in the glosses. The rationale is that less frequent lexemes might be more discriminative in describing a sense, thus leading to a more informative sense graph. Again, a similar argumentation is used by Resnik (1995). Our experiments on small development sets (100 random samples of each gold standard presented in Section 3.5) indeed show that a strict filter leads to discriminative edges resulting in high precision, while at the same time the graph sparsity decreases recall. Independently of the resource pair,

<sup>1</sup>We use the OpenNLP POS tagger available at <https://opennlp.apache.org/>

Pair	$\phi$	$P$	$R$	$F_1$
WordNet-OmegaWiki	1/1000	0.628	0.595	0.611
WordNet-OmegaWiki	1/200	0.561	0.657	0.605
WordNet-OmegaWiki	1/100	0.553	0.671	0.606
WordNet-Wiktioary	1/1000	0.821	0.204	0.327
WordNet-Wiktioary	1/200	0.733	0.236	0.357
WordNet-Wiktioary	1/100	0.696	0.256	0.374
GermaNet-Wiktioary	1/1000	0.901	0.755	0.822
GermaNet-Wiktioary	1/200	0.895	0.765	0.825
GermaNet-Wiktioary	1/100	0.896	0.772	0.829
FrameNet-Wiktioary	1/1000	0.788	0.254	0.384
FrameNet-Wiktioary	1/200	0.739	0.277	0.403
FrameNet-Wiktioary	1/100	0.654	0.320	0.430
Wiktioary-OmegaWiki	1/1000	0.771	0.337	0.469
Wiktioary-OmegaWiki	1/200	0.742	0.363	0.488
Wiktioary-OmegaWiki	1/100	0.740	0.389	0.510

Table 5.2: Influence of the frequency limit  $\phi$  (relative to the resource size) for linking monosemous lexemes in the LSRs, calculated on development sets of 100 random samples from a selection of gold standards not including Wikipedia (all other parameters fixed to appropriate values for each resource pair). Note that we refrained from experimenting with monosemous links for Wikipedia due to prohibitive computation times; instead, we utilize the existing links within the articles.

we discovered that setting this frequency limit value  $\phi$  to about 1/100 of the graph size (e.g. 1,000 for a graph containing 100,000 senses) gives the best balance between precision and recall. Larger values of  $\phi$  usually lead to no significant improvement in recall while the precision is continuously degrading. Note that Wikipedia was excluded from these experiments as the identification and linking of monosemous lexemes in all Wikipedia articles proved computationally too expensive. Instead, we decided to use only the explicitly encoded links (see next section). Table 5.2 illustrates this behavior on a selection of datasets.

### Graph Configurations

Following the considerations and observations made in the previous section, we experimented with three options for the construction of the graphs:

**Semantic relations only (*SR*)** This configuration directly corresponds to the graphs discussed in the analysis in Section 3.4.2, with one notable exception: intuitively, not all links in a Wikipedia article are equally meaningful. While the most central aspects are usually discussed at the beginning of an article and salient related articles are linked there, the later sections often contain references to articles which are less important in comparison. Thus, for the *SR* configuration, we decided to retain only the category links and the links within the first paragraph of the article. We assume that the targets of these links are most closely related to the sense which an article represents as the first paragraph usually includes a concise definition of a concept, and the category links allow determining the topic an article belongs to.

Resource	Senses	Config.	Relations	Rel./Sense	Isol. Senses
WordNet	117,659	<i>SR</i>	570,696	4.85	25%
		<i>LM</i>	1,414,940	12.05	7%
		<i>SR+LM</i>	1,985,636	16.88	2%
FrameNet	11,942	<i>SR</i>	76,315	6.39	2%
		<i>LM</i>	355,372	29.76	24%
		<i>SR+LM</i>	431,687	36.14	0%
VerbNet	31,891	<i>SR</i>	197,824	6.20	0%
		<i>LM</i>	0	0.0	100%
		<i>SR+LM</i>	197,824	6.20	0%
GermaNet	74,612	<i>SR</i>	193,669	2.60	0%
		<i>LM</i>	48,003	0.64	92%
		<i>SR+LM</i>	241,672	3.24	0%
Wiktionary en	421,848	<i>SR</i>	5,132	0.01	98%
		<i>LM</i>	2,360,187	5.59	32%
		<i>SR+LM</i>	2,365,319	5.61	30%
Wiktionary de	72,752	<i>SR</i>	44,523	0.61	69%
		<i>LM</i>	1,039,855	14.29	18%
		<i>SR+LM</i>	1,084,378	14.90	15%
Wikipedia en	2,921,455	<i>SR</i>	13,790,422	4.72	6%
		<i>LM</i>	69,429,790	23.77	5%
		<i>SR+LM</i>	83,220,212	28.49	4%
Wikipedia de	838,428	<i>SR</i>	3,721,150	4.44	7%
		<i>LM</i>	9,243,998	11.02	5%
		<i>SR+LM</i>	12,965,148	15.46	4%
OmegaWiki en	45,137	<i>SR</i>	62,104	1.38	41%
		<i>LM</i>	90,268	2.00	33%
		<i>SR+LM</i>	152,372	3.38	4%
OmegaWiki de	24,509	<i>SR</i>	32,705	1.33	45%
		<i>LM</i>	60,090	2.45	40%
		<i>SR+LM</i>	92,795	3.79	13%

Table 5.3: This table describes, among other statistics, what percentage of nodes remains isolated (i.e. with no attached edges) in different graph configurations using semantic relations only (*SR*), monosemous linking (*LM*,  $\phi = 1/100$ ) or both (*SR+LM*). Note that this number is maximal for the English Wiktionary as the few semantic relations and many missing glosses do not offer many possibilities for connecting nodes, while the German Wiktionary and OmegaWiki do not suffer from this problem as much. GermaNet is fully linked via relations, but has only few glosses which makes monosemous linking ineffective. WordNet, FrameNet and Wikipedia are densely linked in all configurations. Also note that for Wikipedia, *SR* means that we used category links and links from the first paragraph, while links from the rest of the article were used for the *LM* configuration. VerbNet has no glosses, so that the *LM* configuration is not applicable.



Table 5.3 gives an overview of the resulting graphs for each resource in this and the other discussed configurations.

**Linking of monosemous lexemes only (*LM*)** For this configuration, the limiting parameter  $\phi$  was set to 1/100 of the graph size for every resource except Wikipedia as described above. As our experiments show, linking the monosemous lexemes in the glosses while disregarding semantic relations results in well-connected graphs for all resources but GermaNet and Wiktionary. Only about 10% of the GermaNet senses have a gloss, thus this option was completely disregarded in this case. Note that Henrich et al. (2011) also have to construct pseudo-glosses (“lexical fields”) for a sense by collecting the lemmas from all senses which are reachable via semantic relations. For both Wiktionaries, the reason for the high number of isolated nodes are also missing glosses (cf. Section 3.4.1), although this issue is not as severe as for GermaNet. For Wikipedia, we refrained from using monosemous linking due to the prohibitive computation time. This would have required to process all article texts in order to discover valid link targets for every non-stopword. Instead, we decided to use all links from Wikipedia (excluding the links used for the *SR* configuration) in this case. The rationale is that in the majority of articles many meaningful terms link to the corresponding articles anyway, so that the resulting graph is comparable with those for the other LSRs.

**Combining both (*SR+LM*)** This configuration always yields the maximum number of available edges. In the evaluation, we report the results for GermaNet only for this configuration and omit the *SR* results for the sake of brevity as the influence on the F-measure for the GermaNet-Wiktionary alignment (see Section 5.3.3) is not statistically significant. For the English Wiktionary, this configuration only slightly increases the number of connected nodes in comparison to the *LM* configuration simply because very few edges for Wiktionary can be derived from semantic relations (cf. Section 3.4.2). For the German Wiktionary and OmegaWiki, on the other hand, well-connected graphs can be constructed.

## 5.3.2 Computing Sense Alignments

### Initialization

After resource graphs for both LSRs  $A$  and  $B$  are created, the trivial alignments are retrieved and introduced as edges between them. Trivial alignments are those between senses which have the same attached lexeme in  $A$  and  $B$  and where this lexeme is also monosemous within either resource. E.g., if the noun phrase *programming language* is contained in either resource and has exactly one sense in each one, we can directly infer the alignment. For Wikipedia, a lexeme is considered monosemous if there is exactly one article with this title, also counting titles with a bracketed disambiguation (e.g., *Java (programming language)* and *Java (island)* are two distinct senses of *Java*). While this method does not work perfectly, we observe a precision  $> 0.95$  for monosemous gold standard senses, which is in line with observations made, for instance, by Henrich et al. (2011).

---

Dijkstra-WSA( $A, B, \lambda$ )	
1	ASenseSet = A.senses
2	BSenseSet = B.senses
3	UnalignableSenses = $\emptyset$
4	
5	<b>foreach</b> sense $s \in$ ASenseSet
6	<b>if</b> ( $s.isMonosemous$ )
7	$t = findTrivialMatch(s, BSenseSet)$
8	<b>if</b> ( $t \neq null$ )
9	ASenseSet.remove( $s$ )
10	BSenseSet.remove( $t$ )
11	createEdge( $s, t$ )
12	
13	<b>foreach</b> sense $s' \in$ ASenseSet
14	ASenseSet.remove( $s'$ )
15	$T = findCandidatesWithSameLexeme(s', B)$
16	<b>if</b> ( $T \neq \emptyset$ )
17	$t' = findShortestPathToCandidates(s', T, \lambda)$
18	<b>if</b> ( $t' \neq null$ )
19	createEdge( $s', t'$ )
20	<b>else</b>
21	UnalignableSenses.put( $s'$ )
22	<b>else</b>
23	UnalignableSenses.put( $s'$ )

---

Table 5.4: Pseudocode of the Dijkstra-WSA algorithm.

### Alignment Procedure

We consider each sense  $s \in A$  which has not been aligned in the initialization step. For this, we first retrieve the set of possible target senses  $T \subset B$  (those with matching lemma and part of speech) and compute the shortest path to each of them with Dijkstra’s shortest path algorithm (Dijkstra, 1959). The candidate  $t \in T$  with the shortest distance is then assigned as the alignment target, and the algorithm continues with the next still unaligned sense in  $A$  until either all senses are aligned or no path can be found for the remaining senses. The intuition behind this is that the trivial alignments from the initialization serve as “bridges” between  $A$  and  $B$ , such that a path starting from a sense  $s_1$  in  $A$  traverses edges to find a nearby already aligned sense  $s_2$ , “jumps” to  $B$  using a cross-resource edge leading to  $t_2$  and then ideally finds an appropriate target sense  $t_1$  in the vicinity of  $t_2$ . Note that with each successful alignment, edges are added to the graph so that a different ordering of the considered senses yields different results. Also, in case of a tie (i.e. two candidates with the same distance) only the first one found is assigned as target sense. While we consequently observe slight differences for repeated runs using the same configuration, these are in no case statistically significant. The pseudo code of this algorithm is given in Table 5.4, while an example can be found in Figure 5.3.

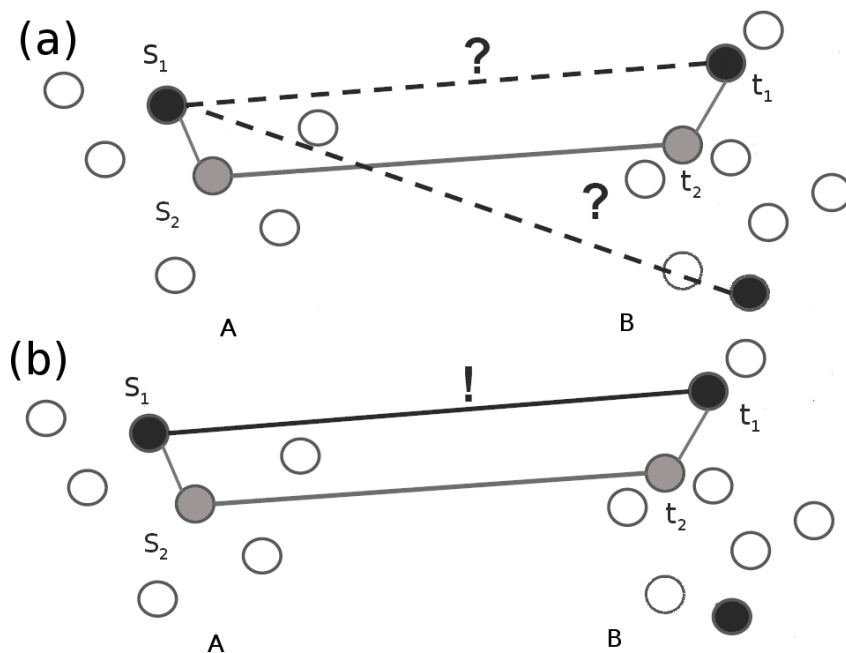


Figure 5.3: An example of how Dijkstra-WSA works. While there exist 2 candidates for aligning a sense  $s_1 \in A$  (dashed lines) (a), the correct one  $t_1 \in B$  can be determined by finding the shortest path using an already established edge between two monosemous senses  $s_2 \in A$  and  $t_2 \in B$  (solid line) (b).

### Parameter Influence

Apart from the already mentioned parameter  $\phi$  which influences the construction of the graph, an important variable in the actual alignment process is the maximum allowed path length  $\lambda$  of Dijkstra's algorithm. In general, allowing an unbounded search for the candidate senses is undesirable as long paths, while increasing recall, usually also lead to a decrease in precision. This is because the nodes which can be reached in many steps are usually also semantically distant from the source sense. In this respect, we found notable differences between the optimal configuration for individual resource pairs. However, the general observation is that short paths ( $\lambda \leq 3$ ) lead to a very high precision, while paths longer than 10 do not increase recall significantly any more. Table 5.5 shows some example configurations and results based on a development sets of 100 random examples per selected resource pair. If no development set is available to determine this parameter or a completely unsupervised setup is desired, a threshold can be estimated based on the size and density of the LSRs. For instance, for pairs including the English Wiktionary (which is rather sparse, cf. Section 3.4.2), short path lengths of 3 or 4 are advisable, while longer paths are useful in case of the more densely linked German Wiktionary.

A modification of the algorithm is to not only align the closest target sense, but all senses which can be reached with a certain number of steps (e.g. align all candidate senses connected by a path of length 3 or less). This caters for the fact that, due to different sense granularities, one coarser sense in  $A$  can be represented by several senses in  $B$  and vice versa (see Table 3.17 for the fraction of 1:n alignments in

Pair	$\lambda$	$P$	$R$	$F_1$
WordNet-OmegaWiki	3	0.936	0.348	0.507
WordNet-OmegaWiki	5	0.735	0.462	0.567
WordNet-OmegaWiki	10	0.737	0.466	0.571
WordNet-Wiktioary	2	0.952	0.128	0.225
WordNet-Wiktioary	3	0.696	0.256	0.374
WordNet-Wiktioary	4	0.214	0.419	0.283
GermaNet-Wiktioary	4	0.989	0.488	0.654
GermaNet-Wiktioary	8	0.890	0.755	0.817
GermaNet-Wiktioary	12	0.896	0.772	0.829
WordNet-Wikipedia	3	0.824	0.537	0.651
WordNet-Wikipedia	4	0.750	0.674	0.710
WordNet-Wikipedia	5	0.649	0.740	0.691
Wiktioary-Wikipedia en	3	0.860	0.493	0.627
Wiktioary-Wikipedia en	4	0.782	0.573	0.662
Wiktioary-Wikipedia en	5	0.495	0.666	0.568
Wiktioary-Wikipedia de	6	0.870	0.702	0.777
Wiktioary-Wikipedia de	7	0.823	0.738	0.778
Wiktioary-Wikipedia de	8	0.805	0.749	0.776
FrameNet-Wiktioary	3	0.913	0.242	0.383
FrameNet-Wiktioary	4	0.654	0.320	0.430
FrameNet-Wiktioary	5	0.571	0.365	0.445
Wiktioary-OmegaWiki	3	0.763	0.389	0.516
Wiktioary-OmegaWiki	4	0.583	0.500	0.538
Wiktioary-OmegaWiki	5	0.525	0.547	0.536

Table 5.5: Influence of the allowed path length  $\lambda$  for Dijkstra’s algorithm (all other parameters fixed to appropriate values for each resource pair).

the datasets). Regarding this modification, we made the observation that the recall improved (sometimes considerably), but at the same time the precision decreased, sometimes to an extent where the overall F-measure got worse (see Table 5.6). In the evaluation section, we explain which setting is used for which datasets and configurations. Generally, in situations where recall is more important than precision or if the task is explicitly defined as finding multiple alignments and not only the best one, it should be considered to allow multiple alignments. This is, for instance, the case for the sense clustering application we present in Section 8.1. However, as the dataset characteristics show, having only one candidate for a specific sense is, by a large margin, the more common case, so that 1:1 alignment should be the default setting.

Note again that for each alignment task (i.e. each pair of resources), we tuned the parameters for the algorithm on 100 random samples from each gold standard for a result balancing precision and recall as discussed above. Individual tuning of parameters was necessary for each pair due to the greatly varying properties of the LSRs. Our hope is however that the extensive analysis of different resource pairs, along with the detailed analysis of the alignment results, will facilitate the

Pair	Conf.	$P$	$R$	$F_1$
WordNet -OmegaWiki	1:1	0.728	0.471	0.572
WordNet -OmegaWiki	1:n	0.597	0.543	0.569
WordNet -Wiktionary	1:1	0.714	0.224	0.341
WordNet -Wiktionary	1:n	0.696	0.256	0.374
GermaNet -Wiktionary	1:1	0.890	0.755	0.817
GermaNet -Wiktionary	1:n	0.623	0.885	0.736
WordNet -Wikipedia	1:1	0.750	0.674	0.710
WordNet -Wikipedia	1:n	0.381	0.731	0.501
Wiktionary -Wikipedia en	1:1	0.807	0.560	0.661
Wiktionary -Wikipedia en	1:n	0.782	0.573	0.662
Wiktionary -Wikipedia de	1:1	0.975	0.620	0.758
Wiktionary -Wikipedia de	1:n	0.823	0.738	0.778
FrameNet-Wiktionary	1:1	0.654	0.320	0.430
FrameNet -Wiktionary	1:n	0.473	0.366	0.413
Wiktionary-OmegaWiki	1:1	0.583	0.500	0.538
Wiktionary-OmegaWiki	1:n	0.462	0.537	0.496

Table 5.6: Influence of allowing 1:1 or 1:n alignments (all other parameters fixed to appropriate values for each resource pair).

choice of appropriate parameters in the future, also in cases where no development set is available. Furthermore, while it would have been ideal to train and test on disjoint sets, we calculated the overall results on the full gold standards including the development sets to ensure comparability with the previous work.

### Hybrid Approach

Manual inspection of the results revealed that the alignments found by Dijkstra-WSA are usually different from those based on the gloss similarity. While the latter precisely recognizes alignments with similar wording of glosses, Dijkstra-WSA is advantageous if the glosses are different but the senses are still semantically close in the graph. In Section 5.3.4, we analyze this in greater detail. Exploiting this fact, we experimented with a hybrid approach: we perform an alignment using Dijkstra-WSA, tuned for high precision (i.e. using shorter path lengths) and fall back to using the results of the similarity-based approaches for those cases where no alignment target could be found in the graph. These results are marked with *+SB* in the result overview tables in the next section.

## 5.3.3 Evaluation

### Baselines and Previous Approaches

Apart from the naive baselines (see Section 2.2.2), we also report the similarity-based results (*SB*) for Wiktionary-OmegaWiki and WordNet-OmegaWiki as described in Sections 4.3 and 4.4, respectively. We calculated such an alignment for both Wiktionary-Wikipedia datasets using the same framework. For the datasets

	WordNet-Wiktionary				WordNet-Wikipedia			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>Acc.</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>Acc.</i>
Random	0.21	0.59	0.31	0.67	0.49	0.62	0.53	0.86
<i>1:1</i>	0.68	0.19	0.30	0.88	0.68	0.52	0.59	0.91
<i>1st</i>	0.33	0.51	0.40	0.80	0.51	0.72	0.60	0.88
<i>SB</i>	0.67	0.65	0.66	0.91	0.78	0.78	0.78	<b>0.95</b>
<i>SR</i>	<b>0.95</b>	0.13	0.23	0.89	<b>0.82</b>	0.63	0.71	0.93
<i>LM</i>	0.72	0.24	0.36	0.89	0.65	0.66	0.65	0.91
<i>SR+LM</i>	0.68	0.27	0.39	0.89	0.75	0.67	0.71	0.93
<i>SR+SB</i>	0.68	0.67	0.68	<b>0.92</b>	0.75	<b>0.87</b>	<b>0.81</b>	<b>0.95</b>
<i>LM+SB</i>	0.68	0.70	<b>0.69</b>	<b>0.92</b>	0.70	<b>0.87</b>	0.78	0.94
<i>SR+LM+SB</i>	0.68	<b>0.71</b>	<b>0.69</b>	<b>0.92</b>	0.75	<b>0.87</b>	<b>0.81</b>	<b>0.95</b>
<i>A</i> <sub>0</sub>	-	-	0.78	0.93	-	-	0.87	0.97

Table 5.7: Alignment results for the WordNet-Wiktionary and WordNet-Wikipedia datasets reported by Meyer and Gurevych (2011) and Niemann and Gurevych (2011): using semantic relations (*SR*), monosemous links (*LM*) or both (*SR+LM*). The similarity-based (*SB*) baselines, also used as a back-off for the hybrid approaches (*+SB*), were created as described in Section 4.2 and also originally reported by Meyer and Gurevych (2011) and Niemann and Gurevych (2011). For Wikipedia, *SR* means that only category links and links within the first paragraph were used, while *LM* uses links from the full article. Baselines and the inter-annotator agreements are given for reference.

that were not created in our work, we used the numbers reported in the original papers (Niemann and Gurevych, 2011; Meyer and Gurevych, 2011; Hartmann and Gurevych, 2013).

The sole exception is the GermaNet-Wiktionary dataset. The automatic alignment results (i.e. the outcome of the algorithm without manual post-correction) reported by Henrich et al. (2011) were unavailable for us as a baseline, as was the exact definition and composition of the lexical fields that were used as sense description in cases where the gloss was missing (cf. Section 3.5.1). This is why we were not able to calculate similarity values ourselves. The only information available to us were the absolute gloss overlap values which were the foundation for their alignment decision. Thus, reimplementing their original approach, we directly align senses regardless of their similarity if the decision is trivial (i.e. if there is only one candidate, see Section 5.3.2). We also do not train a machine learning classifier on a gold standard as we did for the other datasets. Instead, we closely follow the idea of Henrich et al. (2011) to align the most similar candidate regardless of the absolute value. We experimented with a threshold-based classifier applied to the overlap values, but could not achieve improved results in this way.

### WordNet-Wiktionary

Experiments using only the semantic relations (*SR*) yield a very low recall. The small number of sense relations with monosemous targets in Wiktionary makes the graph very sparse. Nevertheless, the alignment targets which Dijkstra-WSA finds are

mostly correct, with a precision greater than 0.95 even when allowing 1:n alignments. Using only monosemous links (*LM*) improves the recall considerably, but unlike the WordNet-OmegaWiki alignment, it stays fairly low. Consequently, even when using semantic relations and monosemous links in conjunction (*SR+LM*), the recall can only be increased slightly, leading to an overall F-measure of 0.39. As mentioned above, this is due to the Wiktionary glosses. A substantial fraction of senses has no gloss at all (see Table 3.11), and even if a gloss is present, it is typically short, containing few monosemous words as “link anchors”. This leads to many isolated nodes in the graph with no or only very few connecting edges. The ideal, rather short path length  $\lambda$  of 2 or 3 stems from the relatively high polysemy of the gold standard (see Table 3.17). We experimented with  $\lambda \geq 4$ , achieving reasonable recall, but in this case the precision was so low that this configuration, in conclusion, does not increase the F-measure. However, 1:n alignments work well with these short paths as the correct alignments are mostly in the close vicinity of a sense, hence we achieve an increase in recall in this case without too much loss of precision.

For the hybrid approach, we achieve an F-measure of 0.69 when using all edges (*SR+LM+SB*), setting the path length to 2, and also allowing 1:n alignments. This is a statistically significant improvement over (Meyer and Gurevych, 2011) which confirms the effectiveness of the hybrid approach and supports our hypothesis that the similarity-based approach effectively complements the graph-based algorithm in case of an insufficient graph structure.

### WordNet-Wikipedia

The *SR* configuration (WordNet relations + Wikipedia category/first paragraph links) yields the best precision (0.82), even outperforming the *SB* approach, and an F-measure of 0.71. This again shows that, using an appropriate parametrization ( $\lambda \leq 4$  in this case), Dijkstra-WSA can detect alignments with high confidence if both resources are sufficiently linked. The relatively low recall of 0.63 could be increased by allowing longer paths, however, as hyperlinks do not express relatedness as reliably as semantic relations, this introduces many false positives and thus lowers precision considerably. This issue becomes even more prominent when the links from the full articles are used as edges (*LM*). While the increase in recall is relatively small, the precision drops substantially. However, using all possible links (*SR+LM*) allows us to balance out precision and recall to some extent, while yielding the same F-measure as the *SR* configuration. Note that 1:1 alignments were enforced in any case, as the high polysemy of the dataset in conjunction with the dense Wikipedia link structure rendered 1:n alignments very imprecise.

Using the hybrid approach, we can increase the F-measure up to 0.81 (*SR+SB*), outperforming the results reported by Niemann and Gurevych (2011) by a significant margin. The F-measure for (*LM+SB*) is slightly worse due to the lower precision. Combining all edges (*SR+LM+SB*) does not influence the results any more, but in any case the hybrid configuration achieves the best overall recall (0.87). These results are plausible considering the resource properties discussed in Section 3.4. The LSRs have a very high lexical overlap of their glosses (which are present for almost all senses) and are both very well linked. These are ideal conditions for both angles to tackle the issue of WSA, so that combining them is very effective.

### GermaNet-Wiktionary

As stated above, we used the *SR+LM* configuration for GermaNet in every case, as only few GermaNet synsets have glosses (see Table 3.11) and the impact on the graph was thus insignificant (see Table 5.3). For the German Wiktionary, the much greater number of relations compared to its English counterpart is directly reflected in the results, as using the semantic relations only (*SR*) not only yields a good precision of 0.91 but also a decent recall of 0.59. Using the semantic relations together with monosemous links (*SR+LM*) yields the F-measure of 0.78, which is slightly lower than the similarity-based (*SB*) approach and the baselines. However, these baselines are already very strong in this case due to the low polysemy of the evaluation dataset. This is especially true for the *1st* baseline.

In the hybrid configuration, we can increase the performance to an F-measure of up to 0.85 (*SR+LM+SB*), outperforming all graph-based and similarity-based configurations as well as the baselines by a small, but still significant margin. The high lexical overlap between the available glosses of the resources (see Section 3.4.1) is a beneficial factor in this case.

In general, results for this pair of LSRs are high in comparison with other LSR pairs. We at least partly attribute this to the fact that the German Wiktionary and GermaNet both are densely linked with semantic relations which is especially beneficial for the recall of Dijkstra-WSA. This is also reflected in the ideal  $\lambda$  of 10-12. Many high-confidence edges allow long paths which still express a considerable degree of relatedness. However, while the results for 1:n alignments are already good, restricting oneself to 1:1 alignments gives the best overall results as the precision can then be pushed well above 0.80 without decreasing recall too much. An important factor in this respect is that the GermaNet-Wiktionary dataset, as already mentioned, has a relatively low degree of polysemy (compared, for instance, to WordNet-Wiktionary) and only few 1:n alignments (compared to WordNet-OmegaWiki), two facts which make the task easier.

### FrameNet-Wiktionary

This dataset is especially interesting, as FrameNet, with its focus on frame semantics, is inherently different from the other resources. The question is if the linking of senses via frames (see Sections 3.4.2 and 5.3.1 for more details) sufficiently expresses semantic relatedness, as semantic relations in WordNet or Wikipedia links do.

The answer to this question is yes – the results we observe for Dijkstra-WSA alone reach a good precision, while the unsatisfactory recall (also if monosemous linking is applied) can be attributed to the low density of the English Wiktionary. In this respect, the results look very similar to the other alignment of an expert-built LSR (WordNet) to this Wiktionary edition. The impression of a strong resemblance between the two datasets is reinforced even more by the fact that also for FrameNet-Wiktionary only very short paths ( $\lambda \leq 3$ ) yield reasonable results. Thus, we can tentatively conclude that, at least for our purposes, FrameNet frame relations are sufficiently accurate in expressing relatedness.

With the hybrid approach, we can again alleviate the problem of low recall, and also achieve a modest increase in precision, which seems plausible due to high lexical overlap between glosses in FrameNet and Wiktionary (see Table 3.8). Thus, using



	GermaNet-Wiktionary				FrameNet-Wiktionary			
	$P$	$R$	$F_1$	$Acc.$	$P$	$R$	$F_1$	$Acc.$
Random	0.44	0.51	0.47	0.54	0.56	0.54	0.55	0.53
$1:1$	<b>0.95</b>	0.58	0.72	0.74	0.90	0.22	0.35	0.78
$1st$	0.77	0.81	0.79	0.79	0.66	0.64	0.65	0.81
$SB$	0.90	0.76	0.83	0.81	0.73	0.75	0.74	0.86
$SR$	0.91	0.59	0.72	0.72	<b>0.97</b>	0.20	0.33	0.78
$LM$	0.85	0.72	0.78	0.76	0.66	0.28	0.40	0.76
$SR+LM$	0.85	0.72	0.78	0.76	0.73	0.31	0.44	0.78
$SR+SB$	0.87	0.81	0.84	<b>0.82</b>	0.77	0.75	0.76	0.87
$LM+SB$	0.83	<b>0.86</b>	<b>0.85</b>	0.81	0.77	0.76	0.76	0.87
$SR+LM+SB$	0.83	<b>0.86</b>	<b>0.85</b>	0.81	0.77	<b>0.79</b>	<b>0.78</b>	<b>0.88</b>
$A_0$	-	-	N/A	N/A	-	-	0.80	N/A

Table 5.8: Alignment results for GermaNet-Wiktionary and FrameNet-Wiktionary. Note that for GermaNet, the  $SR+LM$  configuration was always used. The different configurations given for this alignment thus only apply to Wiktionary. Also note that not all agreement measures were reported for the annotations of these datasets, which is why we mark them as “not available” (N/A).

	WordNet-OmegaWiki				Wiktionary-OmegaWiki			
	$P$	$R$	$F_1$	$Acc.$	$P$	$R$	$F_1$	$Acc.$
Random	0.46	0.35	0.40	0.51	0.35	0.40	0.37	0.57
$1:1$	0.36	0.64	0.46	0.55	0.49	0.61	0.54	0.67
$1st$	0.34	<b>0.80</b>	0.48	0.47	0.41	<b>0.77</b>	0.54	0.57
$SB$	0.55	0.53	0.54	0.73	0.67	0.65	0.66	0.78
$SR$	<b>0.66</b>	0.45	0.53	0.76	<b>0.83</b>	0.28	0.42	0.75
$LM$	0.62	0.54	0.58	<b>0.77</b>	0.58	0.50	0.54	0.72
$SR+LM$	0.56	0.69	0.62	0.74	0.58	0.51	0.54	0.72
$SR+SB$	0.60	0.65	0.63	0.76	0.71	0.69	0.70	<b>0.81</b>
$LM+SB$	0.60	0.70	0.64	0.76	0.71	0.71	<b>0.71</b>	<b>0.81</b>
$SR+LM+SB$	0.57	0.75	<b>0.65</b>	0.75	0.68	0.73	<b>0.71</b>	0.80
$A_0$	-	-	0.84	0.85	-	-	0.80	0.85

Table 5.9: Alignment results for WordNet-OmegaWiki and Wiktionary-OmegaWiki.

all linking possibilities in conjunction with the backoff ( $SR+LM+SB$ ), the strong  $SB$  baseline is outperformed. this is also the case for the other backoff configurations, but in these cases the improvement is not statistically significant.

### WordNet-OmegaWiki

When using only semantic relations ( $SR$ ), we achieve an F-measure of 0.53 which is comparable with the 0.54 for the  $SB$  approach. Notably, Dijkstra-WSA has a high precision, while the recall is considerably lower due to the relative sparsity of the resulting OmegaWiki resource graph. When adding more edges to the graph by linking monosemous lexemes ( $SR+LM$ ), we can drastically improve the recall, leading to an overall F-measure of 0.62, which is a significant improvement over

our previous results. Using monosemous links only (*LM*), the result of 0.58 still outperforms *SB* due to the higher precision. Building a graph from glosses alone is thus a viable approach if no or only few semantic relations are available. Regarding the path lengths,  $\lambda = 10$  works best when semantic relations are included in the graph, while for the *LM* configuration shorter paths ( $\lambda \leq 5$ ) were more appropriate. The intuition behind this is that for semantic relations, unlike monosemous links, even longer paths still express a high degree of semantic relatedness. Also, when semantic relations are involved allowing multiple alignments increases the overall results (which is in line with the relatively high number of 1:n alignments in the gold standard), while this is not the case for the *LM* configuration. Here, the edges again do not sufficiently express relatedness.

Using the hybrid approach (*+SB*), we can increase the F-measure up to 0.65 if semantic relations and monosemous linking are combined (*SR+LM*) and the parameters are tuned for high precision ( $\lambda \leq 3$ , 1:1 alignments). This is significantly better than Dijkstra-WSA alone in any configuration. In this scenario, we also observe the best recall of all non-baseline configurations. In summary, though, the sparsity of OmegaWiki, together with the low lexical overlap between WordNet and OmegaWiki glosses (which already impairs the *SB* results) leads to the lowest F-measure across all resource pairs. Plainly spoken, the two LSRs are not a good choice for alignment with the approaches presented here. Note that we already made this assumption based on the differences between the two resources in our analysis in Section 3.4.

### Wiktionary-OmegaWiki

The sparsity of both the English Wiktionary and OmegaWiki (see Table 5.3) is directly reflected in the low recall for the *SR* configuration, which one the other and yields the precision – this is in line with the observations for most other datasets. It is notable though that adding additional edges is not quite as effective as expected. While the recall can again be substantially improved, the precision drops sharply. Here, the presence of shorter glosses with lower quality in both collaboratively constructed resources seems to impair the results.

Nevertheless, the glosses provide enough additional information for the backoff (*+SB*) to be effective, so that we still reach a significant improvement over our own previous results in any hybrid configuration. While the *LM+SB* setup yields the best balance between precision and recall, the full set of available edges (*SR+LM+SB*) maximizes the recall. This is an observation that can also be made for several other datasets such as WordNet-Wiktionary and Wiktionary-Wikipedia. Moreover, it is noteworthy that, due to the inherently less accurate linking of monosemous lexeme as compared to semantic relations, short paths ( $\lambda \leq 4$ ) and a restriction to 1:1 alignment yields the best results in this case.

### Wiktionary-Wikipedia English

Comparable to the WordNet-Wiktionary alignment, the *SR* configuration reaches very high precision (the best overall), but suffers from low recall due to the small number of semantic relations in the English Wiktionary. However, the recall is

	Wiktionary-Wikipedia en				Wiktionary-Wikipedia de			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>Acc.</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>Acc.</i>
Random	0.41	0.49	0.45	0.48	0.68	0.40	0.51	0.46
1st	0.17	0.56	0.26	0.33	<b>1.0</b>	0.63	0.77	0.75
1:1	0.23	<b>0.88</b>	0.36	0.37	0.93	0.66	0.78	0.74
<i>SB</i>	0.61	0.65	0.63	0.84	0.85	0.46	0.60	0.57
<i>SR</i>	<b>0.88</b>	0.29	0.44	0.84	0.85	0.61	0.71	0.66
<i>LM</i>	0.78	0.55	0.65	0.87	0.82	0.74	0.78	0.71
<i>SR+LM</i>	0.78	0.57	0.66	<b>0.88</b>	0.80	0.75	0.77	0.70
<i>SR+SB</i>	0.63	0.75	0.68	0.86	0.90	0.71	0.79	0.74
<i>LM+SB</i>	0.61	0.79	<b>0.70</b>	0.86	0.86	0.77	<b>0.81</b>	<b>0.75</b>
<i>SR+LM+SB</i>	0.62	0.81	<b>0.70</b>	0.86	0.80	<b>0.79</b>	0.79	0.72
<i>A</i> <sub>0</sub>	-	-	0.79	0.95	-	-	0.85	0.89

Table 5.10: Alignment results for Wiktionary-Wikipedia in both English and German. For Wikipedia, *SR* means that only category links and links within the first paragraph were used, while *LM* uses links from the full article.

slightly better, which can be attributed to the way in which the gold standard was automatically created, as more common words, which are better linked in Wiktionary, tended to be retained (see Section 3.5.2). This is also beneficial for the *LM* and *SR+LM* configurations, where the recall is again substantially better than for WordNet-Wiktionary, and the still good precision of 0.78 (at a path length  $\lambda$  of 4) yields an overall F-measure which outperforms the *SB* approach. The lower precision of Wikipedia links in the *LM* configuration is, unlike for WordNet-Wikipedia, only an issue for  $\lambda > 4$ , which is why we refrained from using longer paths.

In the hybrid approach, the overall F-measure can be increased up to 0.70, which is again a significant improvement over Dijkstra-WSA alone and the baselines. Especially the recall can be improved in this way, while the precision remains relatively low. An analysis of the hybrid approach revealed that, for this dataset, Dijkstra-WSA already covers many cases which are also recognized by the *SB* backoff. Thus, the similarity-based classification only achieves low precision on the remaining examples. Nevertheless, the results confirm that Dijkstra-WSA effectively works on this dataset of two large collaboratively constructed resources.

### Wiktionary-Wikipedia German

On this dataset, the recall for the *SB* approach we employed is fairly low due to the richer morphology and peculiar formation of compounds in German. We did not use a compound splitter (an obvious extension for future work), so that, for instance “*Kinderspiel*” and “*Spiel für Kinder*” (both meaning “a game for children”) could not be lexically matched. Interestingly, the baseline results for the other German dataset (GermaNet-Wiktionary) are seemingly not affected by this as much. One reason for this might be that the similarity for GermaNet synsets was, for most part, calculated on lexical fields, i.e. artificial glosses (cf. Section 5.3.3), which are more robust due to the much lower number of inflections and compounds in comparison to regular glosses.

	FN-WKT		WN-WKT		GN-WKT		WN-WP	
<i>SB</i>	561	191	203	110	20,727	6,400	178	49
	214	1,823	98	2,012	2,341	16,168	51	1,537
<i>SR</i>	153	622	41	272	16,086	11,041	143	84
	4	2,010	2	2,108	1,592	16,917	32	1,556
<i>LM</i>	222	555	76	237	19,426	7,701	149	78
	112	1,902	29	2,081	3,400	15,109	79	1,509
<i>SR+LM</i>	241	534	85	228	19,455	7,672	153	74
	88	1,926	41	2,069	3,424	15,085	51	1,537
<i>SR+SB</i>	580	195	210	103	22,009	5,118	197	30
	169	1,845	99	2,011	3,196	15,313	65	1,523
<i>LM+SB</i>	588	187	219	94	23,433	3,694	197	30
	180	1,834	106	2,004	4,788	13,721	83	1,505
<i>SR+LM+SB</i>	616	159	221	92	23,434	3,693	197	30
	189	1,825	104	2,006	4,792	13,717	65	1,523

Table 5.11: Confusion matrices for datasets from previous work. The dataset sizes are stated in Table 3.17. For each cell, top left: true positives, top right: false negatives, bottom left: false positives and bottom right: true negatives.

On the other hand, the baselines are very strong, due to the disproportionately large number of positive examples. This is especially true for the *1:1* setup which reaches perfect precision. In other words, whenever there is only one alignment candidate, it is already the correct one. The *SR* approach in isolation fails to outperform this strong baseline, as the density of the German Wiktionary (while still significantly higher than for the English one) is not sufficient to reach good recall. This is alleviated by applying monosemous linking (*LM*), which gives a significant boost in recall, at least reaching parity with the baselines. Although we observed for the other datasets that linking monosemous lexemes in the glosses yields a substantial decrease in precision, this is not the case here; the German compounding which hampers *SB* performance is beneficial in this case, as it yields more accurate link targets. For example, the above mentioned “*Kinderspiel*” is unambiguously linked to the only Wikipedia article with this title. This is also in line with the observation that long paths still work well on this dataset (see Table 5.5), which is a hint that most edges accurately express semantic relatedness.

Using the hybrid approach, we can increase both the precision and the recall, thanks to the high precision of the components which complement each other well in this case. In this way, using the *LM+SB* approach, we are able to outperform the strong baselines by a small but still significant margin. Note that the difference between the *1:1* and *1:n* configuration is very small in this case, as most Wiktionary senses in the gold standard have only one candidate article anyway (see Table 3.17).

## Summary

In conclusion, our experiments consistently demonstrate that combining Dijkstra-WSA with a similarity-based approach as a backoff yields the strongest performance. This supports our hypothesis that the graph-based and similarity-based alignment

	WN-OW	WKT-OW	WKT-WP de	WKT-WP en
<i>SB</i>	112 98	124 66	10,024 11,820	49 26
	90 383	61 335	1,743 8,210	31 261
SR	94 116	53 137	13,269 8575	22 53
	48 425	11 385	2,302 7,651	3 289
LM	113 97	95 95	16,119 5,725	44 31
	68 405	68 328	3,460 6,493	12 280
<i>SR+LM</i>	145 65	97 93	16,391 5,453	43 32
	116 357	70 326	4,074 5,879	12 280
<i>SR+SB</i>	137 73	132 58	15,497 6,347	56 19
	91 382	55 341	1,811 8,142	33 259
<i>LM+SB</i>	146 64	134 56	16,798 5,046	59 16
	98 375	56 340	2,769 7,184	38 254
<i>SR+LM+SB</i>	157 53	139 51	17,174 4,670	61 14
	118 355	64 332	4,319 5,634	38 254

Table 5.12: Confusion matrices for datasets we created in our work. The dataset sizes are stated in Table 3.17. For each cell, top left: true positives, top right: false negatives, bottom left: false positives and bottom right: true negatives.

approaches are complementary in the sense that they succeed in finding different alignment targets, and that a combination of them is superior to using either of them in isolation. The combination is especially useful if either source of information is unsatisfactory. For missing or short glosses, structural information boosts the results significantly, and this also applies the other way round. However, we also see that the graph-based approach works well on its own if the resources are densely linked, as in these cases the hybrid approach leads to only further modest improvement. Generally speaking, we observe that the results largely reflect the observations (and subsequent hypotheses) made about the LSRs in Section 3.4, and as we covered resource pairs with many different properties in our experiments, we are confident that our analysis facilitates choosing the best approach for WSA experiments on further resource pairs in the future.

The results on the full resources for the best alignment configurations (along with even higher performing ones presented in the next chapter) are freely available as part of UBY (see Section 7.4) and on our website.<sup>2</sup>

### 5.3.4 Error Analysis

The by far most significant error source for Dijkstra-WSA, reflected in the relatively low recall for different configurations, is the high number of false negatives, i.e. sense pairs which were not aligned although they should have been (see Tables 5.11 and 5.12 for a detailed breakdown). This is especially striking for the alignments which involve Wiktionary. As discussed earlier, Wiktionary contains a significant number of empty glosses, and in cases where a gloss is present, it is often rather short and contains only few monosemous terms. A prototypical example is the first sense of

<sup>2</sup><https://www.ukp.tu-darmstadt.de/data/lexical-resources/>

*seedling*: “A young plant grown from seed”. This gloss has no monosemous words which could be linked, and as there are also no semantic relations attached to this sense which could be exploited, the node is isolated in the graph. Our experiments show that for the English Wiktionary, even when optimizing the parameters for recall, around 30% of the senses remain isolated, i.e. without edges. This is by far the highest value across all resources (see Table 5.3). Solving this problem would require making the graph more dense, and especially finding ways to include isolated nodes as well. An approach which tackles this issue by also linking polysemous lexemes in glosses was recently presented by Pilehvar and Navigli (2014). We will present further details about this approach in Section 6.3.3.

However, the *seedling* example also shows why the hybrid approach works so well: The correct WordNet sense “young plant or tree grown from a seed” was recognized by the similarity-based approach with high confidence, where the alternatives for this lexeme would have been “One grown in a nursery for transplanting” and “A tree smaller than a sapling”.

With regard to false positives, Dijkstra-WSA and the similarity-based approaches perform comparably. This is because senses with very similar wording are likely to share the same monosemous words, leading to a close vicinity in the graph and the false alignment. Moreover, if two senses within a resource are very similar, they are likely to be transitively connected via semantic relations anyway (for instance, as subsenses of the same broader sense), so that even a limitation to the *SR* configuration, i.e. refraining from monosemous linking, can only partially address the issue of false positives.

As an example, consider two senses of *bowdlerization* in WordNet (“written material that has been bowdlerized”) and Wiktionary (“The action or instance of bowdlerizing; the omission or removal of material considered vulgar or indecent.”). While these senses are clearly related, they are not identical and should not be aligned. Nevertheless the similar wording (and especially the use of the highly specific verb “bowdlerize”) results in an alignment. As for the similarity-based approaches, it is an open question how this kind of error can be effectively avoided (cf. (Meyer and Gurevych, 2011)).

There is a considerable number of examples where Dijkstra-WSA recognizes an alignment which similarity-based approaches do not detect.<sup>3</sup> Consider the two senses of *Thessalonian* in Wiktionary and WordNet: “A native or inhabitant of Thessalonica” and “Someone or something from, or pertaining to, Thessaloniki”. These are (mostly) identical and should be aligned, but there is no word overlap due to the interchangeable usage of the synonyms “Thessalonica” and “Thessaloniki”. However, those terms are both monosemous in WordNet as well as in Wiktionary, sharing the also monosemous noun “Greece” in their glosses. This yields the bridge between the resources to find a path and correctly derive the alignment.

---

<sup>3</sup>This can be derived by comparing the true positive (TP) numbers for the *SB* baselines in the confusion matrices to the “+*SB*” configurations. The difference between these numbers are the TPs found by Dijkstra-WSA alone.

### 5.3.5 Issues with VerbNet Alignments

As a final remark to this chapter, we want to separately discuss alignment efforts concerning VerbNet, which remained largely unsuccessful. We already mentioned in the previous chapter that VerbNet senses do not have glosses, so that we expected a lot from graph-based approaches. A major obstacle, however, are the extremely fine-grained sense distinctions (cf. Table 3.17). A typical example is the verb *to bleed* which has 21 senses, with very subtle differences which are reflected in none of the other LSRs. This makes the alignment task, or more specifically, beating a naive baseline very hard. If we just align all valid alignment candidates, we already achieve a precision of 0.77 for VerbNet-FrameNet and even 0.82 for VerbNet-WordNet, at a perfect recall of 1.0. In other words, on average 4 out of 5 of the possible alignment candidates are correct. In our example, 14 of the 21 senses of *to bleed* can be correctly aligned to only one single WordNet sense “lose blood from one’s body”. Moreover, the extreme polysemy entails only very few lexemes which are monosemous, i.e. which have only one sense. Identifying these to compute trivial alignments is, however, a prerequisite of the Dijkstra-WSA algorithm, which is thus rendered very ineffective. As mentioned in the introduction to this chapter, it is possible to construct a meaningful and well-connected graph for VerbNet by linking verbs in the same Levin verb class (Levin, 1993), but without a sufficient number of trivial bridges to the other LSRs the resulting paths are not informative for an alignment decision.

Due to the lack of glosses, VerbNet can also not be tackled with the joint approaches we describe in the following chapter, so that in summary, all approaches we investigate in this thesis proved to be ill-suited for VerbNet. Nevertheless, identifying ways to include this LSR into the WSA framework is still planned for future work. One idea to address the extremely skewed class distribution is to use the “align all” scenario as a starting point and try to confidently identify incorrect alignments to further increase precision. In other words, the usual task of identifying alignments would be reverted to identifying non-alignments based on the structural properties of VerbNet. This would, however, require a deeper analysis of the LSR, as well as novel algorithmic approaches which are tailored to this particular problem. Considering the syntax seems imperative for these investigations due to the strong syntactic focus of VerbNet.

## 5.4 Chapter Summary and Contributions

In this chapter, we present Dijkstra-WSA, a graph-based algorithm for word sense alignment. We show that this algorithm on its own performs competitively on 6 out of 8 evaluation datasets. This supports our hypothesis that WSA solely based on the structural properties of LSRs (and without external knowledge in the form of annotated training data or corpora) is effective given they are sufficiently linked, using either semantic relations, monosemous links or (most effectively) a combination of both. These experiments also show that good results are possible without making assumptions about the semantics of the edges, other than the general notion of relatedness they express, which makes our method flexibly applicable to different pairs of LSRs. Moreover, while estimating parameters on a development set proved

helpful in order to optimize for precision or recall, reasonable results can still be achieved by using default parameters. These are significant insights which will be very valuable for investigating other WSA datasets and approaches in the future.

A hybrid approach, i.e. combining Dijkstra-WSA with the similarity-based approach presented in Chapter 4, leads to a statistically significant improvement over the previous results on *all* of the considered datasets. This supports our original hypothesis that both ways of assessing the equivalence of senses in different resources are valid and complement each other. This also motivates our experiments on more sophisticated methods to combine them which we present in the next chapter.

To summarize, the contributions presented in this chapter are:

**Contribution 1** We present Dijkstra-WSA, a novel graph-based algorithm for word sense alignment which solely depends on the graph structure and shows competitive performance in case when sufficiently dense graphs can be derived.

**Contribution 2** We combine Dijkstra-WSA with the previously presented similarity-based alignment in a backoff approach and achieve state-of-the-art performance on every considered dataset, effectively proving that both methods complement each other.



# Chapter 6

## Joint Approaches to Word Sense Alignment

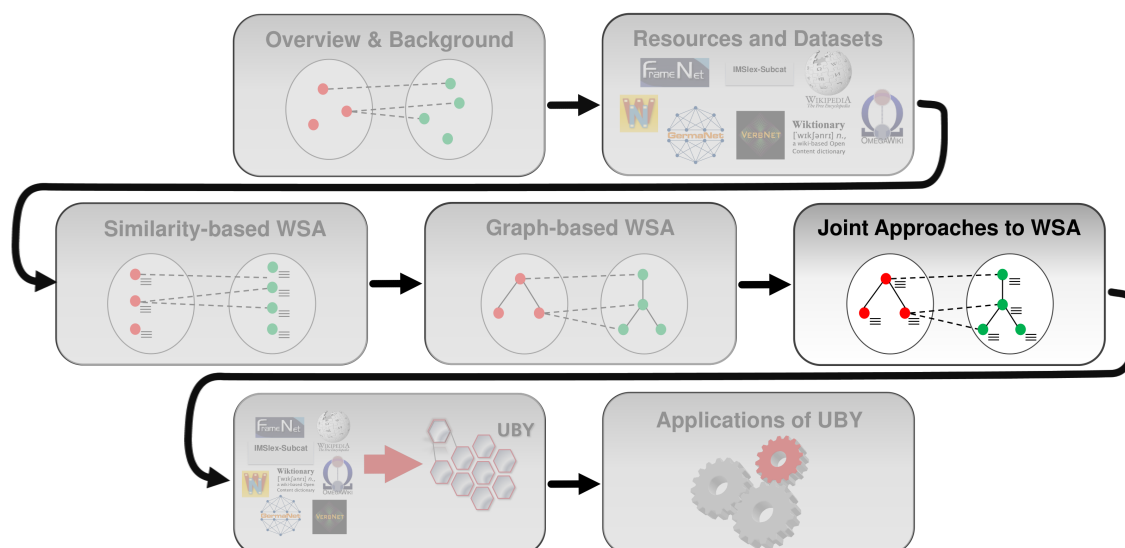


Figure 6.1: Visual outline of the thesis.

### 6.1 Introduction

We have seen in the last two chapters that similarity-based and structure-based approaches to WSA both have their benefits when covering different aspects of sense similarity, and these benefits are mostly orthogonal: the former approaches work generally well, but they are at a loss if the phrasing of sense descriptions is too different. The latter approaches can alleviate that, but are limited in their performance in case of sparsely connected LSRs such as Wiktionary.

As we already discussed, a simple fallback approach which applies both ideas in two separate steps (Section 5.3.2) leads to significant improvements in F-measure. However, there exist only few previous works which effectively combine both notions of similarity in a more elaborate framework for WSA or other purposes. In other

Work	Resource (pair)
(Bond and Foster, 2013)	WordNet-Wiktionary
(Navigli and Ponzetto, 2012a)	WordNet-Wikipedia
(Ferrandez et al., 2010)	WordNet-FrameNet
(De Melo and Weikum, 2008a)	Wordnet construction

Table 6.1: Previous work on WSA using combined features.

words, the different approaches to compute sense similarity have mostly been used in isolation, or combined in a shallow or restricted way.

More complex approaches (which we will present in the next section) usually require extensive feature engineering, mostly on WordNet-specific information types, which makes their applicability to other resources difficult. In contrast, our guiding idea is investigating WSA from a broad perspective, i.e. examining the applicability of approaches to as many heterogeneous resources as possible.

Thus, after discussing the previous work in Section 6.2, we present a framework which fills this gap in the body of WSA research by combining different dimensions of similarity in a generic and flexible machine learning approach (Section 6.3). It efficiently achieves state-of-the-art WSA performance on a variety of resource pairs. We also briefly discuss some experiments on aligning several resources at once (N-way alignment) in Section 6.4. We conclude in Section 6.5 with a summary of our contributions.

## 6.2 Previous Work

Bond and Foster (2013) aim to enrich wordnets in many different languages by aligning them to Wiktionary in the course of the *Open Multilingual Wordnet* project. While their alignment algorithm is based on gloss similarity (following Niemann and Gurevych (2011), see Section 4.2), they apply the additional feature of translation overlap, i.e. two senses are assumed to be equivalent if they share many translations. While the accuracy proved to be sufficient (around 0.90), a major issue is the relatively low coverage of translations in Wiktionary, especially for smaller languages, which impairs the effectiveness of this approach. This is in line with our own observations regarding this feature (see Section 6.3.1). Also, although Bond and Foster (2013) briefly describe the relational structure of the various wordnets, they do not make use of any structural features for the actual alignment.

Building on their own previous work on Wikipedia categories discussed earlier (Ponzetto and Navigli, 2009), Navigli and Ponzetto (2012a) align WordNet with Wikipedia articles in the course of creating the large-scale multilingual LSR *Babel-Net*, reaching an F-measure of 0.78 on their own gold standard data set. Besides using bag-of-words overlap to compute gloss similarity (similar to our similarity-based approach), they build “disambiguation contexts” for Wikipedia articles by, for instance, using redirect links, and then disambiguate the lexemes in these contexts. To this end, a graph structure is built from WordNet semantic relations covering all possible WordNet senses of all lexemes contained in such a context, and local vicinity is used to derive the correct alignment. However, in contrast to our approach, the

information contained in the graph structure of Wikipedia is largely disregarded, as only a subset of Wikipedia links is used to compose the disambiguation contexts. Moreover, for the actual alignment step, just a locally restricted subset of WordNet relations is used to make the decision, not the full WordNet graph, which would potentially provide additional valuable information about the senses to be aligned. As this approach is most closely related to our alignment algorithm, we will provide an evaluation of the WordNet-Wikipedia alignment contained within *BabelNet* on our own gold standard. Note again that the other alignments contained in *BabelNet* (WordNet-Wiktionary, WordNet-OmegaWiki) were not accessible to analysis due to a different representation of identifiers, and neither were the original gold standards, which is why we have to restrict ourselves to the gold standard available to us.

Ferrandez et al. (2010) align FrameNet LUs and WordNet synsets by combining different features in a machine learning approach. For a candidate pair, they first traverse the relations in both resources independently to construct “neighborhood graphs”, with the starting word sense at the center. Then, for each neighbor (appearing in any or in both neighborhoods) they calculate the distance to the centering word of each neighborhood and produce a normalized similarity score based on this, hence incorporating structural information from both resources. Plainly spoken, if both senses have similar neighbors in the respective LSRs, they are also assumed to be similar; this idea is in line with the idea for Dijkstra-WSA we presented earlier. As an additional feature, they also consider the textual similarity between glosses, but only on the character level. Using 100 examples for training their classifier, they achieve an accuracy of 0.77. While this approach is similar to the idea we present in this chapter, the algorithm is less generic as it heavily relies on the particular relation types in WordNet (e.g. hyponymy, meronymy) and FrameNet (e.g. inheritance, causative) to assign optimal edge weights for the graph, which impairs the applicability to other LSRs. Moreover, Ferrandez et al. (2010) do not elaborate on the behavior of their classifier in cases where either distances or gloss similarities are (partially) missing, as these cases are negligible when examining the expert-built resources FrameNet and WordNet (cf. Section 3.4). For collaboratively constructed resources such as Wiktionary, however, this possibility also needs to be considered.

De Melo and Weikum (2008a) also use a machine learning approach – not with the goal of aligning existing LSRs, but creating new ones. More precisely, they aim to create wordnets in a target language  $L_0$  other than English by using the structure of the Princeton WordNet as a “scaffold”. They tackle this issue by first providing a set of candidate translations for the lexemes contained in a WordNet synset from translation dictionaries, and then deciding for each translation if it is appropriate for this synset or not, based on a manually annotated training set. They train the classifier on a large variety of features based on the structure and the content of WordNet, and thereby reach a precision of 0.81. However, this approach is not easily generalizable as they also use WordNet-specific features such as corpus frequencies which are not readily available for most LSRs. Moreover, the task is inherently easier than (cross-lingual) WSA, because deciding if a lexeme  $l$  is a valid lexicalization for a concept in WordNet and then *creating* a new corresponding synset in  $L_0$  circumvents the more challenging step of *choosing* the correct target in an existing LSR.

## 6.3 Joint Modeling of Features

To address the disadvantages of the similarity- and the graph-based approaches which we exhaustively discussed earlier, and to truly leverage their benefits, we jointly model the different aspects of sense similarity by applying machine learning techniques to WSA.<sup>1</sup> Unlike previous approaches, we do not engineer our features towards a specific resource pair or application scenario, rendering the approach proprietary. Instead, we aim to combine the generic features discussed in Chapters 4 and 5 which are applicable to a variety of resources. Thus, we take advantage of both (orthogonal) ways of identifying equivalent senses and develop a very robust and flexible WSA framework. We show that this combination leads to state-of-the-art WSA performance without the need for extensive, resource-specific feature engineering.

The basic steps of our alignment framework are:

1. For each sense in one resource, all possible candidates in the other resource are retrieved as discussed in Section 2.2.
2. For each candidate pair, we calculate a set of features describing different dimensions of similarity.
3. For a subset of candidate pairs (the gold standard), the alignment decision is made by human annotators; we again rely on the gold standards already described in Section 3.5.
4. A machine learning classifier is trained on the gold standard, and an alignment decision is made for the remainder of the candidate pairs to produce a complete alignment of the LSRs. In our setup, we use a 10-fold cross validation to evaluate the classifiers.

We explain the different steps of the algorithm in more detail in the following sections.

### 6.3.1 Feature Engineering

As stated above, the selection of features for our machine learning approach was driven by the premise to keep the framework as generic and resource-agnostic as possible, in order to ensure applicability to many different LSRs without additional engineering effort.

Consequently, following our earlier WSA efforts, we focus on the similarity measures COS and PPR (explained in more detail in Section 4.2) for glosses, at least for the LSRs for which these could be calculated. As pointed out earlier, this was not the case for the GermaNet-Wiktionary alignment; here, only gloss overlap similarity values were available and thus used as a feature. While we had no choice for GermaNet-Wiktionary, we deliberately refrained from using gloss overlap for all

---

<sup>1</sup>The algorithm presented in Section 4.2 also uses a machine learning component. However, only two features expressing the same notion (gloss similarity) are employed in a simple threshold-based setup.

	Part of speech	Sense index	Translation overlap	Example sentences
WordNet	✓	✓	✗	✓
FrameNet	✓	✗	✗	✓
GermaNet	✓	✗	✗	✗
Wiktionary	✓	✓	✓	✓
OmegaWiki	✓	✗	✓	✓
Wikipedia	✗	✗	✓	✗

Table 6.2: Available machine learning features for different LSRs.

other datasets. Preliminary experiments showed that this additional measure had no significant impact on the results.

With regard to expressing distance between senses, we rely on the Dijkstra-WSA algorithm presented in the previous chapter. However, while Dijkstra-WSA is designed to directly align candidate senses which are closest to the source senses in the resource graph, in this setup we save the distance for each candidate sense and directly use it as a feature, expressing semantic relatedness based on the structure of both underlying resources. When no distance can be computed (in case of a disconnected graph), we assume infinite distance.

### Additional Features

We also experimented with other features which were accessible directly from the resources, i.e. without the need for external knowledge or extensive computational effort. We believe that this reflects a realistic alignment setup between two arbitrary resources. These features were not available for every resource pair, thus we state in Table 6.2 for each feature which dataset it could be applied to.

**Part of speech** The part of speech was incorporated following the well-known fact that different parts of speech have different characteristics, e.g. regarding the degree of polysemy. This feature was disregarded for Wikipedia, as it only contains nouns.

**Sense index** The sense index marks the position of a sense in the list of senses for a certain lexeme (cf. Section 2.2.2). While this position reflects corpus frequencies in WordNet (i.e. the first sense is the most frequent sense), no such explicit statement can be made for Wiktionary, although it is assumed that more frequent senses are added earlier on in the creation of an entry page and thus have a lower index (Meyer, 2013). For OmegaWiki, due to the structure of the database with mixed senses from different languages and no additional information, this information is not reliably inferable, as it is for FrameNet and Wikipedia, which is why this feature is not used in these cases. For Wikipedia, articles which cover the same lemma are usually listed on a disambiguation page, however, the order and grouping of the articles is motivated by coherence rather than frequency.

**Translation overlap** Wikipedia, Wiktionary and OmegaWiki all offer translations into other languages; thus, we can compare the number of translations which

two senses have in common. The assumption is that many common translations indicate equivalent senses. This feature was, for instance, suggested by Bond and Foster (2013) for aligning wordnets in different languages to Wiktionary.

**Example sentence patterns** For LSRs containing usage example sentences for senses, we investigated whether the examples which demonstrate the correct usage of word senses in context could be exploited to calculate the similarity between these senses. This follows the intuition that word senses are probably similar if they are used in the same context, a notion which, for instance, is the foundation of lexical substitution approaches (Cholakov et al., 2014a) and also for creating sense labeled corpora (Cholakov et al., 2014b).

As direct comparison of example sentences from different resources did not seem reasonable due to the presumably high lexical variation, we decided to use a more abstract approach and employed part of speech patterns, i.e. we only consider the sequences of parts of speech before and after the target senses of the example sentences. This is the approach employed by Cholakov et al. (2014b), and these patterns strike a balance between a fully lexicalized setting and a syntactic parse of the sentences, which would both be too specific representations for a meaningful comparison. A similar approach (dubbed “shallow frame structures”) was recently proposed by Caselli et al. (2013) for aligning verb senses in two Italian resources. While, obviously, part of speech tagging of the usage examples is necessary for this pattern-based approach, it still seemed a moderate effort.

## Discussion

Unfortunately, for none of the above features we could observe any significant improvement on the machine learning results when applying them in combination with gloss similarity and sense distance. Thus, we do not report the results here. While this was expected for part of speech and sense index, we investigated the other sets of features we tried in more detail .

For the translation overlap feature, the problem was mainly the coverage – only a small number of word senses from OmegaWiki, Wikipedia and Wiktionary share translations into the same languages, so that only very few instances of the gold standard were affected by this feature.

The example sentence pattern feature suffered not only from the lack of example sentences (especially in OmegaWiki, cf. Section 3.3), but also from the very large heterogeneity in the sentences provided due to the many degrees of freedom when composing a sentence. Although the patterns were supposed to provide a sufficient level of abstraction, this was not the case. The situation was further aggravated by the fact that, even for senses which do have example sentences, there are usually only a few of them. If these happen to differ from the examples in the other resource, the results of the pattern overlap are in many cases not significantly distinguishable from the overlap with a random pattern, which is of course not useful as a feature value. While more example sentences would certainly help (for instance for calculating the maximum similarity between elements in two sets of example sentences and using this as feature), out of the resources we covered only FrameNet had a substantial number of examples per sense. This was not enough to render this feature effective.

We consider the lack of impact of the additional features as an indicator that gloss similarity and distance in the resource graph already sufficiently capture the similarity between senses, making the generic approach we present in this chapter generally effective.

### 6.3.2 Machine Learning Classifiers

We experimented with different machine learning classifiers using WEKA (Hall et al., 2009). While a detailed discussion of these classifiers is beyond the scope of this work, we will at least give a short description of the ones we eventually used. For more details, please refer to textbooks such as Murphy (2012). We used WEKA's standard configuration in every case, avoiding the step of training hyperparameters – this again is in line with our goal of creating a flexible and generic alignment approach.

**Naive Bayes** classifiers assume that features are independent (i.e. the value of one feature is unrelated to the value of any other feature), and are thus able to learn reliable classification probabilities on relatively small training sets. While the independence assumption can be considered an oversimplification, the algorithm is widely used due to its efficiency and good precision.

**Bayesian Networks** (or *belief networks*) also classify based on probabilities learned from training data, however, they offer the advantage of modeling dependencies between features, hence allowing a more accurate representation of the data. Technically, such a network is a directed acyclic graph where the nodes are the variables and the edges are used for modeling the conditional dependencies between variables.

**Perceptrons** are classifiers which map a real-valued input vector to a binary output by means of an artificial neural network. Perceptrons are so-called online approaches, which adapt the model gradually when new training data is seen, and they are commonly used for pattern recognition, also in NLP (Collins, 2002).

**Support Vector Machines** (SVMs) construct a hyperplane in a multi-dimensional space which yields a good separation between positive and negative training examples represented as data points. Non-linear classification is also possible by transforming the feature space via kernels; this is not applied in our case, however.

**Decision Trees** are built from training input by iteratively splitting the set of samples based on their attribute values so that the resulting subset is as homogeneous as possible with regard to the class label. Unseen examples can be classified by testing the attribute values and following different branches of the tree. One of the main advantages (e.g. in comparison to SVMs) is that this approach is easily interpretable by human inspection.

### 6.3.3 Experimental Results and Analysis

#### Baselines

Apart from the naive baselines (Section 2.2.2), we report three additional reference results which have been discussed in previous sections: i) *SB*: A similarity threshold is learned for gloss similarity values as described in Section 4.2, ii) *DWSA*: The closest candidate sense in the resource graph is aligned as described in Section 5.3, iii) *HYB*: A hybrid approach of using *DWSA* first and then *SB* as a backoff, also described in Section 5.3.

#### Overview

Tables 6.3, 6.4, 6.5 and 6.6 present the results for the baselines and for the different machine learning setups. The joint approach outperforms the previous results for the hybrid approach and the baselines on four of the eight datasets in terms of F-measure, and at least achieves an improvement in precision on most of the other datasets. However, there is no consistent pattern in the results across different LSRs and classifiers. One reason is that the range of feature values varies substantially between different datasets. For instance, Dijkstra-WSA distances tend to be greater when Wikipedia is involved simply by its virtue of being larger than the other LSRs, and gloss similarities also differ depending on the average length of the glosses and the language (cf. Section 3.4.1). Another factor are the gold standards used, which are quite different in terms of size and composition (see Table 3.17). The different classifiers seem to be sensitive to this kind of variation so that none of them is the undisputed “winner”. However, Bayesian Networks proved most robust in our experiments, showing competitive results in every setup. As training them is also computationally inexpensive (compared to SVMs, for instance), we would generally recommend this kind of classifier for WSA tasks. In the following, we provide a more detailed discussion of the results for each individual dataset.

#### WordNet-Wiktionary

On this dataset, the recall of the alignment is satisfactory for every classifier, while especially Dijkstra-WSA struggles because of its issues with low graph density, exhaustively discussed in Chapter 5. The strength of the machine learning approach becomes apparent especially in comparison with the *HYB* approach: while the latter merely combines independent alignment decisions, hence achieving better recall but failing to improve precision (cf. Section 6.2), the joint usage of features leads to an improvement in both respects, especially for the Bayesian classifiers. Analysis of the decision tree classifier shows that, as we suspected, the “borderline cases” are explicitly reflected in the learned model, i.e. examples with high gloss similarity but also a high Dijkstra-WSA distance (or vice versa) are ruled out with higher confidence. This observation generally also holds for the other datasets. The combination of distances and gloss similarities is also able to alleviate the low density of Wiktionary to some extent, as examples with missing Dijkstra-WSA distance can still be aligned in case of sufficient gloss similarity. SVMs show the best precision here, but are challenged by the suboptimal separability of the feature space.



### WordNet-Wikipedia

In this setup, the machine learning approach is not able to achieve an improvement with regard to F-measure, as the *HYB* approach already achieves very good recall (0.87) with a still good precision; none of the classifiers can match these results. Nevertheless, as it can also be observed for several of the other datasets, the alignment precision can generally be improved, with the sole exception of the Naive Bayes classifier. The Decision Tree gives the best overall precision, and the manual analysis of the tree reveals that, as expected, the combination of sense distances and gloss similarities is able to rule out false positives with very high confidence. Dijkstra-WSA distances are especially informative in this case, as both WordNet and Wikipedia are densely connected (see Table 5.3), which yields almost no candidate pairs with infinite distances in the graph. Thus, the classifiers do not have to rely on gloss similarity alone to make the decision.

As mentioned before, for this data set we were able to compare our alignment to the full alignment contained in *BabelNet*. To make this possible, we extracted the WordNet synset identifiers as well as the Wikipedia article titles from the *BabelSynsets* which conflate WordNet synsets as well as Wikipedia senses. Although no explicit alignment information is given, we can safely assume that a *BabelSynset* containing both information from WordNet and Wikipedia constitutes an alignment between the corresponding concepts. Note again that this procedure was not possible for the other alignments contained in *BabelNet* due to the different representation of identifiers. The results for this alignment have been added to Table 6.3, and we can see that, while the precision is on par with the best machine learning approach, the recall is fairly low. One reason for this may be that this alignment was explicitly tuned for high precision to serve as a reliable foundation for the further construction of *BabelNet* (Navigli and Ponzetto, 2012a). Moreover, we also hypothesize that the exclusive reliance on the WordNet structure for creating an alignment might prevent this approach from achieving a higher recall, as our own experiments confirm that the structural properties of both resources are largely different (cf. Section 3.4.2). It would be interesting to compare the results on our data set to the ones originally reported by Navigli and Ponzetto (2012a), but due to the unavailability of the original gold standard this is not possible at this time.

### GermaNet-Wiktionary

For this pair, an improvement in F-measure can be observed, which does not come as a surprise considering the already very strong results of the *SB* and *HYB* approaches, although we can reach an improvement over the strong baselines (cf. Section 5.3.3) in some configurations. Interestingly, unlike for the other datasets, not even a significant improvement in precision can be achieved. While the Naive Bayes and Perceptron classifiers reach parity with the very strong precision of the similarity-based baseline (0.90), they do so at the cost of unacceptable recall. The Naive Bayes classifier seems to be especially challenged in this case, and manual analysis of the dataset revealed that this can be attributed to the simple gloss similarity measure we had to apply for this dataset. We used word overlap (cf. Section 5.3.3) instead of the more complex measures COS and PPR, which were not applicable in lieu of the originally used lexical fields and glosses. The uneven distribution of feature

	WordNet-Wiktionary				WordNet-Wikipedia			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>
Random	0.21	0.59	0.31	0.67	0.49	0.62	0.53	0.86
<i>1:1</i>	0.68	0.19	0.30	0.88	0.68	0.52	0.59	0.91
<i>1st</i>	0.33	0.51	0.40	0.80	0.51	0.72	0.60	0.88
<i>BabelNet</i>	–	–	–	–	0.85	0.31	0.47	0.90
<i>SB</i>	0.67	0.65	0.66	0.91	0.78	0.78	0.78	<b>0.95</b>
<i>DWSA</i>	0.68	0.27	0.39	0.89	0.75	0.67	0.71	0.93
<i>HYB</i>	0.68	0.71	0.69	0.92	0.75	<b>0.87</b>	<b>0.81</b>	<b>0.95</b>
SVM	<b>0.82</b>	0.61	0.70	0.93	0.81	0.67	0.73	0.94
Naive Bayes	0.71	0.79	0.75	0.92	0.71	0.82	0.76	0.94
Bayesian Network	0.70	<b>0.84</b>	<b>0.77</b>	<b>0.94</b>	0.77	0.78	0.78	0.94
Perceptron	0.74	0.72	0.73	0.92	0.76	0.81	0.78	0.94
Decision Tree	0.78	0.66	0.72	0.93	<b>0.86</b>	0.73	0.79	<b>0.95</b>
Agreement			0.78	0.93			0.87	0.97

Table 6.3: Alignment results for WordNet-Wiktionary and WordNet-Wikipedia: using baselines (top), approaches from previous work (middle) and different machine learning classifiers (bottom). Best results for each value and dataset are marked in bold. The annotator agreements  $A_0$  and  $F_1$  are given as plausible upper bounds. The results for the full alignment derived from *BabelNet* are given for comparison.

values for this overlap measure seems to make accurate modeling of the probabilities hard. Nevertheless, it stands to reason that using the same similarity measures as for the other datasets an improvement on precision might be possible, especially if the issues regarding morphology and compounds in German observed, for instance, in the German Wiktionary-Wikipedia dataset (see Section 5.3.3) can be effectively addressed.

### FrameNet-Wiktionary

Here, we observe that the precision is higher at the expense of recall, comparable for instance to the WordNet-Wikipedia and WordNet-OmegaWiki datasets. The machine learning classifiers are again significantly better at sorting out false positives by jointly considering the sense distances and similarities. However, the machine learning approach fails to increase the overall F-measure – as the *SB* and *HYB* already achieve good precision, the modest further increase via machine learning cannot make up for the loss in recall. This is only slightly disappointing, as the inter-annotator agreement suggests there is not much room for improvement anyway. It is still notable, though, that at least the overall accuracy is improved for the Decision Tree classifier. Considering this, in a scenario where precision is more important than recall (i.e. if we want to ensure that as many alignments as possible are correct), using the joint learning approach is still a viable option – this is, of course, also true for the other datasets where the precision can be improved.

	GermaNet-Wiktionary				FrameNet-Wiktionary			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>
Random	0.44	0.51	0.47	0.54	0.56	0.54	0.55	0.53
<i>1:1</i>	<b>0.95</b>	0.58	0.72	0.74	<b>0.90</b>	0.22	0.35	0.78
<i>1st</i>	0.77	0.81	0.79	0.79	0.66	0.64	0.65	0.81
<i>SB</i>	0.90	0.76	0.83	0.81	0.73	0.75	0.74	0.86
<i>DWSA</i>	0.85	0.72	0.78	0.76	0.73	0.31	0.44	0.78
<i>HYB</i>	0.83	<b>0.86</b>	<b>0.85</b>	<b>0.81</b>	0.77	<b>0.79</b>	<b>0.78</b>	0.88
SVM	0.87	0.77	0.82	0.79	0.82	0.65	0.72	0.86
Naive Bayes	0.91	0.48	0.63	0.76	0.80	0.65	0.72	0.86
Bayesian Network	0.84	0.80	0.82	0.79	0.79	0.66	0.72	0.85
Perceptron	0.90	0.65	0.75	0.74	0.78	0.70	0.74	0.86
Decision Tree	0.87	0.77	0.82	0.79	0.86	0.71	0.77	<b>0.89</b>
Agreement	-	-	N/A	N/A	-	-	0.80	N/A

Table 6.4: Alignment results for GermaNet-Wiktionary and FrameNet-Wiktionary.

	WordNet-OmegaWiki				Wiktionary-OmegaWiki			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>
Random	0.46	0.35	0.40	0.51	0.35	0.40	0.37	0.57
<i>1:1</i>	0.36	0.64	0.46	0.55	0.49	0.61	0.54	0.67
<i>1st</i>	0.34	<b>0.80</b>	0.48	0.47	0.41	<b>0.77</b>	0.54	0.57
<i>SB</i>	0.55	0.53	0.54	0.73	0.67	0.65	0.66	0.78
<i>DWSA</i>	0.56	0.69	0.62	0.74	0.58	0.51	0.54	0.72
<i>HYB</i>	0.57	0.75	0.65	0.75	0.68	0.73	<b>0.71</b>	0.80
SVM	<b>0.95</b>	0.32	0.48	0.79	<b>0.82</b>	0.54	0.65	0.81
Naive Bayes	0.73	0.62	0.67	0.82	<b>0.82</b>	0.53	0.64	0.81
Bayesian Network	0.75	0.72	<b>0.74</b>	<b>0.84</b>	0.79	0.64	<b>0.71</b>	<b>0.83</b>
Perceptron	0.73	0.58	0.65	0.81	0.79	0.60	0.68	0.82
Decision Tree	0.68	0.63	0.66	0.80	0.76	0.66	<b>0.71</b>	0.82
Agreement			0.84	0.85			0.80	0.85

Table 6.5: Alignment results for WordNet-OmegaWiki and Wiktionary-OmegaWiki.

### WordNet-OmegaWiki

For this dataset, the results look similar to WordNet-Wiktionary as far as the improvement of precision is concerned, since the joint usage of features helps to make a correct decision on borderline examples. As an example, the two senses of *genome* in biology (“The non-redundant genetic information stored in DNA sequences that defines an individual organism”) and algorithmics (“In the context of a genetic algorithm, the information that defines an individual entity”) have similar glosses; they are, however, quite far apart in the graph and thus not aligned. The Bayesian Network achieves the best results as it comprehensively models this interdependence of gloss similarity and sense distance. The SVM achieves the best precision, but the distribution of feature values does not lend itself well to linear separation in this case, leading to unsatisfactory recall. On a more general note, we observe the

highest absolute improvement in F-measure across all datasets, which is a strong indicator that even quite heterogeneous resources (with regard to their structure and gloss vocabulary) can be aligned with satisfactory results, if the available features are intelligently exploited.

### Wiktionary-OmegaWiki

The results for this dataset also fail to improve the overall F-measure – however, as it is the case for FrameNet-Wiktionary and WordNet-Wikipedia, the precision can be massively increased, which leads to a significant improvement in accuracy. The *SB* and the *DWSA* approach both struggle to achieve good precision due to the relatively low graph connectivity and low gloss quality of both resources. In conjunction both kinds of features work reasonably well, which again indicates that these features are sufficient if correctly combined. While the SVM and Naive Bayes classifiers excel in terms of precision at the expense of recall because of the not linearly distributed feature values, the Bayesian Network and Decision Tree classifiers manage to model the feature space more accurately and yield a very balanced result. Thus, as mentioned for the other datasets, the usage of machine learning seems advisable if precision is preferred over recall, i.e. if mostly correct alignments are desired.

### Wiktionary-Wikipedia (English)

As mentioned earlier, the low connectivity of Wiktionary is not as much an issue here as for WordNet-Wiktionary, as higher-frequency words were usually retained in the gold standard (see Section 3.5.2), which are more densely linked within Wiktionary. This leads to reasonable results for Dijkstra-WSA alone. The hybrid approach reaches the best recall, but due to the relatively low precision of the *SB* alignment, the overall result leaves room for improvement. This improvement is again achieved via joint modeling of features. As for the datasets discussed above, the precision is improved significantly; this is especially true for the Bayesian Network classifier. Precision and recall for the SVM classifier are also satisfactory in this case (due to the better linear separability of the feature space), making it the best overall classifier along with the Perceptron. In general, however, there is not much difference between different classifiers on this dataset.

### Wiktionary-Wikipedia (German)

As explained in Section 5.3.3, the naive baselines are already very strong, due to the bias towards positive examples on this dataset. Nevertheless, the *HYB* approach yields better results thanks to the high precision of its two components. Recall is insufficient for the *SB* setup as the richer morphology and compound formation in German impairs the reliability of the similarity values.

However, when machine learning is applied, the recall can again be significantly improved with minimal loss of precision; this is especially true for the Perceptron classifier. Here, as for several of the other datasets, the joint consideration of distance and gloss similarity allows correct alignment of more inconclusive examples.

	Wiktionary-Wikipedia en				Wiktionary-Wikipedia de			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>
<i>Random</i>	0.41	0.49	0.45	0.48	0.68	0.40	0.51	0.46
<i>1:1</i>	0.17	0.56	0.26	0.33	<b>1.0</b>	0.63	0.77	0.75
<i>1st</i>	0.23	<b>0.88</b>	0.36	0.37	0.93	0.66	0.78	0.74
<i>SB</i>	0.61	0.65	0.63	0.84	0.85	0.46	0.60	0.57
<i>DWSA</i>	0.78	0.57	0.66	0.88	0.82	0.74	0.78	0.71
<i>HYB</i>	0.62	0.81	0.70	0.86	0.86	0.77	0.81	0.75
SVM	0.82	0.70	<b>0.76</b>	0.92	0.76	0.84	0.80	0.71
Naive Bayes	0.77	0.72	0.74	0.91	0.85	0.54	0.66	0.62
Bayesian Network	<b>0.91</b>	0.63	0.74	<b>0.93</b>	0.86	0.81	0.83	0.77
Perceptron	0.82	0.70	<b>0.76</b>	0.92	0.75	<b>0.90</b>	0.82	0.73
Decision Tree	0.79	0.69	0.73	0.92	0.89	0.81	<b>0.85</b>	<b>0.80</b>
Agreement			0.79	0.95			0.85	0.89

Table 6.6: Alignment results for Wiktionary-Wikipedia for English and German.

	FN-WKT		WN-WKT		GN-WKT		WN-WP	
SVM	502	273	191	122	20,796	6,331	153	74
	111	1,903	42	2,063	3,036	15,473	37	1,551
Naive Bayes	505	270	247	66	13,091	14,036	187	40
	129	1,885	101	2,004	1,366	17,143	76	1,512
Bayesian Network	513	262	263	50	21,625	5,502	177	50
	139	1,875	113	1,992	4,021	14,488	53	1,535
Perceptron	545	230	225	88	17,486	9,641	183	44
	152	1,862	80	2,025	2,047	16,462	58	1,530
Decision Tree	548	227	207	106	20,795	6,332	166	61
	93	1,921	58	2,047	3,039	15,470	28	1,560

Table 6.7: Confusion matrices for datasets reported in previous work. The dataset sizes are stated in Table 3.17. For each cell, top left: true positives, top right: false negatives, bottom left: false positives and bottom right: true negatives.

While, intuitively, the strong bias towards the positive class makes this alignment task comparably easy, the observation that the only other comparably large dataset (GermaNet-Wiktionary) shows a similar distribution of positive and negative examples (see Table 3.17) at least suggests that this dataset is representative of a full alignment task, which is the eventual goal of WSA. Hence, the fact that our result still beats the strong baselines in terms of F-measure indicates that WSA, and especially our joint approach, works well on such a large-scale dataset, at least if the calculation of appropriate similarity measures is possible. This is the case for this pair, due to the already discussed density of the resource graphs and completeness of the glosses (see Section 3.4).

	WN-OW		WKT-OW		WKT-WP de		WKT-WP en	
SVM	67	143	103	87	18,472	3,372	57	24
	4	469	23	373	5,883	4,070	12	274
Naive Bayes	130	80	101	89	11,745	10,099	58	23
	48	425	23	373	2,014	7,939	18	268
Bayesian Network	151	59	122	68	17,598	4,246	51	30
	50	423	32	364	2,984	6,969	5	281
Perceptron	122	88	114	76	19,673	2,171	57	24
	45	428	30	366	6,554	3,399	12	284
Decision Tree	132	78	125	65	17,752	4,092	56	25
	62	411	39	357	2,218	7,735	15	271

Table 6.8: Confusion matrices for datasets we created in our work. The dataset sizes are stated in Table 3.17. For each cell, top left: true positives, top right: false negatives, bottom left: false positives and bottom right: true negatives.

## Summary

In summary, our experiments show that by modeling features in a joint machine learning approach further improvement of the WSA performance is possible. This again supports our claim that the graph-based and similarity-based alignment approaches are complementary, and that combining their benefits is superior to using them in isolation or in the simple backoff approach we presented earlier. In this way, we consistently observe better alignment decisions on borderline examples (i.e. examples where the feature values are contradictory). This includes examples where features are uninformative, because of missing or short glosses or nodes disconnected from the graph. This is especially an issue for collaboratively constructed resources (cf. Section 3.4). In this case, the two distinct approaches are obviously struggling to make correct decisions on their own, and the machine learning provides the means to still achieve satisfactory alignment results. A good example for this is the WordNet-OmegaWiki dataset, which proved challenging for both the similarity- as well as the graph-based approach: for this pair, we can improve the F-measure from 0.65 to 0.74, proving that even very heterogeneous resources can be effectively aligned by our joint approach.

The alignment results on the full resources for the best configurations (as for the Dijkstra-WSA results) is freely available as part of UBY (see Section 7.4) and on our website.<sup>2</sup>

## Error analysis

Generally speaking, the error sources for the joint approach are the same as for the individual approaches we discussed in the previous chapters (see Sections 4.4.2 and 5.3.4). This is inevitable as we exploit the same basic features. However, as we extensively discussed in the previous section, the machine learning is able to (at least to a certain degree) rule out erroneous or misleading feature values so that the results in general, and especially the precision, can be substantially improved.

<sup>2</sup><https://www.ukp.tu-darmstadt.de/data/lexical-resources/>

However, this is obviously not possible for cases when both types of feature are unreliable. For instance, if equivalent concepts are described very differently (known as the “lexical gap”, e.g. the senses “divulge confidential information” and “to confess under interrogation” of the verb *to sing*) and at the same time happen to be not very close in the resource graph, they will not be aligned which leads to false negatives (see Tables 6.7 and 6.8) and hampers further improvement of the recall.

Similarly, false positives (leading to decreased precision) occur for examples such as *Brand*, which is the name of districts in two different German cities (Aachen and Zwickau). The sense descriptions are very much alike, and the senses are also located in similar regions of the resource graphs (roughly speaking, *German geography*), which makes the distinction hard.

Addressing these issues might, on the one hand, be possible by computing more informative gloss similarity values. A sensible way would be using lexical expansion to enrich the glosses with semantically related terms (Cholakov et al., 2014a) or adapting existing measures like PPR for other alignment scenarios. In its original implementation, PPR uses WordNet as a foundation for calculating semantic similarity, and we pointed out on several occasions that this is not ideal in cases where it is not involved in the alignment, as WordNet neither reflects all senses nor lexemes which are present in other LSRs. This leads to unreliable PPR similarity values for these. Recently, Pilehvar and Navigli (2014) suggested exactly this idea of adapting the PPR algorithm to different resources, ridding themselves of this “WordNet bias” and achieving a substantial improvement of WSA performance in this way. The comparability of the resources despite the different graph structures is ensured by only considering the intersection of monosemous concepts in both resources (akin to our “trivial alignments” step) and calculating a ranking-based score over these using the weighting obtained by PPR. “Ranking-based” in this case means that not the absolute weights are considered for comparison, but the positions of the relevant concepts in lists sorted by these weights. In this way, differences in the weight values occurring due to different resource sizes and densities can be mitigated. Such a resource-agnostic approach would be especially interesting for German, as no comparable semantic similarity measure has been implemented so far. GermaNet seems ideally suited for this, although our analysis of the graph structure (see Section 3.4.2) suggests that the German Wiktionary is a promising candidate as well.

Improving the graph structure (and subsequently, the meaningfulness of distances) is the other possible avenue for obtaining better feature values. While using semantic relations is obviously useful for identifying related senses, monosemous linking is not entirely reliable, and more severely, only partially successful if glosses are missing or short, which is especially an issue for collaboratively constructed resources. For this, Pilehvar and Navigli (2014) propose to further enrich the graph structure by also considering polysemous lexemes in glosses for linking. They apply different similarity metrics and heuristics to determine if such a link is plausible, which leads to denser graphs and better WSA results. For comparison, their results are given in Table 6.9, and we can see that by using the described enhancements and the novel ranking-based similarity score, an improvement over our results is possible on the WordNet-OmegaWiki and WordNet-Wikipedia data sets. In these cases, their smart graph enrichment leads to high recall while maintaining high precision. For the WordNet-Wiktionary case the results are slightly lower than ours, and we

	WordNet-OmegaWiki				WordNet-Wiktionary				WordNet-Wikipedia			
	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>	<i>P</i>	<i>R</i>	<i>F</i> <sub>1</sub>	<i>A</i>
<i>SB</i>	0.55	0.53	0.54	0.73	0.67	0.65	0.66	0.91	0.78	0.78	0.78	0.95
<i>DWSA</i>	0.56	0.69	0.62	0.74	0.68	0.27	0.39	0.89	0.75	0.67	0.71	0.93
<i>HYB</i>	0.57	<b>0.75</b>	0.65	0.75	0.68	0.71	0.69	0.92	0.75	<b>0.87</b>	0.81	0.95
<i>ML</i>	0.75	0.72	0.74	0.84	<b>0.70</b>	<b>0.84</b>	<b>0.77</b>	<b>0.94</b>	<b>0.86</b>	0.73	0.79	0.95
PN14	<b>0.78</b>	0.73	<b>0.75</b>	<b>0.90</b>	0.67	0.80	0.73	0.93	0.85	0.84	<b>0.84</b>	<b>0.97</b>

Table 6.9: Our alignment results in comparison to the approach recently presented in (Pilehvar and Navigli, 2014), marked as PN14. *ML* stands for our best machine learning results as presented in this chapter. By further enriching the resource graphs, adapting the PPR measure and introducing a new ranking-based similarity score, they achieve improvements for the WordNet-OmegaWiki and WordNet-Wikipedia data sets, the latter of which is statistically significant. Note that the evaluation data sets were provided by us and thus are identical.

hypothesize that this is due to the many gaps and short glosses in Wiktionary. Even when considering polysemous lexemes, only few reliable hints for accurately enriching the graph structure can be found, so that our machine learning approach seemingly still caters best to this particular resource pair.

The last possible source of errors is, of course, the element we introduce in this chapter to combine the two different dimensions of sense similarity – the machine learning classifiers. As we pointed out in the previous section, different classifiers struggle on different datasets due to their different perspectives on the data. The SVM classifiers, for instance, show good results in many cases (and SVMs are also renowned for being effective in many other application scenarios), if the feature space is not easily linearly separable however, the recall drops considerably. This is especially striking for the WordNet-OmegaWiki and Wiktionary-OmegaWiki datasets. While for the former, the gloss similarity values are not optimal due to the low lexical overlap between glosses (cf. Section 3.4), we observe many missing or high Dijkstra-WSA distances for the latter. The SVM classifier is seemingly sensitive to this, and a possible solution might be the application of kernels to allow a non-linear classification (Scholkopf and Smola, 2001). Moreover, the relatively weak results on the same two datasets show that Perceptron classifiers are also not well-suited in this case.

The generally good results of the Bayesian Networks are not surprising as, intuitively, the similarities of glosses and the distances in the graphs are not independent – after all, this is the main idea we base this joint approach on. Thus, the explicit modeling of this dependency is beneficial, which is proven by the fact that the Naive Bayes classifiers, which basically work with the same probabilities but consider them independently, fail to outperform Bayesian Networks in every setup. These dependencies between feature values are only implicitly modeled in Decision Trees, but this is seemingly still enough for these classifiers to be effective in most cases.

As a final remark, we would like to point out that, although we have described several flaws in the features and classifiers which merit further investigation, the achieved accuracy is close to the human inter-annotator agreement in the majority of cases, which suggests that, at least for the datasets considered, there is not much head room for major improvements.



## 6.4 Experiments on N-way Alignment

Motivated mainly by the construction of the graph framework described in Chapter 5, and also by the automatic creation of the Wiktionary-Wikipedia gold standard we described in Section 3.5.2, a further direction for our research is N-way alignment, i.e. joint WSA with more than two LSRs at the same time. The intuition behind this is that, generally speaking, the more information is available, the more informed an alignment decision could be made. However, we still do not want to abstain from our guiding principle of general applicability and flexibility of the algorithms, so that we focus on the same basic features we discussed thus far, and especially on the graph representation of LSRs – theoretically, we can combine as many graphs as desired without impairing the applicability of the underlying principles we exploited for our pairwise approaches. We investigate a couple of ideas in that direction, which we will briefly describe in this section.

### 6.4.1 Using Existing Alignments

A very straightforward idea is to construct an N-way alignment out of existing pairwise alignments between LSRs; a construction of 3-way alignment between WordNet, Wikipedia and Wiktionary from the alignments we discussed earlier was investigated by Miller and Gurevych (2014) from a set-theoretic perspective. However, the results are disappointing in the sense that the accuracy of the derived alignment is rather low, as errors are transitively magnified. This is in line with our observations for the transitive construction of the English Wiktionary-Wikipedia gold standard dataset using WordNet as pivot (Section 3.5.2) and the preliminary experiments for automatic construction of a WordNet-OmegaWiki alignment (Section 4.4). Another issue is that such an alignment construction approach is only possible for lexemes which are contained in all of the resources involved (i.e. in the intersection), which is a strong constraint considering the small lexical overlap between particular pairs of resources (cf. Section 3.3). Thus, this approach seems not reasonable for producing alignments of acceptable quality and size.

### 6.4.2 Considering Multiple Graphs at once

A direct extension of Dijkstra-WSA is to include more than two resource graphs at the same time, i.e. calculating trivial alignments between all of them and then finding pairwise shortest paths, but with allowing “detours” across other resources. However, while a moderate increase in recall can be observed, there is no significant impact on the overall results regarding F-measure; our investigation of the dataset shows that for most pairs the “direct path” between the two LSRs is the shortest one, so that adding additional edges has little effect.

### 6.4.3 Clustering Word Senses for a Lemma

The inspiration for this approach is the notion we discussed at length in the past chapters: senses which are related, or more specifically, equivalent, are by some measure of similarity identifiable as such. The options we focus on in this thesis are gloss

similarity and distance in the graph representations, and this works reasonably well for the binary decision we discussed thus far, i.e. decisions if a pair of two senses is equivalent or not. If this idea is pursued further, it is obvious that senses which have a high similarity score can be grouped together into clusters, even across resource boundaries. This is also the rationale for the application of WSA we describe in Section 8.1.

To leverage this intuition for WSA, we employ the following setup for N-way alignment ( $N > 2$ ):

- For a specific lexeme, extract all senses from all LSRs involved. This is similar to the candidate extraction process described earlier. We call this set of sense the *initial sense graph* for a lexeme.
- For all the senses in this initial graph, calculate a similarity score between them. The similarity values serve as edge weights.
- Run a graph clustering algorithm which groups highly similar senses into smaller, strongly connected components. For these components, pairwise sense alignments are created between all senses they include, as these intuitively represent the same meaning. The expected benefit from this approach is that, by means of transitive closure, alignments can be found which would usually have been disregarded due to insufficient pairwise similarity.

This largely follows the approach which was first described by Kirschner (2012), which is why we limit ourselves to a brief discussion. Our main extension in this respect is the employment of the more elaborate similarity values based on Dijkstra-WSA distance and PPR, while the original framework only considered COS similarity.

Note that the final step (creation of pairwise alignments from the clustered senses) is inherently necessary for the evaluation. For pairwise WSA numerous gold standards and established evaluation metrics exist, while this is not the case for sense clusters. An option we considered was to create “gold clusters” by leveraging the existing gold standards, i.e. by exploiting the fact that different gold standards are lexically overlapping and alignments can thus be transitively inferred. This approach, however struggles with the issue that the lexical overlap can be extremely small depending on the (number of) resources involved (Miller and Gurevych, 2014). Moreover, even if such gold clusters are manually defined, it is an open question how an automatically created clustering can be meaningfully labeled “correct” or “incorrect” in relation to the gold standards. If the results are absolutely identical, this is obviously a true positive, but if only part of the answer is correct (for instance, 6 out of 10 items overlap with the gold standard), it is debatable if (and to what extent) this can be considered correct. A preliminary study on a small sample also showed that human annotators largely disagree on this notion, i.e. the annotator agreement was extremely low when allowing to mark clusters as correct that were only partially overlapping with a gold standard one. Thus, we default to using the existing pairwise gold standards, while the definition of a comprehensive and meaningful notion of correctness of sense clusters would be an interesting subject for future work.

## Experiments

For evaluating the clustering approach described below, we experiment with 3-way alignment of WordNet, Wiktionary and OmegaWiki. This combination is convenient as we have access to gold standards for all three pairwise combinations, as well as similarity scores between them calculated for all relevant measures, i.e. Dijkstra-WSA distances as well as COS and PPR gloss similarities, which we use for edge weights.

However, there are two problems which need to be addressed:

- Combining multiple similarity scores: only one edge weight can be applied for a pair of senses. While this was no problem for the original framework as only COS similarity was used, we have to find a way to merge similarity measures into a single value in a meaningful way. Following the approach successfully applied by Bär et al. (2012) for text reuse detection, we employ a linear regression approach to combine the feature values.
- Score normalization: as already briefly described in Section 6.3.3, and also by Kirschner (2012), different similarity measures show a different distribution of values depending on the resource pair. Simply speaking, a cosine similarity of 0.5 between WordNet and OmegaWiki must not be interpreted in the same way as the same score between WordNet and Wiktionary, due to different gloss lengths etc. Thus, to be able to compare these values, we normalize them by first calculating the average similarity for 1000 samples of each resource pair, and then adapting the average  $a$  to 0.5. Formally, for an average similarity value  $x$ , we calculate  $a$  so that the result of  $f(x) = \log_2(x^a + 1)$  is 0.5.

For the actual graph clustering, we employ two different algorithms, which are also described in detail by Kirschner (2012). While there exist many other graph clustering algorithms, we limited ourselves to these ones as they are intuitively understandable and commonly used.

1. Hierarchical Agglomerative Clustering (HA) (Jain et al., 1999) initially considers each sense as one cluster and iteratively groups clusters with similarity above a certain threshold which needs to be provided as a parameter for the algorithm. After each step, similarities between each cluster are recalculated, where usually the average of the old values is used to obtain the new edge weight. We also use this setup.
2. Newman Clustering (NC) (Newman, 2004) follows the inverse approach of considering the whole initial graph as one cluster, and iteratively dividing it into smaller clusters based on the betweenness measure of singles nodes, i.e. the number of pairwise shortest paths between all nodes that lead through these particular nodes. Nodes with a high betweenness are considered to be good “breaking points” for a graph into smaller subgraphs. The stopping criterion for the betweenness measure is also a parameter which needs to be provided.

For the evaluation, we use four different setups regarding similarity values: COS only, PPR only, Dijkstra-WSA only, and a combination of all three measures using

linear regression. For the clustering algorithms, we iteratively increase the threshold values for stopping in a range of reasonable values.

However, no matter which configuration or algorithm is used, no improvement over the different pairwise approaches we presented earlier can be achieved. On the contrary, even for the best configuration, the results are significantly worse than for the pairwise similarity-based setup. We achieve an F-measure of 0.66 for WordNet-Wikipedia, of 0.65 for WordNet-Wiktionary and of 0.59 for Wiktionary-Wikipedia using the combined similarity measure and HA clustering. This is in line with the results that were reported by Kirschner (2012). However, he only compares against a naive baseline and not the state of the art as we do. As these results are disappointing, we refrain from exhaustively listing all configurations and results here, but we still briefly explain the source of the errors.

A major source of false positives are the salient errors described earlier which are inherent to the LSRs and hard to address, also by this approach (too similar glosses/too small distances) – if a similarity measure produces a high value for a pair, it will almost inevitably end up in the same cluster, i.e. the precision can not be increased by using this clustering approach. The combination of different similarity values by linear regression is not as helpful as expected, while in any case it seemed implausible to reach higher precision than the pairwise machine learning approach as the same features are used as input.

However, while for the pairwise alignment only singular instances are affected by this, a single edge with too high a weight in the clustering setup can, in the worst case, connect two clusters (with sizes  $n$  and  $m$ , respectively) which are otherwise unrelated, resulting in  $n \times m$  incorrect alignments, i.e. false positives. In other words, the “noise” introduced by a single edge is magnified and multiplied, resulting in error propagation. The recall is increased using this approach, as some originally disregarded alignments can be recovered using the additional information provided by more than one resource (which was the main objective of this approach), but this by no means outweighs the observed loss in precision. Only for very restrictive threshold values, the precision can be brought to levels which are on par with the pairwise approaches, but then the recall is unacceptably low.

Another issue, which is however not as pertinent, is the different range of similarity values depending on the LSRs. As stated above, we address this problem by normalization, but it stands to reason that the values are still not fully comparable, which further impairs the discovery of meaningful clusters.

In summary, the clustering approach as it is presented here does not seem to be a promising direction for N-way alignment. The precision cannot be increased, and the possible gain in recall is negated by the disproportional propagation of alignment errors. Our intuition is that this is not easily addressable using the similarity measures we discuss here and which are commonly used, so that an entirely different approach might be required to leverage the additional knowledge provided by multiple resources. A possible idea would be to not use N-way alignment (or more specifically, the clustering approach) in isolation, but to combine it with the pairwise alignment methods presented in the previous chapters, e.g. as a fallback for cases which suffer from low confidence. In this way, the clusters would provide an additional source of information for making an alignment decision which takes more than one LSR at a time into account. Such information might also directly be used

in the machine learning approach. For instance, the information that two senses in resources  $A$  and  $B$  share a strong resemblance with another sense in resource  $C$  could be expressed by an additional feature.

## 6.5 Chapter Summary and Contributions

In this chapter, we have shown that through joint modeling of different similarity measures for WSA, the overall alignment quality in terms of F-measure can be significantly improved over the state of the art for four of the considered eight datasets; on the three others, we could at least significantly improve the alignment precision, reaching an improvement in accuracy. This proves that such a joint usage of the structural similarity as well as the gloss similarity of the LSRs is indeed preferable over using either of them in isolation or combining them in a simple backoff approach. The joint approach effectively covers the dependencies of both kinds of similarity. This is especially true if applications are addressed that require high precision, i.e. that are sensitive to too many wrong alignments. An example is WSD, which benefits from aligned senses if the alignment is accurate, but is impaired in case of too noisy information (Cholakov et al., 2014b).

Our experiments not only indicate that our features are effective, but also that they are sufficient. Experiments with additional features did not yield further improvements, which supports our major claim that structure and glosses, if cleverly used in combination, are enough to achieve satisfactory alignment quality. By means of machine learning, we push the resulting F-measure well above 0.70 on each of the considered datasets, getting close to the upper bounds implied by the inter-annotator agreements, while the features we apply are commonly available and not dependent on a particular LSR or usage scenario. While the heterogeneity of the LSRs with regard to their structure (density, path lengths) and content (gloss lengths, vocabulary) makes this task challenging, we see that the machine learning approach is able to alleviate these issues and provides a powerful tool for aligning all kinds of LSRs in the future – this is a central insight with regard to the previous work concerning WSA or related tasks such as ontology matching, which usually make much more restrictive assumptions about the input data, leaving the corresponding algorithms proprietary.

In further experiments, we also investigate different efforts in aligning more than two resources at a time, i.e. N-way alignment. However, we do not achieve an improvement in alignment performance, due to disproportional magnification of alignment errors when more than two resources are involved and mismatches in resource coverage.

In summary, our contributions in this chapter are as follows:

**Contribution 1** We present a joint machine learning approach for WSA which effectively combines gloss similarity as well as sense distance to improve upon state-of-the-art performance (and especially precision) on the considered datasets, proving that these features are sufficient for achieving satisfactory WSA results on heterogeneous resources.

**Contribution 2** We perform several experiments on N-way alignment and discuss

properties of the approaches and the resources which prohibit successful alignment of more than two resources at once.

# Chapter 7

## Putting the Pieces Together: UBY – A Unified Lexical-Semantic Resource

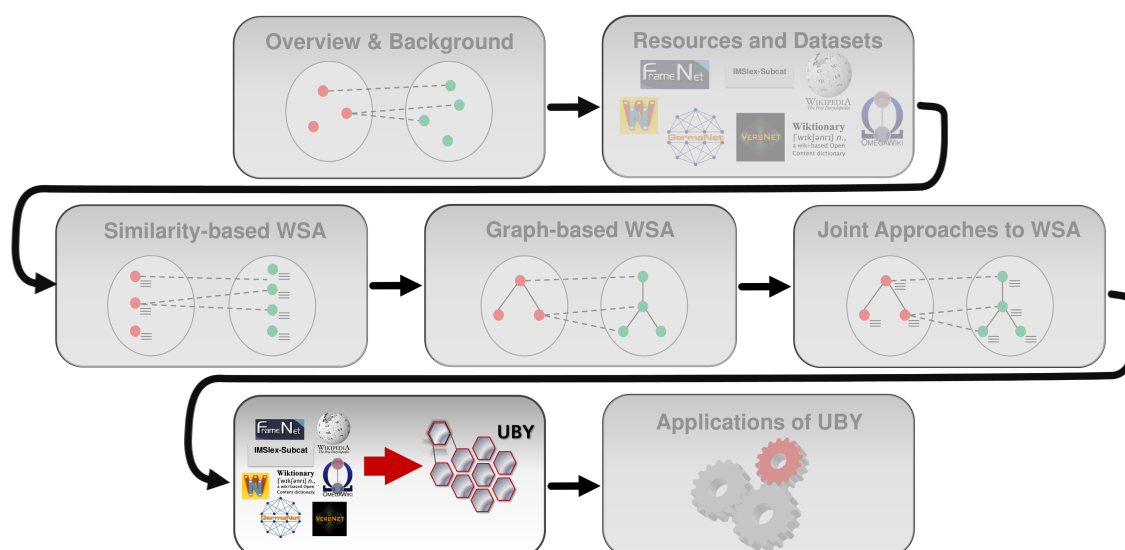


Figure 7.1: Visual outline of the thesis.

### 7.1 Introduction

In Chapter 3, as well as in the introduction, we already briefly introduced UBY – a lexical-semantic resource which was designed and constructed to subsume multiple, previously independent and structurally different resources to allow easy access and synergy effects by using those resources in combination. While we already discussed the resources that are contained in UBY and the motivation for including them, as well as the computation of alignments between these, we now want to focus on the design of UBY itself. To this end, we will briefly discuss previous efforts in standardizing, combining and accessing resources in Section 7.2, and then present the unique features of UBY:

- The structure of UBY is determined in a Lexical Markup Framework (LMF)-based model (Francopoulo et al., 2009) for large-scale multilingual LSRs called UBY-LMF. It models the lexical-semantic information down to a fine-grained level of information (e.g. syntactic frames) and uses standardized definitions of linguistic information types from ISOcat as proposed in the LMF standard. UBY-LMF covers various types of lexical-semantic information from all the heterogeneous resources presented earlier, and thus also accommodates multiple languages. We will present in Section 7.3 the general idea and layout of UBY-LMF, and illustrate how LSRs are mapped to UBY by using OmegaWiki as an example. Ensuring an appropriate (i.e. lossless) representation of OmegaWiki in UBY-LMF and the subsequent mapping of OmegaWiki to this standard are our main contributions in this respect.
- Apart from the unified representations, the already extensively discussed sense alignments between different resources are the other outstanding feature of UBY. These are provided to enable resource interoperability on the sense level, e.g. by providing access to the often complementary information for a sense in different resources. We will discuss how these alignments are represented in UBY-LMF (Section 7.3) and also discuss the alignments which were integrated into the first and current versions of UBY (Section 7.4).
- After discussing how resources are mapped to the UBY-LMF format, we present the result of the integration effort: the actual resource UBY. We especially focus on the database design which reflects the UBY-LMF classes, and also describe how the results of our work are made available to the research community (Section 7.5). The chapter and our contributions are summarized in Section 7.6.

Accompanying UBY, we also developed a Java-based API which allows easy and fast access to the information contained in the unified resource and a web interface which allows one to browse the resource without the need to set up the UBY database or access it programmatically. We will describe the implementation of the API and the interface in Appendix A.2. There, we also discuss how the standardization of LSRs in UBY and the API enabled the construction of the generic WSA framework we used for our own alignment experiments (Appendix A.2.3).

UBY is a group effort, to which many people contributed on many levels (Eckle-Kohler et al., 2012; Gurevych et al., 2012; Eckle-Kohler and Gurevych, 2012; Hartmann and Gurevych, 2013). Thus, we will focus on presenting our own identifiable contributions to UBY in detail, while presenting the other information about UBY in a concise but complete manner in order to make the context and the importance of our contributions understandable.

## 7.2 Related Work

As we already discussed, there have previously been several independent efforts to combine existing LSRs to enhance their coverage concerning the number of lexical items, and the types of lexical-semantic information contained (Shi and Mihalcea,



2005; Johansson and Nugues, 2007; Ponzetto and Navigli, 2010). However, as these efforts often targeted particular applications, they focused on aligning selected, specialized information types and frequently, the presented models were thus not easily scalable and applicable to other LSRs. The previous work also lacked the aspects of lexicon format standardization and API access, which we deem crucial to ensure acceptance and broad applicability in NLP.

The shortcomings are what motivated the design of UBY, and we will now briefly discuss the related work for the areas which have influenced its creation, with the exception of automatic WSA efforts; these already have been discussed separately in the respective chapters.

**Standardization of resources** Recently, multiple standards have been developed to represent dictionaries and language resources. Apart from LMF, which we will describe in the next section, the most important ones are the *Resource Description Framework* (RDF), the *Web Ontology Language* (OWL) and the *Text Encoding Initiative* (TEI).

The TEI standard was first established in 1987 and provides a set of formats and guidelines for exchanging texts. It is based on XML and is mainly used for modeling common annotation schemes for a variety of texts including, but not limited to, dictionaries. However, although a large part of TEI specifically addresses the standardization of dictionaries, its recommendations focus on running text as it is found in printed dictionary articles. This limits the modeling possibilities for LSRs such as WordNet, which have a graph-based, non-linear structure.

RDF and OWL are based on the notion of triples of the form (subject, predicate, object) which encode formal statements about concepts and their relations, and have emerged in the context of the Semantic Web (Berners-Lee et al., 2001). While many Semantic Web resources focus on world (i.e. encyclopedic) knowledge, initiatives such as the *Open Linguistics Working Group* aim at representing LSRs and other language resources in this format to build a *Linguistic Linked Open Data* cloud (Chiarcos et al., 2012).

One of the earliest works using RDF (De Melo and Weikum, 2008b) integrates a set of information types extracted from Wiktionary into the RDF web service *Lexvo*. Because of the differences between the different Wiktionary language editions, *Lexvo* is limited to a small subset of possible information types, mostly translations. This issue has been recently addressed by Hellmann et al. (2013) with a more comprehensive RDF representation of Wiktionary. An open research question in their work is, however, how lexical resources with differing schemata can be linked, as RDF does not provide a fixed set of predicates and literals for representing the necessary information types for LSRs. Recommendations such as *SKOS* (Miles et al., 2005) and the *LexInfo* model (Buitelaar et al., 2009) have been proposed to fill this gap, but only recently McCrae et al. (2011b) propose *lemon*, a conceptual model for lexicalizing ontologies, thus operationalizing *LexInfo* and providing an LMF implementation in OWL. The goal of *lemon* is to support the linking between ontologies and lexicons.

Regarding other implementations of LMF, the focus was mostly on single resources and information types. For instance, Soria et al. (2009) define *WordNet-LMF*, an LMF model for representing wordnets used in the *KYOTO* project as well as the *Open Multilingual Wordnet* project (Bond and Foster, 2013). Henrich and

Hinrichs (2010) do this for GermaNet and Toral et al. (2010) for Italian. These models are similar, but they present different implementations of the LMF meta-model, which impairs interoperability between the resources. UBY, on the other hand, goes beyond modeling a single LSR and aims at representing a large number of very heterogeneous resources in the same format. Also, UBY is one of the first efforts for completely modeling collaboratively constructed resources in a standardized format. While Serasset (2012), for instance, proposes a transformation of Wiktionary to LMF, this transformation does not include all information encoded in Wiktionary – translations are, for example, modeled at the level of words rather than at the level of word senses. The same holds for the approach proposed by McCrae et al. (2012), who focus on linking lexical information to ontologies and hence model only a small part of Wiktionary’s lexicographic information in their LMF model. Declerck et al. (2012) represent Wiktionary data using the TEI standard, and although their model is able to represent translations and many other lexicographic information types found in Wiktionary, the model does not contain information such as pronunciations. We are not aware of any works representing OmegaWiki in a standardized model.

**Large-scale integration of resources** Most previous efforts on integration of resources targeted encyclopedic (not lexical-semantic) knowledge. Prominent examples are *YAGO* (Suchanek et al., 2007) and *DBPedia* (Bizer et al., 2009).

Atserias et al. (2004) present the *Meaning Multilingual Central Repository* (MCR) which integrates five local wordnets based on the manually defined Interlingual Index of *EuroWordNet* (Vossen, 1998). It is restricted to a single type of resource, however, and features only a single type of lexical information (semantic relations) specified upon synsets. The same is true for the *Open Multilingual Wordnet*, which integrates the MCR as well as over 20 additional wordnets. Most of them are only linked to 5,000 “core concepts” of the Princeton WordNet, though, which represent only a fraction of its content (Bond and Foster, 2013) – this impairs the usability of this resource.

De Melo and Weikum (2009) create a multilingual wordnet by integrating wordnets, bilingual dictionaries and information from parallel corpora. They also do not integrate lexical-semantic information, such as syntactic subcategorization or semantic roles. McFate and Forbus (2011) present *NULEX*, a syntactic lexicon automatically compiled from WordNet, Wiktionary and VerbNet. Their goal is to create an open resource for syntactic parsing. Thus, they use only a small part of the lexical information present in each resource.

Padró et al. (2011) present their work on lexicon merging within the *Panacea* Project. One goal of *Panacea* is to create a lexical resource development platform that supports large-scale lexical acquisition and can be used to combine existing lexicons with automatically acquired ones. To this end, Padró et al. (2011) explore the automatic integration of subcategorization lexicons. Although they mention the LMF standard as a potential data model, they do not make use of it.

As already mentioned in the introduction, Shi and Mihalcea (2005) integrate FrameNet, VerbNet and WordNet, and Palmer (2009) presents a combination of Propbank, VerbNet and FrameNet in a resource called *SemLink* in order to enhance semantic role labeling. Similar to our work, multiple resources are integrated, but

the formats are proprietary, as is the sense alignment effort, which is mostly based on manual labor.

Lastly, one of the most comprehensive approaches to resource integration thus far, which also covers encyclopedic information, is *BabelNet* (Navigli and Ponzetto, 2012a). This resource (which has been mentioned on several occasions throughout this thesis) integrates WordNet, *Open Multilingual Wordnet*, Wikipedia, OmegaWiki, Wiktionary and Wikidata. The result is a multilingual network (created for large parts via automated alignment, see Section 6.2) containing over 9 million entries and covering 50 languages; it has been used for tasks such as WSD (Navigli and Ponzetto, 2012d), creation of semantic predicates (Flati and Navigli, 2013) and semantic relatedness computation (Navigli and Ponzetto, 2012c). The rationale behind the construction of *BabelNet* is to not only combine knowledge from the integrated resources, but also enrich the knowledge representation for better recall, at the possible expense of precision. For instance, if lexicalizations (i.e. synonyms or translations) of an existing concept are missing, machine translation of sense-annotated text snippets is used to fill these gaps, and additional relations within a single resource are inferred by using other resources as a “scaffold” (Flati et al., 2014). The integration of concepts is also taken one step further than for resources like UBY as all information is encapsulated into so-called *BabelSynsets*. While this makes it straightforward to use the combined knowledge, the selective use of particular resources or the restriction to knowledge originally contained in the resources (i.e. knowledge that was not automatically added in the enrichment steps) is not easily possible, as this distinction is not made transparent in the data model. This is one of the main issues we aimed at when creating UBY in order to produce a unified resource which can be used as a wholesale replacement for the LSRs it contains while at the same time providing enriched sense representations.

*BabelNet* is freely available as part of the aforementioned *Linguistic Linked Open Data* cloud (Ehrmann et al., 2014). The corresponding data representation is based on the *lemon* lexicon model (McCrae et al., 2011b) which ensures interoperability to other resources and easy integration into applications. A similar representation based on *lemon* was recently proposed for a subset of the resource UBY we discuss in this chapter (Eckle-Kohler et al., 2014). However, as we were not directly involved in the creation of this *lemonUby*, we omit a detailed discussion and comparison of these lexicon models. It suffices to say that both models ensure compatibility to other resources by focusing on the core elements of *lemon* (lexical entries, senses and synsets) and the relations between them while omitting certain resource-specific peculiarities.

**Programmatic access to resources** An important factor to the success of resources in NLP research is a public API, which facilitates the access to the information. Prominent examples are, for instance, the Java WordNet API,<sup>1</sup> the Java-based Wikipedia API,<sup>2</sup> the Wiktionary API<sup>3</sup> and our own JOWKL API which allows access to OmegaWiki (see Appendix A.1). As an example for accessing a linked resource, Navigli and Ponzetto (2012e) present the *BabelNet API*, which is specifically tailored

---

<sup>1</sup><http://sourceforge.net/projects/jwordnet/>

<sup>2</sup><http://code.google.com/p/jwpl/>

<sup>3</sup><https://code.google.com/p/jwktl/>

towards easy use in WSD applications. Following this, a major design objective of UBY is to create such an API, in spirit of Pradhan et al. (2007), who present integrated access as a main goal of their work on standardizing and integrating corpus annotations in the *OntoNotes* project.

**Interfaces for resources** Web interfaces have been traditionally used for electronic dictionaries, such as the *Oxford Dictionary of English* or the *American Heritage Dictionary*. Lew (2011) reviews the interfaces of the most prominent English dictionaries. These interfaces have also largely influenced the development of web interfaces for LSRs, such as the ones for WordNet, FrameNet, Wiktionary, or the recently presented *DANTE* (Kilgarriff, 2010) which directly adapted the dictionary interface models. Two other examples for accessing WordNet are *Visuwords*<sup>4</sup> and *WordNet explorer*<sup>5</sup> that allow browsing of the WordNet synset structure. An example for a cross-lingual graph-based interface is *VisualThesaurus*<sup>6</sup> which shows related words in six different languages. All of these interfaces have been designed in adherence to a specific, single LSR. Only a few interfaces are able to display information from multiple LSRs. The majority of them are limited to show preformatted lexical entries one after another without making any attempt to connect them. Popular examples are *Dictionary.com*<sup>7</sup> and *TheFreeDictionary*.<sup>8</sup> Similarly, the *DWDS* interface (Klein and Geyken, 2010) displays its entries in small rearrangeable boxes. The *Wörterbuchnetz* (Burch and Rapp, 2007) is an example of a web interface that connects its entries by hyperlinks – however, only at the level of lemmas and not word senses. The *BabelNetExplorer* (Navigli and Ponzetto, 2012b) enables access to the semantic network in *BabelNet*, but it does not allow determining the source of the information or obtaining additional knowledge about the presented senses.

In summary, related work mostly focuses either on the standardization of single resources (or a single type of resource), or on the integration of several resources in proprietary and heterogeneous formats. Collaboratively constructed resources have received little attention in previous work on resource standardization, and the level of detail of the modeling is insufficient to fully accommodate different types of lexical-semantic information. On top of that, complete API or UI access is rarely provided for integrated resources, which follows immediately from the heterogeneous ways in which access to the single LSRs has been realized. As a result, the comparative exploration of different LSRs, and in particular, of sense-aligned LSRs in order to assess their quality and usefulness for particular tasks is not easy in practice; neither is their orchestrated usage. This makes it hard for the community to exploit these LSRs on a large scale, diminishing the impact that these projects might achieve.

---

<sup>4</sup><http://www.visuwords.com>

<sup>5</sup><http://faculty.uoit.ca/collins/research/wnVis.html>

<sup>6</sup><http://www.visualthesaurus.com>

<sup>7</sup><http://www.dictionary.com>

<sup>8</sup><http://www.thefreedictionary.com>

## 7.3 The Lexical Markup Framework and UBY-LMF

Our analysis in Chapter 3 showed that LSRs represent encoded lexicographic information in different ways, with regard to:

- the structure. For instance, Wiktionary is built around wiki pages describing a lexical item, while resources like WordNet and OmegaWiki are centered around synsets.
- the encoded information types: Wiktionary encodes pronunciations and etymologies, while OmegaWiki encodes, for instance, ontological relations and FrameNet encodes semantic arguments.
- the coverage and granularity: Different resources cover different languages, lexemes and senses.
- the terminology: Different terms are used to refer to the same things. The Wiktionary term *sense definition*, for instance, corresponds to the term *gloss* used for WordNet and *paraphrase* used for GermaNet.
- the data format: Wiktionary is released as an XML database dump, WordNet is shipped in an idiosyncratic database, and OmegaWiki is available as an SQL database dump (see Section 3.3).
- the access paths: Different interfaces offer different search options for human users, and multiple software tools are available for accessing an LSR from applications.

In order to make use of the data in the LSRs, it is necessary to harmonize their heterogeneous representations and thus make them interoperable. Interoperability is a prerequisite for a smooth integration of resources into applications and for making them accessible in a unified way.

Ide and Pustejovsky (2010) distinguish *syntactic interoperability* and *semantic interoperability* as the two types of interoperability of computer systems. The former addresses the degree of the heterogeneity of the formats used to store and retrieve the language data. The latter represents the reference model for interpreting the language data. In terms of lexical resources, we need a structural model for storing and retrieving the lexicographic data and a set of standardized information types for encoding it. For this purpose, the ISO standard *Lexical Markup Framework* (LMF) (ISO24613, 2008), a standard with a particular focus on lexical resources for natural language processing (Francopoulo and George, 2013), is an obvious choice. It has emerged as a result of multiple projects such as *ACQUILEX*, *EAGLES/ISLE*, *MILE*, and *PAROLE*. As mentioned above, LMF has proven useful for modeling wordnets (Soria et al., 2009; Henrich and Hinrichs, 2010), but has only rarely been used for representing collaboratively constructed resources. One of the main challenges of the integration work is thus to develop a model based on LMF that is standard-compliant, yet able to express the information contained in diverse LSRs, and that also supports the integration of new resources.

LMF is a meta-model for lexical resources which is usually expressed using the *Unified Modeling Language* (UML). That is to say, LMF introduces a number of *classes* and *relationships* between them. The classes are organized in multiple packages (called *extensions*) that may be chosen according to the type of resource that is to be modeled. These include, for instance, the *Morphology* extension, the *Machine readable dictionary* extension, the *NLP syntax* extension and the *NLP semantics* extension. Each package provides a number of predefined classes and relations, and the *core* package represents the essence of the standard and is to be used for each instance of LMF. It includes, among others, the `LexicalEntry` class for modeling lexical entries in accordance to dictionaries, the `Form` class for representing different orthographic variants of a lexical entry, and the `Sense` class for modeling one of multiple possible meanings of a lexical entry. This, for most part, corresponds to the definitions we have introduced in Section 1.2.

Since LMF is conceived as a meta-model, the standard does not state which classes and attributes should be used to encode the language data in the resources. This is defined by the actual *lexicon model* – that is, an *instantiation* of the LMF standard. Developing a lexicon model such as UBY-LMF thus involves first selecting appropriate classes (e.g. `LexicalEntry`) from the LMF packages, second defining attributes for these classes (e.g. `part of speech`), and third linking the attributes and other linguistic terms (such as the attribute values `verb`, `noun`, `adjective`) to a well-defined reference source, in our case Data Categories (DCs) selected from ISOcat<sup>9</sup>, a Data Category Registry compliant with the ISO standard 12620 (Broeder et al., 2010) where a large amount of linguistic vocabulary is encoded. We can distinguish between closed data categories for which all possible values can be enumerated (such as `part of speech`) and open data categories which can take arbitrary values (possibly with certain constraints). The `writtenForm` of a `LexicalEntry` is an example for this case.

As the main development goal for our lexicon model is to standardize divergent and multilingual resources, representing a wide range of information types, the selection of classes and attributes has to ensure it is *comprehensive* (that is, the model covers all the information present in the resource) and *extensible*. As a result of these considerations, 39 LMF classes are chosen, along with 116 attributes for representing lexicographic information. Each attribute is registered in ISOcat. UBY-LMF also extends LMF in several ways, as it employs two new classes and several new relationships between classes for different information which is not covered in the standard – for instance, a `SemanticLabel` class, which is an optional subclass of `Sense`, `SemanticPredicate`, and `SemanticArgument`. This selection of a set of LMF classes and the relationships between them allows for structural interoperability, while the selection of data categories ensures the semantic interoperability of the lexicon model with respect to linguistic terminology and hence of our standardized representation of the LSRs. The final lexicon model can be expressed in the form of a UML class diagram, a Document Type Definition (DTD), or a similar form of schema description.

One major asset of UBY is that the semantic interoperability of resources is ensured not only at the terminology level, but also at the sense level, which is en-

---

<sup>9</sup><http://www.isocat.org/>

abled by pairwise alignments between senses in different LSRs; this is also explicitly modeled in UBY-LMF via the `SenseAxis` class. We will discuss the integration of alignments into UBY in Section 7.4, while the actual algorithmic process of automatically aligning senses and thus creating new alignments (which is the main part of our work) was exhaustively described in Chapters 4, 5 and 6.

More information and a complete discussion of the UBY-LMF lexicon model, including the mechanism for modeling alignments, can be found in the reference paper we published (Eckle-Kohler et al., 2012). Nevertheless, to illustrate how the creation of UBY and UBY-LMF was operationalized, we use the example of OmegaWiki to explain how concepts and terms in a resource are mapped to such a unified format and what obstacles have to be overcome. This mapping process was preceded by a thorough analysis of all resources involved, and OmegaWiki was the main focus of our own work (see also Section 3.3).

The final step of the LMF process is the population of the lexicon model – i.e. the transformation of the original resource into the classes and data categories of the defined lexicon model. The populated, standardized resource can be made available as, for example, a database dump. We discuss the population of our lexicon model in Section 7.4.

### A Subset of UBY-LMF for OmegaWiki

In this section, we describe the subset of the UBY-LMF model which is used to represent OmegaWiki, including an extension we deemed necessary for properly representing translation information. Figure 7.2 shows an overview of all classes and data categories we used for this subset.

#### Lexicon

In the UBY-LMF model, one unique `LexicalResource` instance which represents the complete resource consists of one or more `Lexicon` instances, i.e. each integrated resource is modeled as a separate `Lexicon`. Note further that LMF requires each `Lexicon` instance to belong to exactly one language (that is, having exactly one language identifier) – a requirement that reflects the diversity of different languages at the morphosyntactic and lexical-syntactic level. However, as mentioned before, OmegaWiki does not have separate editions for each language. Instead, OmegaWiki is based on the notion of multilingual synsets – that is, language-independent concepts to which lexicalizations of the concepts are attached. Thus, we have to split OmegaWiki’s defined meanings to create artificial language editions. For example, when populating our LMF model with a `Lexicon` for the German OmegaWiki, we iterate over all defined meanings and only create those `LexicalEntry`, `Sense`, etc. instances for which German lexicalizations are present. In turn, this means that concepts which are not lexicalized in German are simply left out of this `Lexicon`. The lexicalizations in the other languages are, however, not lost, but stored as translations using the `Equivalent` class (see below). For other multilingual LSRs such as Wiktionary and Wikipedia, each language edition constitutes one `Lexicon`, as does every monolingual LSR.





### Lexical Entry, Synset and Sense

The lexical information is modeled using the `LexicalEntry` class, which is characterized by a `Lemma` (that is, a written form) and a part of speech. For OmegaWiki, the `LexicalEntry` corresponds to each lexicalization of a particular defined meaning. Each `LexicalEntry` may also be connected to multiple instances of the `Sense` class modeling a certain meaning of the lexical entry. In our model, these senses are subsequently grouped into `Synsets`. This reflects the fact that the different lexicalizations of the same defined meaning describe the same concept and are thus synonyms.

### Lexicographic Information

An integral part of our LMF model is the representation of the variety of lexicographic information found in the various resources, which is represented by different classes attached to `Sense`: While `Definition` and `SenseExample` are self-explanatory, the `Statement` class contains further knowledge about a `Sense`, such as etymological information. The `SemanticLabel` class contains labels for many different dimensions of semantic classification (for example, domain, register, style or sentiment) for word senses. Such labels are very useful, as they contain valuable characteristics of the situations or contexts in which a word sense is usually used.

Relationships between word senses can be represented by means of paradigmatic relations, such as synonymy, antonymy or hyponymy that are modeled in the `SenseRelation` class. As all relations in OmegaWiki are encoded between defined meanings, the paradigmatic relations expressed by `SenseRelation` instances can be trivially transferred to `SynsetRelation` instances. That is to say, the structure of OmegaWiki stipulates that paradigmatic relations between synsets also hold for the contained senses and vice versa. Another fact worth mentioning is that, in contrast to most other resources, OmegaWiki also contains ontological (as opposed to linguistically motivated) relations – for instance, the *borders on* relation is used to represent neighboring countries. This is very much in the spirit of OmegaWiki, being a collection of lexicalized concepts rather than a classic dictionary. Those relations are also modeled using the `SenseRelation` and `SynsetRelation` classes.

### Translations

In addition to the elements contained in the first version of UBY-LMF described above, we also introduced a new `Equivalent` class which, for instance, is useful for translation applications. In this class, we store translation equivalents of a `Sense`, for example, the German translation *Barsch* of *bass*. Using the `Equivalent` class for this has been suggested before by Serasset (2012), but – as opposed to our model – they represent translations at the word level rather than at the level of word senses. For OmegaWiki, these translation equivalents are directly available via the lexicalizations in different languages attached to the same defined meaning. In other resources, translation equivalents are usually encoded as links to other language editions or plain text. Besides the written form of the translations and the target language, the `Equivalent` class includes the following additional attributes: `transliteration` to encode different scripts (such as Cyrillic), `geographicalVariant` for representing

a certain region in which the translated word is predominantly used (for example, Moscow), and `orthographyName` for storing a certain orthographic variant, such as the German orthography reform of 1996.

### SenseAxis

The sense alignments in UBY are represented by the `SenseAxis` class. Its role is twofold: first, it represents monolingual sense alignments, i.e. sense alignments between different lexicons in the same language, and second, it links the corresponding word senses from different languages, e.g. English and German. The latter is a novel interpretation of `SenseAxis` introduced by UBY-LMF. An alignment via a `SenseAxis` can thus be interpreted as an equivalent relation between senses in different resources.

In OmegaWiki, some of the defined meanings provide links to the corresponding Wikipedia page; these can be directly used to infer alignments between the two resources. More importantly, though, if more than one artificial language edition is created (see above), there naturally exists a considerable overlap of concepts which are lexicalized in different languages. To express that the corresponding word senses in these languages refer to the same meaning, we utilize the `SenseAxis` class to link them. In other words, the information originally contained in OmegaWiki's defined meanings is preserved by modeling it as a cross-lingual sense alignment between the artificial language editions.

### Syntactic Properties

To a small extent, OmegaWiki allows encoding syntactic properties such as verb valency. While this affects only a small fraction of the entries for now, we assume that the importance of this will increase as the resource is edited and extended by the crowd. Thus, we integrated this information to make the transformation as complete as possible or even lossless, and also to prepare the ground for integrating OmegaWiki with resources which specifically focus on syntactic properties. To cater for this, we are utilizing the classes `SubcatFrame`, `LexemeProperty` and `SyntacticBehavior` which enable us to model all of the syntactic information available in OmegaWiki.

## 7.4 UBY – The Final Resource

After mapping a resource to the UBY-LMF model as described above, this mapping is operationalized by a converter which takes as an input the original resource (in our case, the OmegaWiki database dump) and outputs the same data in the unified format, which is XML specified by a DTD. The same is true for existing alignments from third parties (see below), for which converters had to be created as well.

XML, however is only an intermediate format to ensure coherence with the underlying model, i.e. to enable the validation of the transformation process. For productive usage, the data is stored in an SQL database backend, using the Hibernate<sup>10</sup>

---

<sup>10</sup><http://www.hibernate.org/>

framework. Hibernate has the attractive property that we can seamlessly match instances encoded in XML documents to SQL tables and, even more importantly, Java classes. This lets us exploit the advantageous properties of all representation formats at the same time: i) the formal and sound definition of the model (ensuring interoperability) in XML, ii) the efficient storage and access to the knowledge in a database, with the added value of consistency enforced by database constraints and iii) easy programmatic usability via direct access to instances and their attributes in Java. The Java-based API we created based on this will be discussed in Appendix A.2.1.

The final resource UBY (i.e. the combined output of all converters) holds standardized and hence interoperable versions of the LSRs previously listed in Chapter 3, namely English WordNet 3.0, Wiktionary, Wikipedia, FrameNet 1.5, VerbNet 3.0, German Wikipedia, Wiktionary, GermaNet,<sup>11</sup> the syntactically rich IMSLex-Subcat,<sup>12</sup> as well as the multilingual OmegaWiki.

In its first version, UBY featured both monolingual and cross-lingual pairwise sense alignments between a subset of LSRs:

- WordNet–Wikipedia (Niemann and Gurevych, 2011) and WordNet–Wiktionary (Meyer and Gurevych, 2011) were created at the UKP lab. The datasets and algorithmic approach for creating the alignments are explained in Sections 3.5.1 and 4.2, respectively.
- WordNet–OmegaWiki (Gurevych et al., 2012) was created as part of this thesis. Our contribution was described and analyzed in Section 4.4.
- VerbNet–WordNet (Kipper et al., 2006) and VerbNet–FrameNet (Palmer, 2009) were created in previous work by other groups. We discuss them briefly in Section 3.5.1. Integrating them involved mapping the sense IDs from the proprietary alignment files to the corresponding sense IDs in UBY.
- OmegaWiki–Wikipedia alignments, as well as the alignments between the English and German parts of OmegaWiki could be straightforwardly derived as described above. Following this idea, interlanguage links between the English and German Wikipedia articles were integrated as well. We thereby gathered a considerable number of alignments which we assume to be correct as they were manually entered by the community. This idea was later extended to deriving a large WSA dataset for an alignment between Wiktionary and Wikipedia (see Section 3.5.2).

On top of these initial alignments, we added many more in the course of our research, as discussed in the previous chapters. For each resource pair we considered, we calculated a full alignment (i.e. covering the entire resources and not only the evaluation datasets) using the best algorithm and configuration in terms of F-measure. The sizes for these alignments as well as for the originally contained

---

<sup>11</sup>Note that in the latest version of UBY (0.6.0) GermaNet 8.0 is used, while GermaNet 7.0 was contained in the first version (0.1.0).

<sup>12</sup>This resource was not present in the first release.

	UBY 0.1.0	UBY 0.6.0	Increase
Lexical entries	4,259,894	4,290,660	0.7%
Senses	4,691,313	4,739,141	1.0%
Semantic relations	5,570,100	5,627,300	1.0%
Sense alignments	758,435	1,002,022	32.1%

Table 7.1: The content of UBY, for the original release (April 2012) and the current version (summer 2014)

alignments are listed in Table 7.2, while Table 7.3 presents a matrix overview of the alignments present and missing in UBY.<sup>13</sup>

In its first release version, UBY contained more than 4.2 million lexical entries, 4.6 million senses, 5.5 million semantic relations between senses and more than 700,000 alignments between senses. There were 890,000 unique German and 3.1 million unique English lemma-POS combinations.<sup>14</sup>

In Table 7.1, we see that until now there has been only a modest increase in lexical and sense coverage (which can mostly be attributed to the addition of IMSLex and a newer version of GermaNet), but a substantial improvement of the alignment density, which can for the large part be attributed to this thesis.

## 7.5 Community Issues

One of the most important reasons for creating UBY was the need for an easy-to-use powerful LSR to advance NLP research and development. Therefore, building an active community which uses the resource is one of the major concerns of the project. To this end, our group offers free downloads of the lexical data presented in this chapter as well as the accompanying software under open licenses from our website,<sup>15</sup> namely: The UBY-LMF DTD, mappings and conversion tools for existing resources and sense alignments, the Java-based API (see Appendix A.2.1), and, as far as licensing allows,<sup>16</sup> already-converted resources which are ready to use within UBY. If resources cannot be made available for download, the conversion tools will still allow users with access to these resources to import them into UBY easily. In this way, it will be possible for users to build their own “custom UBY” containing selected resources. As the underlying resources are subject to continuous change, updates of the corresponding components will be made available on a regular basis. The web interface to UBY<sup>17</sup> (see Appendix A.2.2) is soon to be replaced by an updated and extended version which will offer a more convenient access to the resource for the community.

<sup>13</sup>Note that these tables were already presented in Chapter 3 to help the reader understand the scope of our work. They are repeated here for convenience.

<sup>14</sup>Note that for homonyms, there may be more than one `LexicalEntry` for a lemma-POS combination.

<sup>15</sup><http://www.ukp.tu-darmstadt.de/data/uby>

<sup>16</sup>GermaNet and IMSLex are subject to a proprietary license agreement; hence, the converted UBY versions of them are not freely available, but the converters are.

<sup>17</sup><https://uby.ukp.informatik.tu-darmstadt.de/webui/>

Resource Pair	Source	Abbrev.	Approach	P	R	$F_1$	Size	UBY 0.1.0	UBY 0.6.0
WN-WKT	(Meyer and Gurevych, 2011)	MG11	Gloss similarity	0.67	0.65	0.66	99,662	✓	✗
	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.70	0.84	0.77	138,282	✗	✓
WN-WP	(Niemann and Gurevych, 2011)	NG11	Gloss similarity	0.78	0.78	0.78	50,351	✓	✗
	(Matuschek and Gurevych, 2013)	<b>MG13</b>	Shortest paths	0.75	0.87	0.81	83,192	✗	✓
	(Navigli and Ponzetto, 2012a)	NP12	Gloss/Structure	0.81	0.75	0.78	47,956	✗	✗
WN-OW	(Gurevych et al., 2012)	<b>Gu12</b>	Gloss similarity	0.55	0.53	0.54	23,024	✓	✗
	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.75	0.72	0.74	27,529	✗	✓
FN-WKT	(Hartmann and Gurevych, 2013)	HG13	Gloss similarity	0.73	0.75	0.74	12,326	✗	✗
	(Matuschek and Gurevych, 2013)	<b>MG13</b>	Shortest paths	0.77	0.79	0.78	12,340	✗	✓
GN-WKT	(Henrich et al., 2011)	He11	Gloss overlap	0.84	0.84	0.84	27,127	✗	✗
	(Matuschek and Gurevych, 2013)	<b>MG13</b>	Shortest paths	0.83	0.86	0.85	32,850	✗	✓
WKT-OW	(Matuschek et al., 2013)	<b>Ma13</b>	Gloss similarity	0.67	0.65	0.66	25,742	✗	✗
	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.79	0.64	0.71	25,727	✗	✓
WKT-WP en	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.82	0.70	0.76	66,050	✗	✓
WKT-WP de	(Matuschek and Gurevych, 2014)	<b>MG14</b>	Machine learning	0.89	0.81	0.85	21,872	✗	✓
VN-WN	(Kipper et al., 2006)	Ki06	Manual	N/A	N/A	N/A	40,716	✓	✓
VN-FN	(Palmer, 2009)	Pa09	Manual	N/A	N/A	N/A	17,529	✓	✓
(WN-GN)	(Vossen, 1998)	Vo98	Manual	N/A	N/A	N/A	12,079	✗	✗
(WN-FN)	(Ferrandez et al., 2010)	Fe10	Machine Learning	N/A	N/A	0.77	30,744	✗	✗
(WN-FN)	(Laparra et al., 2010)	La10	Structure	0.76	0.74	0.75	28,352	✗	✗
(OW-WP)	OmegaWiki crowd	OW	Manual	N/A	N/A	N/A	5,057	✓	✓
(OW en-OW de)	OmegaWiki crowd	OW	Manual	N/A	N/A	N/A	58,785	✓	✓
(WP en-WP de)	Wikipedia crowd	WP	Manual	N/A	N/A	N/A	463,311	✓	✓

Table 7.2: List of all available full alignments between resources in UBY. We report papers where these alignments were covered, best results achieved, the sizes of the full alignments as well as availability in the first (0.1.0) and current (0.6.0) version of UBY. For works to which we contributed, the abbreviation is given in bold face. Pairs we did not consider for our experiments are marked with parentheses.

	WN	VN	FN	WP en	WKT en	OW en	GN	WP de	WKT de	OW de
WN	manual Ki06	automatic (Fe10) (La10)	automatic NG11 MG13 NP12	automatic MG11 MG14	automatic Gu12 MG14	manual (Vo98)	✗	✗	✗	automatic Gu12
VN		manual Pa09	✗	✗	✗	✗	✗	✗	✗	✗
FN			✗	automatic HG13 MG13	✗	✗	✗	✗	✗	✗
WP en				automatic MG14	manual (OW)	✗	manual (WP)	✗	✗	✗
WKT en					automatic Ma13 MG14	✗	✗	✗	✗	✗
OW en						✗	✗	✗	✗	manual (OW)
GN								✗	automatic He11 MG13	✗
WP de									automatic MG14	manual (OW)
WKT de										✗
OW de										

Table 7.3: This matrix visualizes what alignments for resources in UBY are available and supplements Table 7.2.

## 7.6 Chapter Summary and Contributions

In this chapter, we present UBY, a unified lexical-semantic resource which holds standardized and thus interoperable versions of ten different, heterogeneous expert-built and collaboratively constructed resources in two languages. We discuss UBY-LMF, the representation format we have developed for this purpose which encompasses all knowledge from the integrated resources, and show how the standardization is operationalized by mapping OmegaWiki to this format, in turn presenting its most important features. This also includes the representation of sense alignments, which are the other key feature of UBY and enable to effectively leverage the combined knowledge from different LSRs.

In summary, our contributions in this chapter are:

**Contribution 1** We present the properties of UBY-LMF, our standardized representation format for LSRs which aims at modeling all knowledge contained in the resources included in UBY.

**Contribution 2** We show how OmegaWiki was modeled in and mapped to this format, exemplifying the mapping effort in the process.

**Contribution 3** We present the content of the final resource UBY, and also briefly discuss its development over time as well as the efforts to make it accessible to the research community.





# Chapter 8

## Applications of Sense Alignments

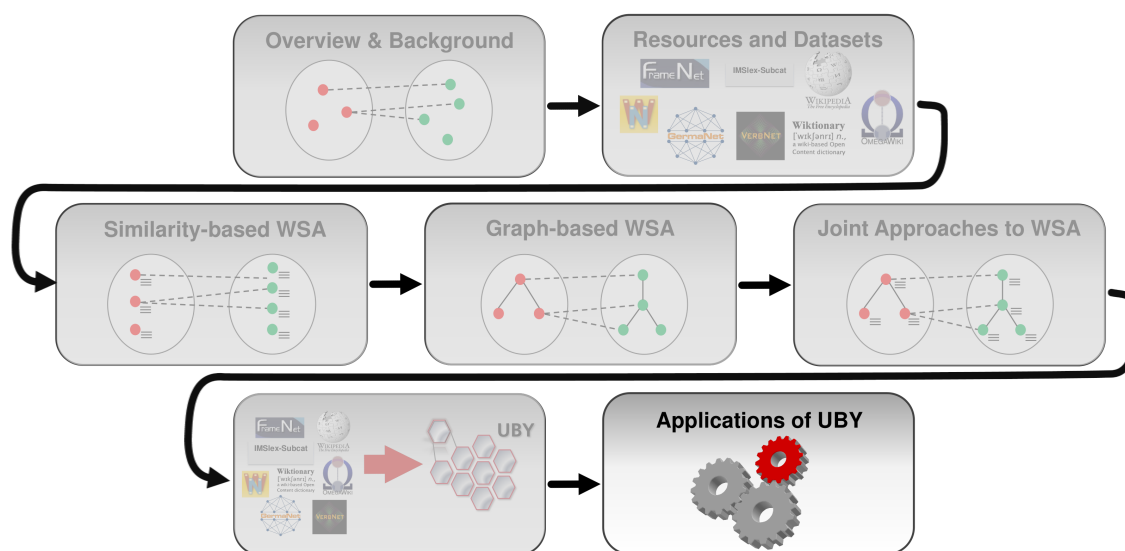


Figure 8.1: Visual outline of the thesis.

In this chapter, we demonstrate how the standardization and unification of LSRs, and especially the creation of sense alignments between them, can be beneficial for natural language processing applications. This is the ultimate goal of our alignment work and the creation of the unified resource UBY. We discuss two different scenarios: the clustering of fine-grained word senses in WordNet and GermaNet to improve word sense disambiguation (WSD) performance (Section 8.1), and the integration of the aligned collaboratively constructed LSRs Wiktionary and OmegaWiki into a computer-aided translation (CAT) environment (Section 8.2). In both sections, we give an overview of related work as well as a detailed description and evaluation of the respective approaches. Section 8.3 summarizes the chapter.

## 8.1 Using Alignments for Sense Clustering

### 8.1.1 Introduction

The driving motivation behind our alignment efforts is the fact that not every resource is equally well suited for each task, e.g. because of different lexical and sense coverage or different information types. For tasks such as word sense disambiguation (WSD), on which we focus in this section, using a combination of resources instead of a single one has already proven beneficial to the performance (Navigli and Ponzetto, 2012a; Cholakov et al., 2014b). On the other hand, the actual tagging of word senses usually happens referencing only one LSR (or sense inventory), and in the majority of cases this is an expert-built one. The Princeton WordNet (Fellbaum, 1998) is the predominant sense inventory for English because of its free availability, its comprehensiveness, and its use in dozens of previous studies and datasets. For German, GermaNet (Hamp and Feldweg, 1997) has positioned itself as the reference resource for WSD, although systematic investigation of German WSD has only recently begun (Henrich and Hinrichs, 2012).

However, there is much evidence to suggest that the sense distinctions of such expert-built wordnets are too fine-grained – i.e. they are more subtle than what is typically necessary for real-world NLP applications, and sometimes even too subtle for human annotators to consistently recognize. This point has been made specifically for WordNet (Ide and Wilks, 2006), but applies to other similar resources as well. This makes improving upon experimental results difficult, while at the same time the downstream benefits of improved WSD based on these LSRs are often not clearly visible.

Using a different sense inventory could solve the problems inherent to expert-built LSRs, and recently collaboratively constructed resources, such as Wiktionary and Wikipedia, have been suggested (Mihalcea, 2007). As pointed out in Chapter 3, these resources are attractive because they are large, freely available in many languages, and continuously improved. However, they also still contain considerable gaps in coverage, there are only few large-scale sense-annotated corpora using them (like the *Wikilinks* corpus (Singh et al., 2012)), and for some parts of speech their senses are also rather fine-grained (cf. Section 3.3). Much prior work has therefore focused instead on enhancing wordnets by decreasing their granularity through (semi-)automatic sense clustering. However, until now, the focus of attention has almost exclusively been the English WordNet. While it has been shown that such clustering significantly enhances both human inter-annotator agreement (Palmer et al., 2007) and automatic WSD performance (Snow et al., 2007), the previous approaches had been specifically tailored towards this resource, for instance making use of WordNet’s particular semantic relations, which makes it difficult to apply them to other LSRs and languages.

In this section, we describe a solution to the granularity problem which taps the benefits of collaboratively constructed LSRs (i.e. the more coarse-grained senses) without the disadvantages of using them as wholesale replacements for other LSRs in WSD (i.e. the lower coverage). We induce a clustering of an expert-built resource’s senses by using the alignments to other resources, more precisely, by grouping source senses which map to the same target sense. This results in a coarse-grained sense

inventory with good coverage.

In contrast to previously used alignment-based clustering techniques (see Section 8.1.2), we create the alignment using Dijkstra-WSA, the algorithm we presented in Chapter 5. This allows us to produce clusterings based on several different resource alignments, for which we conduct in-depth analyses and evaluations. To demonstrate the language-independence of our approach, we produce clusters for GermaNet as well as WordNet, while our algorithm is easily applicable to multiple resource pairs at a time. This again reflects our overarching goal to find solutions which are broadly applicable and do not require language- or resource-specific engineering efforts.

### 8.1.2 Related Work

The clustering of fine-grained senses has been widely researched in the past. Peters et al. (1998), Buitelaar (2000), Mihalcea and Moldovan (2001a), Tomuro (2001) and Ide (2006) use the hierarchical structure of wordnets to group related senses in different ways, Dolan (1994) and Chen and Chang (1998) use heuristics based on text- and metadata to calculate similarity scores for senses, Resnik and Yarowsky (2000) group senses which share translations into another language, and Agirre and Lopez de Lacalle (2003) and McCarthy (2006) exploit the distributional similarity of senses across different contexts. Other methods derive clusters from disagreements between human annotators (Chklovski and Mihalcea, 2003), map senses to learned semantic classes (Kohomban and Lee, 2005) and analyze syntactic patterns and predicate-argument structures (Palmer et al., 2004, 2007).

A prevalent issue with these approaches is that an evaluation of the clusterings on state-of-the-art WSD systems is usually not provided, i.e. it is not possible to determine their usefulness in a practical application. A notable exception is Palmer et al. (2007), who report human inter-annotator agreement for fine-grained and clustered senses and also relate their results to random clusterings of the same granularity. McCarthy (2006) also uses a WSD-based evaluation, but only naive first-sense baselines are considered for comparison.

Recently, approaches which reduce WordNet’s sense granularity by aligning it to another, more coarse-grained resource have received more attention. Navigli (2006) produces an alignment between WordNet and the *Oxford Dictionary of English* (Soanes and Stevenson, 2003) using lexical overlaps and semantic relationships between pairs of sense glosses. WordNet senses which align to the same *Oxford* sense are clustered together, akin to the approach we present here. Snow et al. (2007) and Bhagwani et al. (2013) go beyond Navigli’s approach by training machine learning classifiers to decide whether two senses should be merged. They use a variety of features derived from WordNet as well as external sources, e.g. the already mentioned *Oxford*-WordNet mapping or the *OntoNotes* corpus (Pradhan et al., 2007). Their methods outperform the baseline, but they require a considerable amount of annotated training data. Moreover, the features are largely based on WordNet-specific information types (e.g., shared antonym relations for two senses). This might render the methods’ application to other resources challenging. All of these alignment-based methods are extrinsically evaluated using standard WSD datasets, but in each case the reported results are at least questionable – we elaborate on this in Section 8.1.4.

We go beyond the previous work by employing Dijkstra-WSA (see Chapter 5). As mentioned earlier, it makes the alignment (and hence, the clustering approach) easily applicable to multiple resource combinations, and the flexibility of Dijkstra-WSA allows a deeper comparative analysis of the alignment-based clusterings against not one but three different LSRs – namely, Wiktionary, Wikipedia, and OmegaWiki. We investigate how the different properties of these resources influence the alignments and clusterings, particularly with respect to the performance for different parts of speech. Such an in-depth analysis has not been performed in the previous work. We have chosen to focus on the collaboratively constructed LSRs, as their emergence has led to an ongoing discussion about their quality and usefulness (Zesch et al., 2007; Meyer and Gurevych, 2012b; Krizhanovsky, 2012; Gurevych and Kim, 2012; Hovy et al., 2013). This work aims to contribute to this discussion by further investigating the crucial aspects of granularity and coverage of collaboratively constructed resources, extending the discussions from Section 3.3.

### 8.1.3 Task Definition

*Word sense clustering* is the process, be it manual or automatic, of identifying senses in an LSR which are similar to the extent that they could be considered the same, slight variants of each other, or subsenses of the same broader sense. The purpose of this is to merge these senses (i.e. to consider the set of clustered senses as a single new sense) in order to facilitate the usage of the sense inventory in applications which benefit from a lower degree of polysemy. For example, the two WordNet senses of *ruin*—“destroy completely; damage irreparably” and “reduce to ruins”—are very closely related and could be used interchangeably in many contexts.

One way to produce such a clustering is WSA, which has been precisely defined in Section 2.2. If it is not restricted to 1:1 alignments (i.e., a sense may participate in more than one pair), it is possible that a sense  $s_1$  in one LSR  $A$  is assigned to several senses  $t_1, \dots, t_n$  in another LSR  $B$ . Assuming that all alignments are correct, this implies that  $s \in A$  is more coarse-grained and subsumes the other senses, which in turn can be considered as a sense cluster within  $B$ . For example, the aforementioned senses of *ruin* could both be aligned to the Wiktionary sense “to destroy or make something no longer usable”, which would result in their merging.

### 8.1.4 Evaluation

#### Methodology

A common approach for the evaluation of sense clusterings is to use the output from an existing WSD system and recalculate the scores on a standard evaluation dataset in accordance with the clustering. This means that an answer is also considered correct in case it is just contained in the same cluster as the actual correct answer. As any clustering trivially increases accuracy, it is important to measure the accuracy of each clustering in relation to that of a random cluster with equivalent granularity. While Navigli (2006) disregards this issue, Snow et al. (2007) and Bhagwani et al. (2013) use a mathematical method for determining this random clustering score. However, they do not consider the possibility of multiple correct answers for a given

	Items	Monosemous Items	Deg. of Polysemy
Nouns	1,539	1	2.86
Verbs	962	3	3.75
Adjectives	218	1	2.48
All	2,719	5	3.14

Table 8.1: Statistics about WebCAGe.

instance, and thus underestimate the random baseline on the Senseval-3 dataset they use for their experiments. In (Matuschek et al., 2014), an improved version of this method is presented which is also applicable to the general case. As it is implemented in the freely available DKPro WSD framework (Miller et al., 2013), we use it in our own experiments to compute the WSD performance for the clustered senses.

### Experiments on GermaNet

For the first set of experiments, we align GermaNet to the German editions of the three different collaboratively constructed LSRs Wikipedia, Wiktionary, and OmegaWiki. Our goal in this setup is to demonstrate that effective word sense clustering is possible for resources in languages other than English using the generic and language-agnostic alignment approach we presented earlier. Moreover, we aim to cover two dictionary resources which are at different stages of development regarding their size and coverage (OmegaWiki and Wiktionary, cf. Section 3.3) as well as the most popular collaboratively constructed encyclopedia (Wikipedia) to investigate the influence of resource choice on the clustering results.

As a detailed discussion of the alignment approach is given in Chapter 5, we focus on the clusterings which are derived from the alignment and relate these results to the properties of the LSRs involved. We deem this especially interesting with regard to the collaborative construction process, which is inherently different from the traditional way of building resources and hence is bound to yield different ways of representing meaning.

For evaluation, we use WebCAGe (Henrich et al., 2012), which is currently (to the best of our knowledge) the only freely available, German-language sense-annotated corpus of considerable size. WebCAGe is a “lexical sample” corpus, with sense usage examples taken from Wiktionary example sentences, Wikipedia articles, books published in the German *Gutenberg-Projekt* and texts from German web sites. The latest version (2.0), which we use here, contains 10,429 instances of 2,719 lexical items annotated with GermaNet 8.0 senses. As with the Senseval-3 dataset, many WebCAGe instances specify multiple gold-standard senses. Statistics about the polysemy of the dataset by part of speech are given in Table 8.1.

German-language WSD is still in its infancy; the best results reported so far are for various weakly supervised, Lesk-like systems (Henrich and Hinrichs, 2012). For our extrinsic cluster evaluation, we therefore rescore the sense assignments made by their best-performing *lsk\_Ggw+Lgw* system (in terms of recall and F-measure) when run on the entire WebCAGe 2.0 corpus. These results are given in each result table (Tables 8.4, 8.5, 8.6 and 8.7) in the column *none*, which indicates that no

	aff.	imp.	%
OmegaWiki (DWSA)	438	130	29.7
OmegaWiki (Sim. only)	712	165	23.2
OmegaWiki (w/backoff)	872	205	23.5
Wiktionary (DWSA)	1355	311	23.0
Wiktionary (Sim. only)	1463	349	23.8
Wiktionary (w/backoff)	1797	349	19.4
Wikipedia (DWSA)	773	120	15.5
Wikipedia (Sim. only)	710	158	22.2
Wikipedia (w/backoff)	852	147	17.3

Table 8.2: Number and percentage of lexical items in the German evaluation dataset affected and improved by the clusterings. The slight proportional decrease in improved items in some configurations results from an improved alignment recall using the backoff, at the expense of precision.

	GN	OW	WKT	WP
Nouns cov. (%)	100.0	20.6	99.9	80.6
Verbs cov. (%)	100.0	20.7	99.9	—
Adjs. cov. (%)	100.0	29.8	98.6	—
Items cov. (%)	100.0	21.4	99.8	45.6
Senses / noun	2.86	1.18	3.84	2.25
Senses / verb	3.75	1.31	3.59	—
Senses / adj.	2.48	1.26	3.24	—
Senses / item	3.14	1.23	3.69	2.25

Table 8.3: Coverage of lexical items in the German test set per resource, and the degree of polysemy (i.e., the average number of senses per item in the resource). More information about the dataset is given in Table 8.1.

clustering was applied.

**GermaNet-OmegaWiki** When Dijkstra-WSA without the similarity-based back-off is used for clustering, the clusters are small and few in number. As a consequence, few lexical items in the dataset are affected by the clustering. This is very much in line with the observation made earlier that graph-based alignments usually yield good precision at the expense of recall if one of the graphs (in this case, OmegaWiki) is particularly sparse. So although relatively few senses are aligned and subsequently clustered, the clusters seem mostly correct, which is indicated by the significant overall improvement. The first line of Table 8.2 shows how many of the 10,429 lexical items of the evaluation dataset were actually affected by this clustering configuration, and of these how many show an increase in accuracy over the random clusters of similar granularity (which indicates of the validity of the clusters).

For adjectives, which constitute the smallest part-of-speech group in the dataset, there is almost no clustering at all, as for most senses Dijkstra-WSA identified no

	none	rand.	WSA	$\pm$
noun	51.1	60.9	62.5	<b>1.6*</b>
verb	43.1	45.8	46.6	0.8*
adj.	43.3	45.0	45.0	0.0
all	48.1	55.3	56.5	1.2*
noun	51.1	61.6	62.7	1.1*
verb	43.1	55.5	56.3	0.8*
adj.	43.3	61.6	62.1	0.5
all	48.1	59.8	60.7	0.9*
noun	51.1	66.9	68.5	<b>1.6*</b>
verb	43.1	56.0	57.3	<b>1.3*</b>
adj.	43.3	61.1	62.0	<b>0.9</b>
all	48.1	63.3	64.7	<b>1.4*</b>

Table 8.4: WSD accuracy, by part of speech, using clusterings derived from alignments of GermaNet to OmegaWiki with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results per POS, asterisks denote statistically significant differences from the random baseline.

targets, or only one target. The situation is better for nouns and verbs; while the clusters are not large (usually 2–3 senses), the high-precision clustering did improve the results. Nouns especially saw a statistically significant improvement over the random clustering (1.6 percentage points). The upper third of Table 8.4 shows the full results for this setup. The table shows the original accuracy score without clustering (*none*), the accuracy when our clustering is used (*WSA*), the accuracy when random clustering is used (*rand.*), and the difference between the latter two ( $\pm$ ).

When only gloss similarity is used in isolation, we achieve a higher recall for the alignment and thus larger clusters. This way, we are able to also cluster a substantial number of adjectives, leading to an increase in WSD performance. However, the overall results are worse due to the lower precision for nouns.

When we employ the backoff to improve the recall of the graph-based alignment (i.e. a combination of both approaches), we get more and larger clusters (see third line of Table 8.2) which results in a significant improvement in WSD accuracy for nouns and verbs (lower third of Table 8.4). Although we observed that alignment precision for this setup is generally worse than for Dijkstra-WSA alone, the alignments are seemingly still precise enough to form meaningful clusters which contain only a few errors.

A good example is the verb *markieren* (*to mark*), whose only sense in OmegaWiki (“Auf irgendeine Art kennzeichnen für spätere Bezugnahme”, English: “somehow denote for later reference”) is aligned to two GermaNet senses, one covering the marking in text and the other covering territorial marking. The difference in polysemy between GermaNet and OmegaWiki (see Table 8.3) pays off in this case, as the coarse OmegaWiki sense subsumes the GermaNet senses. This is exactly the

	none	rand.	WSA	±
noun	51.1	75.1	77.2	<b>2.1*</b>
verb	43.1	60.1	61.8	<b>1.7*</b>
adj.	43.3	82.5	83.0	0.5
all	48.1	71.2	73.0	<b>1.8*</b>
noun	51.1	72.3	73.8	1.4*
verb	43.1	58.7	58.7	0.0
adj.	43.3	65.9	66.3	0.4
all	48.1	67.8	68.7	0.9*
noun	51.1	83.2	85.3	<b>2.1*</b>
verb	43.1	73.7	74.3	0.6
adj.	43.3	87.9	87.8	-0.1
all	48.1	80.7	82.2	1.5*

Table 8.5: WSD accuracy, by part of speech, using clusterings derived from alignments of GermaNet to Wiktionary with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results per POS, asterisks denote statistically significant differences from the random baseline.

intended effect when this kind of clustering is performed.

However, even in the setup providing the highest recall, the number of affected lexical items for GermaNet is still small compared to the other resources, i.e. there are many notable gaps in coverage (Tables 8.2 and 8.3). Even quite commonly used terms like *öffentlich* (*public*) are missing from OmegaWiki altogether, which prevents even better clustering results in lieu of appropriate alignment candidates for many GermaNet senses. This underrepresentation of lemmas and senses can be attributed to the fact that OmegaWiki, in comparison to Wiktionary and Wikipedia, is in an earlier stage of development. This is especially true for the German edition, which is substantially smaller than the English one (i.e. there exist fewer lexicalizations of senses). This difference as well as the generally low lexical coverage of OmegaWiki as compared to other resources have already been discussed in Section 3.3.

**GermaNet-Wiktionary** Wiktionary’s coverage of lexical items is almost the same as GermaNet’s (> 99%; see Table 8.3). In conjunction with the much denser Wiktionary graph, this leads to a higher number of affected items in the test dataset and, consequently, significantly better overall results in comparison to OmegaWiki in the Dijkstra-WSA only setup. For nouns and verbs, the clustering yields major improvements (see Table 8.5), while the benefit for adjectives is modest. However, it comes as a surprise that the results are not even better than they already are – if for almost every lexeme alignment targets can be found, the assumption would be that many clusters could be formed. The reason for this being not the case is that on the test dataset, the degree of polysemy is almost the same in both resources, and GermaNet is even substantially less polysemous for verbs. Hence, for many senses in GermaNet there exists an equivalent sense with comparable granularity in



Wiktionary, and no 1:n mapping can be found which would imply a clustering. For instance, GermaNet contains two senses for the aforementioned adjective *öffentlich*, both of which are correctly aligned to exactly one of the two Wiktionary senses.

In other words, GermaNet senses are for most part represented in Wiktionary, and with equally (or even more) subtle sense distinctions, which impairs even better results for our clustering approach. On the other hand, this is a strong indicator of the quality of the German Wiktionary; this superiority as compared to the English version regarding different aspects has, for instance, also been discussed by Meyer (2013). The much larger density of semantic relations as compared to the English version made the German Wiktionary also very suitable for Dijkstra-WSA (see Section 5.3.3).

In our WSA experiments (see Section 5.3.3), we also observed that for the GermaNet-Wiktionary alignment, the results are almost identical with gloss overlap and Dijkstra-WSA. While this is also true for the clustering task regarding the recall as seen by the almost identical number of affected lexical items (Table 8.2), the precision is considerably lower, indicated by the lower increase in WSD performance, especially for verbs. This can be attributed to the much higher ambiguity in the WSD dataset (see Table 8.1) than in the WSA dataset (Table 3.17).

When both approaches are combined, recall is again considerably higher, but the overall results are not – more items are affected, but no more can be improved (see Table 8.2). This is especially true for adjectives: too many senses are grouped together with too low precision, leading to a WSD performance improvement which is indistinguishable from random clustering. Here, we apparently hit the limits of the clustering approach. While large clusters (and many affected items) are generally desirable, a certain level of precision has to be maintained for this approach to be effective.

**GermaNet-Wikipedia** As Wikipedia contains almost exclusively noun concepts, our evaluation for this clustering was restricted to this part of speech (see Table 8.6). We observe that the results for Dijkstra-WSA alone as well as for the similarity-based approach are significantly better than random, but worse than for the Wiktionary and OmegaWiki clusterings. This is explicable by the fact that the polysemy for nouns is comparable for GermaNet and Wikipedia (see Table 8.3). The observation made for Wiktionary that similar granularity implies many 1:1 alignments and thus few and small clusters holds here as well, as many GermaNet noun senses in the dataset have a corresponding entry in Wikipedia. An example is the noun *Filter*, where GermaNet encodes three senses (filter for liquids, air filter and polarization filter) which are all present in Wikipedia and correctly aligned. An interesting side note is that, due to its encyclopedic focus, Wikipedia contains quite a few senses which are rather obscure and unlikely to be found in a dictionary (e.g., *Filter* is also an American band). Our analysis shows, however, that the alignment algorithm reliably rules them out as alignment targets so that they usually do not impair the clustering outcome.

When combining both approaches in the hybrid setup, we get the hypothesized boost in recall, whereas the significantly better WSD result (+2.0 accuracy as compared to the random setup) suggests that the precision is still acceptable. This is again in line with the results for WordNet-Wikipedia alignment, which is comparable

	none	rand.	WSA	$\pm$
noun	51.1	75.1	76.3	1.2*
noun	51.1	70.5	71.6	1.1*
noun	51.1	76.6	78.6	<b>2.0*</b>

Table 8.6: WSD accuracy, by part of speech, using clusterings derived from alignments of GermaNet to Wikipedia with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results per POS, asterisks denote statistically significant differences from the random baseline.

	none	rand.	WSA	$\pm$
noun	51.1	75.1	77.2	<b>2.1*</b>
verb	43.1	60.1	61.8	<b>1.7*</b>
adj.	43.3	61.1	62.0	<b>0.9</b>
all	48.1	69.7	71.6	<b>1.9*</b>

Table 8.7: WSD accuracy, by part of speech, using the best clusterings of GermaNet for each part of speech. Nouns and verbs use clusterings from an alignment with Wiktionary, and adjectives from an alignment with OmegaWiki. Asterisks denote statistically significant differences from the random baseline.

due to the similar structures of WordNet and GermaNet. In this setup, the hybrid approach yielded better recall with the same precision as the individual approaches.

**Combined approaches** Our experiments show that clustering GermaNet against different collaboratively constructed resources using a state-of-the-art WSA algorithm is indeed effective: the WSD results consistently beat comparable random clusterings with few exceptions, and in many cases the improvement is significant.

One of the main observations in this context was that different clusterings do not work equally well on each part of speech. While OmegaWiki works best for adjectives, Wiktionary shows the best results for nouns and verbs. Thus, we performed an additional experiment where optimal clusterings were chosen for each part of speech (see Table 8.7). In this case, we achieve a significant improvement for each part of speech except adjectives, and thus the strongest overall improvement (1.9 percentage points) over the random clustering. The reduction in average polysemy by part of speech for this clustering is shown in Table 8.8.

This demonstrates that our generic and language-independent approach is effective, although it consists solely of an alignment algorithm which does not rely on any resource-specific tuning or knowledge external to any of the resources involved. This is in strong contrast to previous work such as Snow et al. (2007), who employ further external resources, as well as features specifically tailored towards WordNet in a supervised machine learning setup.

	before	after	reduction (%)
noun	2.35	1.84	21.60
verb	2.81	2.37	15.49
adj.	2.19	1.84	9.46
all	2.48	2.02	18.56

Table 8.8: Reduction in average polysemy for all polysemous words in GermaNet with our optimal clusterings.

	GAMBL			SenseLearner			Koç University		
	aff.	imp.	%	aff.	imp.	%	aff.	imp.	%
OmegaWiki	106	29	27.4	102	32	31.4	103	31	30.0
OmegaWiki (Sim. only)	134	29	21.6	150	31	20.7	163	31	19.0
OmegaWiki (w/backoff)	302	115	38.1	295	111	37.6	300	116	38.7
Wiktionary	163	43	26.4	163	48	29.4	167	51	30.5
Wiktionary (Sim. only)	147	23	15.6	148	22	14.9	144	23	16.0
Wiktionary (w/backoff)	224	60	26.8	222	58	26.1	219	62	28.3
Wikipedia	69	12	17.4	72	12	16.7	70	10	14.1
Wikipedia (Sim. only)	54	10	18.5	61	13	21.3	53	9	17.0
Wikipedia (w/backoff)	107	17	15.9	106	18	17.0	103	19	18.4

Table 8.9: Number and percentage of lexical items in the English dataset affected and improved by the clusterings. The slight proportional decrease in improved items in some configurations results from an improved alignment recall using the backoff, at the expense of precision.

### Experiments on WordNet

To demonstrate the validity of our approach for English, we clustered WordNet by aligning it to the English editions of the same three collaboratively constructed LSRs and used the coarse-grained WordNet for WSD.

We use the raw sense assignments of the three top-performing systems in the

	WN	OW	WKT	WP
Nouns covered (%)	100.0	54.5	92.6	84.6
Verbs covered (%)	100.0	49.7	88.0	—
Adjs. covered (%)	100.0	44.6	94.2	—
Lexemes covered (%)	100.0	50.0	90.9	36.3
Senses / noun	4.48	1.51	4.93	5.29
Senses / verb	6.69	1.57	4.88	—
Senses / adj.	3.60	1.34	3.09	—
Senses / lexeme	4.81	1.48	4.36	5.29

Table 8.10: Coverage of lexical items in the English test set per resource, and the degree of polysemy (i.e., the average number of senses per item in the resource).

	GAMBL				SenseLearner				Koç University			
	none	rand.	WSA	±	none	rand.	WSA	±	none	rand.	WSA	±
noun	69.0	70.3	70.9	0.6	68.7	69.9	70.6	0.7	69.3	70.4	71.4	1.0*
verb	59.0	61.7	63.0	1.3	56.1	59.4	61.4	<b>2.0*</b>	54.1	57.3	59.3	<b>2.0*</b>
adj.	67.0	67.6	67.9	0.3	70.4	70.8	70.7	-0.1	70.4	70.8	70.7	-0.1
all	65.2	66.9	67.7	0.8*	64.6	66.4	67.5	1.1*	64.1	65.9	67.0	1.1*
noun	69.0	71.2	72.3	1.1*	68.7	71.1	72.1	1.0*	69.3	71.5	72.7	1.2*
verb	59.0	60.7	59.8	-0.9	56.1	58.2	57.3	-0.9	54.1	56.4	54.9	-1.5*
adj.	67.0	67.3	67.4	0.1	70.4	70.5	70.5	0.0	70.4	70.5	70.5	0.0
all	65.2	66.8	70.0	0.2	64.6	66.4	66.6	0.2	64.1	66.0	66.0	0.0
noun	69.0	78.3	81.2	<b>2.9*</b>	68.7	78.4	79.9	1.5	69.3	78.4	80.5	2.1*
verb	59.0	70.9	69.4	-1.5*	56.1	69.5	65.5	-4.0*	54.1	68.1	65.9	-2.2*
adj.	67.0	78.3	83.4	<b>5.1*</b>	70.4	79.1	81.6	2.5*	70.4	79.5	82.1	2.6*
all	65.2	75.7	77.4	<b>1.7*</b>	64.6	75.4	75.1	-0.3	64.1	74.9	75.5	0.7

Table 8.11: WSD accuracy, by system and part of speech, using clusterings derived from alignments of WordNet to OmegaWiki with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results per POS. An asterisk marks statistically significant improvements.

Senseval-3 English all-words WSD task (Snyder and Palmer, 2004): GAMBL (Decadt et al., 2004), SenseLearner (Mihalcea and Faruque, 2004), and the Koç University system (Yuret, 2004); we provide some implementation details about these algorithms in the evaluation section in relation to our achieved results and also report the average performance for all systems. We have chosen this dataset in order to make our results comparable to the ones obtained by Snow et al. (2007) and Bhagwani et al. (2013). However, this proved to be impossible in light of the issues with their evaluation (see Section 8.1.4). Their full cluster datasets are also not available, which is why we were unable to re-evaluate them ourselves.

**WordNet-OmegaWiki** As for German, we also obtain only few clusters when only Dijkstra-WSA is used, which however seem correct for the most part. The first line of Table 8.9 shows how many of the 2041 lexical items of the evaluation dataset were actually affected by this clustering configuration, and of these how many saw an increase in accuracy.

Clustering adjectives also proved challenging for English, as only few possible alignment targets exist for most senses, while for nouns and verbs a substantial number of clusters could be generated. For verbs, we achieved a major improvement over the random clustering (up to 2.0 percentage points, depending on the WSD algorithm), while for nouns the increase was less remarkable and only statistically significant in case of the Koç University algorithm. The upper third of Table 8.11 shows the full results for this setup for each Senseval-3 system.

When using only gloss similarity, we observe stable results for adjectives and a slight improvement for nouns, but when handling verbs, which have the highest degree of polysemy in WordNet and are thus usually considered the most challenging part of speech, we observe the limits of the clustering against OmegaWiki senses:

the results are worse than with the random clustering (and for the Koç University algorithm significantly so). As our analysis shows, the difference in polysemy is too great in this case, and there are many cases where even quite common verb senses encoded in WordNet are entirely missing from OmegaWiki. In these cases, the system either finds no alignment target, or an incorrect one which leads to the erroneous clustering.

By combining Dijkstra-WSA with the backoff (and hence producing even larger clusters, see third line of Table 8.9 and lower third of Table 8.11), the error on verbs is further magnified, i.e. incorrect alignment targets are introduced into even more clusters, which leads to results significantly worse than random for each system. An example is the verb *pour*, which has only the rather specific sense “to cause a liquid to flow into a container” in OmegaWiki. This is aligned to the WordNet senses “cause to run” (by the similarity-based backoff) and “rain heavily” (by Dijkstra-WSA), which should not be in the same cluster.

Interestingly, this magnification phenomenon also works in the other direction, i.e. to improve the results for nouns and adjectives, as both components (Dijkstra-WSA and backoff) contribute correct, but largely disjoint set of alignments. As for verbs, we get more and larger clusters (reflected by the much larger number of affected lexical items), but in this case this leads to a significant improvement in WSD accuracy – more than 5 percentage points for the GAMBL algorithm on adjectives. Although we have seen earlier that alignment precision for this setup is usually slightly worse than for Dijkstra-WSA alone, the alignments for nouns and adjectives are seemingly still precise enough to form meaningful clusters which contain only a few errors. As an example, the adjective *national* has only one sense in OmegaWiki (“having to do with a nation”), which is aligned to three WordNet senses (“of or relating to or belonging to a nation or country”, “limited to or in the interests of a particular nation”, and “concerned with or applicable to or belonging to an entire nation or country”) which are closely related. As for German, the difference in polysemy between WordNet and OmegaWiki (see Table 8.10) is beneficial here, as the coarse OmegaWiki sense subsumes the WordNet senses with only a negligible loss of information. However, another observation we made for German also holds in this case: many meanings found in other resources are not encoded in OmegaWiki, such as the WordNet sense “feeling or showing no enthusiasm” for *cold*. OmegaWiki contains only the literal meaning “having a low temperature”; this is also reflected by the low lexical coverage of OmegaWiki on the evaluation dataset (see Table 8.10) and in general (see Section 3.3). While the coverage for English is better than for German, it is still substantially higher in Wiktionary and Wikipedia.

In summary, the low polysemy and “clean-cut” senses in OmegaWiki are helpful for adjectives and nouns, but not for verbs, as more (and also more fine-grained) senses are needed to find correct equivalents in this case. As noted before, this is partly due to OmegaWikis’ early stage of development, but it can also be reasoned that less frequent senses for English have not yet been addressed by the OmegaWiki community due to its inherent multilingual structure, so that the focus is rather put on covering more common meanings which are also lexicalized in other languages.

**WordNet-Wiktionary** As for German, Wiktionary’s coverage of lexical items is much better than OmegaWiki’s (> 90%; see Table 8.10). However, this does

	GAMBL				SenseLearner				Koç University			
	none	rand.	WSA	±	none	rand.	WSA	±	none	rand.	WSA	±
noun	69.0	70.6	70.7	0.1	68.7	70.5	70.9	0.4	69.3	71.1	72.4	1.3*
verb	59.0	65.7	66.4	0.7	56.1	64.0	65.2	1.2	54.1	61.7	63.2	1.5*
adj.	67.0	68.4	69.0	<b>0.6</b>	70.4	71.6	72.1	0.5	70.4	71.5	71.8	0.3
all	65.2	68.6	69.0	0.4	64.6	68.5	69.2	0.7	64.1	67.9	69.1	1.2*
noun	69.0	70.6	70.9	0.3	68.7	70.5	70.4	-0.1	69.3	70.6	71.4	0.8
verb	59.0	60.6	62.1	1.5*	56.1	57.6	58.8	1.2	54.1	55.8	57.4	1.5*
adj.	67.0	70.2	70.7	0.5	70.4	72.9	73.2	0.3	70.4	72.9	73.2	0.3
all	65.2	67.0	67.8	0.8*	64.6	66.4	66.8	0.4	64.1	65.8	66.7	0.9*
noun	69.0	72.5	73.3	0.8	68.7	72.7	72.9	0.2	69.3	72.7	74.8	<b>2.1*</b>
verb	59.0	67.2	67.7	0.5	56.1	65.8	66.8	1.0	54.1	63.3	65.0	<b>1.7*</b>
adj.	67.0	72.0	72.6	<b>0.6</b>	70.4	74.5	74.9	0.4	70.4	74.4	74.6	0.2
all	65.2	68.6	69.0	0.4	64.6	70.6	71.1	0.5	64.1	69.7	71.3	<b>1.6*</b>

Table 8.12: WSD accuracy, by system and part of speech, using clusterings derived from alignments of WordNet to Wiktionary with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results per POS. An asterisk marks statistically significant improvements.

not lead to better overall results for Dijkstra-WSA only. For nouns and adjectives, the clusterings yield only minor improvements (see Table 8.12). The reason is, as observed for German in several configurations, the similar polysemy in both resources for nouns as well as adjectives (see Table 8.10); thus, 1:n mappings can only be found for few senses. For instance, WordNet contains thirteen senses for the aforementioned adjective *cold*, most of which are aligned to exactly one of the fifteen Wiktionary senses.

For verbs, the polysemy is notably lower in Wiktionary (4.8 vs. 6.7). This implies more 1:n alignments and subsequently derived clusters, which boosts the WSD results in this case (up to 1.5 percentage points). According to our analysis of the dataset, many of the WordNet verb senses are indeed represented in Wiktionary, but they are subsumed (or condensed) into units which seem sensible from a lexicographic perspective and which are easier for systems to choose between. A typical example is *blur*: none of its meanings enumerated by WordNet are completely missing from Wiktionary, though two nearly indistinguishable ones (“to make less distinct or clear” and “to make unclear, indistinct, or blurred”) are aligned to the single Wiktionary sense “to make indistinct or hazy, to obscure or dim”. Nevertheless, the result for verbs is still worse than for OmegaWiki in the same configuration.

In Section 5.3.3, we have seen that for the WordNet-Wiktionary alignment the recall is considerably higher when gloss similarity is used, while precision is slightly worse than for Dijkstra-WSA. This ought to lead to larger clusters, but in our experiments the approaches are almost on par in this respect (as reflected by the number of affected items – see Table 8.9). Thus, WSD performance also remained mostly stable; a significant improvement was only observed on nouns for the GAMBL algorithm. Consequently, employing both approaches in combination also leads to only a modest increase in coverage and performance.

	GAMBL				SenseLearner				Koç University			
	none	rand.	WSA	±	none	rand.	WSA	±	none	rand.	WSA	±
noun	69.0	71.7	73.0	<b>1.3*</b>	68.7	71.0	71.8	0.8	69.3	71.6	72.6	1.0
noun	69.0	70.3	70.7	0.4	68.7	70.3	70.9	0.6	69.3	70.5	71.0	0.5
noun	69.0	73.8	74.7	0.9	68.7	73.1	73.3	0.2	69.3	73.5	74.8	<b>1.3*</b>

Table 8.13: WSD accuracy by system, using clusterings derived from alignments of WordNet to Wikipedia with Dijkstra-WSA (top), with using gloss similarity (middle) and with the similarity-based backoff (bottom). Boldface marks the best results. An asterisk marks statistically significant improvements.

While this is surprising at first glance, our analysis revealed that the WSD dataset we use differs significantly from the WSA evaluation dataset discussed in Section 3.5.1. The latter dataset (initially introduced by Meyer and Gurevych (2011)) was designed to be balanced with respect to the polysemy and corpus frequency of its lexemes. By comparison, the WSD dataset contains a lot more high-frequency, core vocabulary words such as “take” and “want”, as would be expected of an all-words WSD corpus. For these words, the corresponding Wiktionary articles are usually in a far better condition regarding their link structure as compared to mid- or low-frequency words, because they are edited and extended far more often (Meyer, 2013). This enables the pure Dijkstra-WSA approach to work effectively (i.e. with good recall) for the corresponding word senses, rendering the backoff approach less beneficial. This is also in line with the observation that recall for an alignment between GermaNet and Wiktionary is much higher than for WordNet-Wiktionary due to the much higher density of semantic relations in the German edition of Wiktionary as compared to the English one.

**WordNet-Wikipedia** As for GermaNet, our evaluation for the clustering against Wikipedia was restricted to nouns (see Table 8.13). The results for Dijkstra-WSA alone are on par with the OmegaWiki clustering (the differences are not statistically significant) and slightly better than for Wiktionary. However, as for Wiktionary we also observe many 1:1 alignments and thus few and small clusters due to the similar granularity of both resources. For instance, the noun *radiation*, has not only the expected physics-related sense covered in Wikipedia, but also less common WordNet senses from biology (“the spread of a group of organisms into new habitats”) and medicine (“the treatment of disease (especially cancer) by exposure to a radioactive substance”). As for the German setup, obscure senses (e.g. the “British band” sense of *Radiation*) are reliably disregarded by our alignment algorithm and hence do not interfere with the clustering.

The marginal differences between the results with and without the backoff are also consistent with the observations previously made for Wiktionary. For the items in the evaluation dataset, the recall of Dijkstra-WSA alone is already good, so the backoff has only modest influence on the alignments and the clustering results (see also Table 8.9).

	GAMBL				SenseLearner				Koç University			
	none	rand.	WSA	±	none	rand.	WSA	±	none	rand.	WSA	±
noun	69.0	78.3	81.2	<b>2.9*</b>	68.7	78.4	79.9	1.5	69.3	78.4	80.5	2.1*
verb	59.0	61.7	63.0	1.3	56.1	59.4	61.4	<b>2.0*</b>	54.1	57.3	59.3	<b>2.0*</b>
adj.	67.0	78.3	83.4	<b>5.1*</b>	70.4	79.1	81.6	2.5*	70.4	79.5	82.1	2.6*
all	65.2	71.7	74.4	<b>2.7*</b>	64.6	71.1	72.9	1.8*	64.1	70.4	72.5	2.1*

Table 8.14: WSD accuracy, by system and part of speech, using the best clusterings for each part of speech. Nouns and adjectives use clusterings from an alignment with OmegaWiki with backoff, and verbs from an alignment with OmegaWiki without backoff. Boldface marks the best results per POS. An asterisk marks statistically significant improvements.

	before	after	reduction (%)
noun	2.79	2.68	3.81
verb	3.57	3.49	2.25
adj.	2.71	2.61	3.76
all	2.92	2.48	3.34

Table 8.15: Reduction in average polysemy for all polysemous words in WordNet with our optimal clusterings.

**Combined approaches** Our experiments show that our clustering method is also effective for WordNet; in many configurations, we observe a significant improvement over the granularity-controlled random baseline.

As for German, one striking observation was that different configurations are preferable for different parts of speech: OmegaWiki works best for nouns and adjectives when a backoff is applied, and clustering against OmegaWiki using Dijkstra-WSA alone shows the best relative improvement for verbs. In this light, OmegaWiki can be considered the most effective of the three resources in terms of WordNet clustering. While its low coverage impairs better results, its coarse-grained sense distinctions are still the most helpful for the task at hand.

Following this insight, we also performed an additional experiment with optimal clusterings for each part of speech for English (see Table 8.14). This clustering results in a significant improvement for each part of speech (except adverbs, though these comprise only 15 of the 2041 instances in the dataset). The overall improvement averaged across systems is a statistically significant one of 2.2 percentage points. The reduction in average polysemy by POS for this clustering is given in Table 8.15.

Another important insight from our experiments is that, in general, the three WSD algorithms show similar behaviour when applying the same clusterings – the performance improvement (or lack thereof) on the different parts of speech tends to be the same for all of them. For example, Wiktionary clustering improves the performance for verbs on all systems. We interpret this as evidence that the improvements achieved by our clusterings are generally valid in the sense that they are independent of the algorithm used. However, we also note that different algorithms show different performance on different parts of speech. The GAMBL algorithm works



	<b>GAMBL + Koç University</b>			
	none	rand.	WSA	±
noun	69.0	78.3	81.2	<b>2.9*</b>
verb	54.1	57.3	59.3	<b>2.0*</b>
adj.	67.0	78.3	83.4	<b>5.1*</b>
all	65.5	70.1	73.1	<b>3.0*</b>

Table 8.16: WSD accuracy for an aggregate system which pairs the best-performing system and clustering for each part of speech on WordNet. An asterisk marks statistically significant improvements.

significantly better than the others on adjectives, and it also achieves the highest relative improvement across all configurations for nouns (although the picture is more inconsistent here). Apparently, GAMBL’s “word experts” approach which uses custom classifiers based on local context for each lemma-POS combination is effective when the degree of polysemy is relatively low. But for specific low-frequency verb lemmas, this approach struggles to find enough evidence, and these are usually just assigned the most frequent sense (Decadt et al., 2004). The Koç University system fares better in this case, as global and local context features are combined in a naive Bayes classification. The third system, SenseLearner, does not stand out in any configuration (although it reaches parity with the Koç University system for the best verb clustering), but its solid results are achieved in a minimally supervised fashion – i.e., requiring far less external knowledge than the others.

These observations suggest that not only different clusterings, but also the usage of different algorithms per part of speech might be beneficial to achieve the best performance. While this is arguably a case of overfitting to the dataset and may not apply to other scenarios, we still present the results for this ideal setup in Table 8.16. This configuration yields another boost in performance, achieving results similar to Snow et al. (2007), who reported an improvement of 3.55% over the random clustering – keep in mind though that these numbers are not fully comparable and are flawed due to the aforementioned property of their scoring which assumes only one correct sense per item (see Section 8.1.4).

## 8.2 Using aligned Wiktionary and OmegaWiki for Computer-Aided Translation

### 8.2.1 Introduction

Another use case for UBY that benefits from both sense alignments and standardization is computer-aided translation. While this idea is not yet fully implemented and a task-based evaluation is a crucial step to be taken, we still want to discuss in which way such environments could benefit from aligned resources, and what features of UBY can contribute to the overall user experience. This application scenario was one of the driving motivations for investigating the alignment between Wiktionary and OmegaWiki presented in Section 4.3, as they are both multilingual resources.

It also inspired the way translation information was integrated into UBY-LMF (see Section 7.3).

### 8.2.2 Motivation

Recently, operating internationally has emerged as an increasingly important task for governments, companies, researchers, and many other institutions and individuals. This raises a high demand for translation tools and resources. Statistical machine translation (SMT) systems are widely used nowadays (especially among layman translators), but are usually hard to adapt to specific needs as parallel texts for training are not available for many domains. Thus, SMT systems are mainly useful during the drafting phase of translating a text, or as a supplementary tool to provide additional translations for a word or phrase. High quality translations as they are needed for official documents such as contracts still require human effort and editing (Koehn, 2009; Carl et al., 2010). SMT systems are not sufficient for this purpose, since there is usually no hint of what the translations actually mean and why one alternative is preferable when only a bare probability score is provided.

To produce translations of higher quality, additional tools and resources need to be considered. Translation Memory systems became very popular for this purpose in the 1990s (Somers, 2003). They maintain a database of translations which are manually validated as correct and can be applied if the same or a similar translation is required. They can, to some extent, deal with unseen texts using fuzzy matching (i.e. with tolerance for character-level changes), but while this approach yields a high precision, it does not aid in validating translations for entirely new content and is thus mostly useful in environments where the context does not change much over time. More recently, parallel corpora have been used to identify suitable translations in context; for example, through the *Linguee*<sup>1</sup> service. While this might help to identify the correct translation, pinpointing the exact meaning can be hard because no sense definitions or any other lexicographic information is provided. Moreover, the lack of sufficiently large parallel corpora, especially for uncommon language pairs, is also an issue here.

We argue that, either to support translators directly or to improve SMT, multilingual lexical resources such as bilingual dictionaries or multilingual wordnets are required in addition to the tools mentioned. Using the information contained in those multilingual resources (such as sense definitions) makes it possible to manually or (semi-)automatically assess if a translation is appropriate in context and to perform corrections using a better suited translation found in the resource. As has been shown earlier, this is especially true for unusual language combinations and specific tasks such as cultural heritage annotation (Mörth et al., 2011; Declerck et al., 2012).

Consider, for example, the English noun *bass*. In Google Translate,<sup>2</sup> probably the most popular SMT system to date, only the music-related word sense of *bass* is considered for the example translation into German shown in Figure 8.2. None of the translation alternatives addresses the animal-related word sense, which would be

---

<sup>1</sup><http://www.linguee.com>

<sup>2</sup><http://translate.google.com>

correct in this context. Moreover, there are no sense definitions or validated usage examples for the proposed translations.

In contrast, multilingual lexical resources such as Wiktionary allow one to easily distinguish between the two word senses of *bass* and provide a vast amount of lexicographic information to help identify a good translation. Although in this case of homonymy it would be comparatively easy to pick the correct sense, distinguishing closely related senses that share the same etymology poses a much greater problem. Figure 8.3 shows an excerpt of the animal-related word sense of *bass* in Wiktionary that contains the suitable German translation *Barsch* for the example discussed above. OmegaWiki encodes another possible translation *Seebarsch* and provides additional lexicographic information. An excerpt is shown in Figure 8.4.

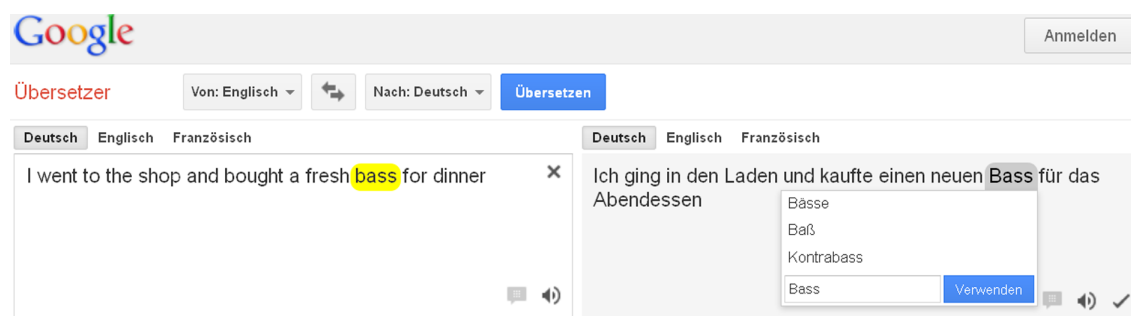


Figure 8.2: The translation alternatives for *bass* in Google Translate.

We identified the following three major requirements for a multilingual lexical resource to be useful for translation applications:


1. The resource should have a high coverage of languages and allow for continually adding or revising information. This is important for covering neologisms or domain-specific terminology, and especially for correcting improper translations or adding missing ones. Terminology-rich resources are especially important for human translators, as SMT systems cannot cope well with domain-specific texts due to the lack of training data.
2. There should be a large variety of lexicographic information types, such as sense definitions, example sentences, collocations, etc. that illustrate the use of a translation without being redundant.
3. Ideally, the resources should be seamlessly integrated into the translation environment via established standards and interfaces.

As our analysis revealed (also see Section 3.2), most expert-built resources such as WordNet fail to fulfill some or all of these requirements. First of all, they need enormous building effort and are in turn rather inflexible with regard to corrections or addition of knowledge. This effort is also the reason why for many smaller languages such resources remain small or do not even exist. Second, expert-built resources usually have a narrow scope of information types. For example, WordNet focuses on synsets and their taxonomy, but mostly disregards syntactic information,

**Etymology 2** [edit]

From Middle English *bas*, alteration of *bars*, from Old English *bærs* ("a fish, perch"), from Proto-Germanic *\*barsaz* ("perch", literally "prickly fish"), from Proto-Indo-European *\*bhars-*, *\*bharst-* ("prickle, thorn, scale"). Cognate with Dutch *baars* ("baars"), German *Barsch* ("perch"). More at *barse*.

**Pronunciation** [edit]

- enPR: bäs, IPA: /bæs/, X-SAMPA: /b{ɚ/
- Audio (US)  (file)

**Noun** [edit]

**bass** (*plural basses or bass*)

1. The **perch**; any of various marine and freshwater fish resembling the perch, all within the order of **Perciformes**.

**Usage notes** [edit]

The plural **bass** refers to multiple fish of a single variety or species, whereas **basses** refers to multiple varieties or species.

**Derived terms** [edit]

- |   |   |
|---|---|
| <ul style="list-style-type: none"> <li>• black bass</li> <li>• black sea bass</li> <li>• largemouth bass</li> <li>• sea bass</li> </ul> | <ul style="list-style-type: none"> <li>• smallmouth bass</li> <li>• spotted bass</li> <li>• striped bass</li> <li>• white bass</li> </ul> |
|---|---|

**Translations** [edit]

**perch** [hide ▲]

Select targeted languages

<ul style="list-style-type: none"> <li>• Bulgarian: <i>костур</i> <sup>(bg)</sup></li> <li>• Cherokee: <i>ᎠᎵᎠ</i> <sup>(chr)</sup></li> <li>• Croatian: <i>grgeč</i> <sup>(hr)</sup> <i>m</i></li> <li>• Dutch: <i>baars</i> <sup>(nl)</sup> <i>m</i></li> <li>• French: <i>perche</i> <sup>(fr)</sup> <i>m</i></li> <li>• German: <i>Barsch</i> <sup>(de)</sup> <i>m</i></li> <li>• Greek: <i>πέρκα</i> <sup>(el)</sup> (<i>perka</i>) <i>f</i></li> </ul>	<ul style="list-style-type: none"> <li>• Hungarian: <i>sügér</i> <sup>(hu)</sup></li> <li>• Italian: <i>branzino</i> <sup>(it)</sup> <i>m</i>, <i>spigola</i> <sup>(it)</sup> <i>m</i></li> <li>• Latvian: <i>asaris</i> <sup>(lv)</sup></li> <li>• Lithuanian: <i>ešerys</i> <sup>(lt)</sup></li> <li>• Norwegian: <i>bass</i> <sup>(no)</sup></li> <li>• Russian: <i>окунь</i> <sup>(ru)</sup> (<i>ókun'</i>) <i>m</i></li> <li>• Spanish: <i>róbalo</i> <i>m</i>, <i>lubina</i> <i>f</i>, <i>perca</i> <i>f</i> (freshwater)</li> </ul>
---	--

• Add translation  :   
 [More](#)

Figure 8.3: An excerpt of the Wiktionary entry on *bass*. <http://en.wiktionary.org/wiki/bass>

### ▼ Definition

Language	Text
Castilian	Pez marino (Percicthyidae o Centrarchidae) popular para la pesca.
Dutch	Een zeevis (Percicthyidae of Centrarchidae) die populair is als sportvis.
English	A marine fish (Percicthyidae or Centrarchidae) that is popular as game.
French	Poisson d'eau de mer (Percicthyidae or Centrarchidae) populaire pour la pêche.
Slovak	Morská ryba (Percicthyidae alebo Centrarchidae) populárna ako lovná zver.

### ▼ Synonyms and translations

Expression	
Language	Spelling
Castilian	lubina
Castilian	róbalo
Castilian	robalo
Dutch	zeebaars
English	bass
French	basse
German	Seebarsch
Italian	spigola
Japanese	バス
Portuguese	robalo
Swedish	bass

### ▼ Annotation

Property	Value
is part of theme	fish

### ▼ Class membership

Class
animal

Figure 8.4: An excerpt of OmegaWiki's defined meaning 5555 on *bass*. [http://www.omegawiki.org/DefinedMeaning:bass\\_\(5555\)](http://www.omegawiki.org/DefinedMeaning:bass_(5555))

which is in turn the focus of VerbNet. Finally, many expert-built resources utilize proprietary or non-machine readable formats, which makes the integration into a translation environment difficult.

The latter issue is solved by the modeling of expert-built and collaboratively constructed resources as described in Section 7.3, and by the subsequent integration into the unified resource UBY, which ensures seamless incorporation into translation applications. To address the other issues, we studied the collaboratively constructed resources Wiktionary and OmegaWiki in detail and described how multilingual lexical-semantic knowledge can be mined from these resources. To the best of our knowledge, these resources have not been discussed in the context of translation applications. There exists a significant amount of previous work using Wikipedia in the context of cross-lingual information retrieval for query expansion or query translation (Potthast et al., 2008; Gaillard et al., 2010; Herbert et al., 2011), but it is primarily an encyclopedic resource, which limits the amount of lexical knowledge available for the application we address here. In previous work, Müller and Gurevych (2009) discussed combining Wiktionary and Wikipedia for cross-lingual information retrieval, but in this case Wiktionary is also merely used for query expansion and most of the lexicographic knowledge encoded in it remains disregarded. However, this knowledge is essential for translation applications in order to make well-grounded decisions (McCrae et al., 2011a).

The results of our study are laid out in detail in (Meyer, 2013) for Wiktionary and in Section 3.3 for OmegaWiki. Nevertheless, we provide a brief summary of the benefits of these resources, considering the application we have in mind:

**Easy contribution.** Wiktionary and OmegaWiki are based on a Wiki system, which allows any Web user to contribute to those resources. This crowd-based construction approach is very promising for the task at hand, since the large body of collaborators can quickly adapt to new language phenomena like neologisms while at the same time ensuring a remarkable quality – this phenomenon known as the “wisdom of crowds” (Surowiecki, 2005) has already been mentioned before.

**Good coverage of languages.** These resources are open to users speaking any language, which is very beneficial to smaller languages. Meyer (2013) found, for instance, that the collaborative construction approach of Wiktionary yields language versions covering the majority of language families and regions of the world, and that it covers a vast number of domain-specific descriptions not found in wordnets.

**Free availability.** All the knowledge in these resources is available for free under non-restrictive licenses. This is a major advantage of these collaboratively constructed resources over efforts like *EuroWordNet* (Vossen, 1998), where the aligned expert-built resources are subject to restrictive licenses. Moreover, the data from these LSRs can be processed automatically as it is available in a machine-readable format.

**Versatility.** They contain multiple different lexicographic information types, such as etymological and grammatical information.

Wiktionary and OmegaWiki are aligned at the level of word senses in order to benefit from the complementary lexicographic information types. The initial, similarity-based effort was described in detail in Section 4.3, while this dataset was also investigated in the subsequent WSA chapters with improved results; the full alignment between Wiktionary and OmegaWiki is distributed as an integral part of UBY as described in Section 7.4.

Thus, we limit ourselves at this point to discussing related work in the area of LSRs for translation applications, and to describing the benefits we gain from the alignment in this context.

### 8.2.3 Related Work

Human translators traditionally utilize monolingual and bilingual dictionaries as a reference. Dictionaries provide many different kinds of lexicographic information, such as sense definitions, example sentences, collocations, idioms, etc. They are well-crafted for being used by humans, but using them computationally poses a great challenge. Although machine readable dictionaries can be processed automatically, computers are often overstrained to properly interpret the structure of an entry or resolve ambiguities that are intuitively clear to humans.

The great success of the Princeton WordNet motivated the creation of a large number of multilingual wordnets, such as *EuroWordNet* (Vossen, 1998), *BalkaNet* (Stamou et al., 2002), *MultiWordNet* (Pianta et al., 2002) or *Open Multilingual Wordnet* (Bond and Foster, 2013). While the nature of these resources seems to perfectly meet our requirements, only few of them gained a size comparable to the English WordNet or provide as many different information types as dictionaries (such as etymology, pronunciation or derived terms) due to their time-consuming and costly construction process.

The limited number of experts also prevents frequent updates with new or updated contents. Automatically induced resources based on the output of Open Information Extraction (OIE) systems such as *KnowItAll* (Banko et al., 2007) can be huge and kept up to date at any time. However, those resources are not sense disambiguated *per se* and, due to the completely automatic creation process, limited in their quality.

Regarding collaboratively constructed resources, Wikipedia has been found to be a very promising resource for a multitude of natural language processing tasks (Zesch et al., 2007; Medelyan et al., 2009). The large size of Wikipedia and the overall high quality of the articles make Wikipedia a valuable resource for translation tasks – for example, as a parallel corpus (Adafre and de Rijke, 2006) or as a source for mining bilingual terminology (Erdmann et al., 2009). However, the vast majority of information in Wikipedia is encyclopedic and almost entirely focusing on nouns. Translators also require lexicographic information types such as idioms, collocations, or usage examples as well as translations for word classes other than nouns – most importantly verbs, adjectives, and adverbs. The unified resource *BabelNet* (Navigli and Ponzetto, 2012a) covers other parts of speech as it contains not only multilingual information from Wikipedia, but also from (among others) Wiktionary and OmegaWiki, the same resources we address. However, *BabelNet* does not include all information from the stand-alone resources which might be useful for

Resource type	Information types	Lexicon size	Computational usage	Update time	Quality
Dictionaries	many	considerable	hard	long	very high
Wordnets	limited	small	easy	long	very high
OIE-based	many	huge	easy	short	low
Wikipedia	encyclopedic	large	medium	short	high
Wiktionary	many	large	medium	short	high
OmegaWiki	many	medium	easy	short	high

Table 8.17: Comparison of the advantages of different resource types (OIE = Open Information Extraction).

this application scenario, such as etymologies or pronunciations.

This is why we investigate the full integration of Wiktionary and OmegaWiki in UBY which combines the advantages of the other resources discussed above (see Table 8.17). Their joint usage offers interesting ways of utilizing the combined multilingual information, such as using ontological knowledge from OmegaWiki to enrich Wiktionary senses. We will discuss this in more detail in the following section.

## 8.2.4 Discussion of Alignment Results

The alignment between Wiktionary and OmegaWiki offers various advantages for translation applications:

- Better coverage as the lexemes and senses from both resources can be considered; this is generally true for all applications which utilize aligned resources.
- Complementary information such as additional example sentences for a sense which help choosing the correct translation or additional translations contained in the additional resource.
- Better structured translation results achieved, for example, by clustering the translations into the same language for aligned senses instead of simply considering all of them in parallel.
- Identical translations in both resources yield combined evidence and thus higher translation confidence; the redundancy in the displayed results can be avoided by collapsing these translations.

The major benefit for our purposes is the availability of additional information, and especially translations, for the aligned resources. Particularly interesting is that, as OmegaWiki is a multilingual resource by design, we obtain an alignment to multilingual synsets. This means that the (disambiguated) translations encoded here apply to the aligned Wiktionary senses. This entails that the correct translation is immediately known once the word sense in the source document can be correctly identified (either by the user or by automatic word sense disambiguation). A similar argument also holds for Wiktionary – all aligned senses from OmegaWiki benefit from the additional translations available in Wiktionary. The only disadvantage in



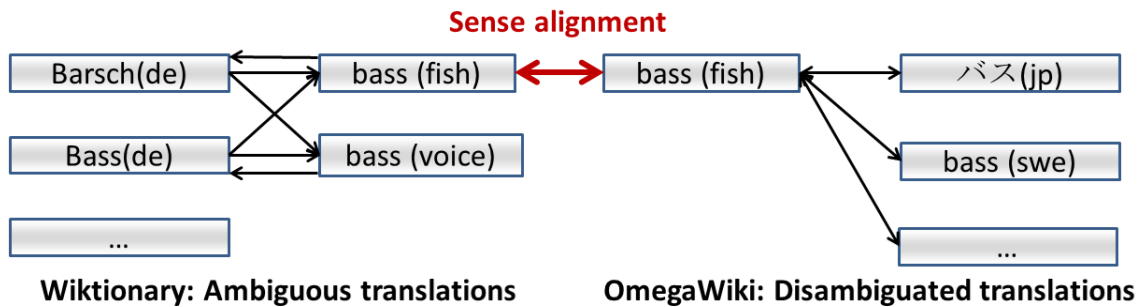


Figure 8.5: Illustration of the sense alignment between Wiktionary and OmegaWiki. As the translations in OmegaWiki are unambiguous, they directly apply to the aligned Wiktionary sense. Although this is not the case for the translations in Wiktionary, they still offer additional translation options. The ambiguity in Wiktionary is exemplified by the arrows pointing from German *Barsch* and *Bass* to both English senses of *bass* – there is no explicit link to the correct sense, only to the lexeme.

this case is that these are not disambiguated. An illustration of these benefits is given in Figure 8.5.

While a task-based evaluation is subject to future work, we would like to further explain the advantages of the derived alignment on the example introduced earlier. Consider again the noun *bass*. The word sense “A male singer who sings in the deepest vocal range” from OmegaWiki is automatically aligned with the sense “A male singer who sings in the bass range” from Wiktionary. While these two different definitions might themselves be useful for pinpointing the exact meaning of the term, there are a number of further valuable information sources:

- Wiktionary offers translations into Spanish, Dutch, Bulgarian, Tatar, Finnish, German, Greek, Hungarian, Italian, Japanese, Russian and Slovene, while OmegaWiki additionally encodes translations into French, Georgian, Korean and Portuguese. Only the Spanish translation *bajo* and the Italian translation *basso* are included in both. Thus, the alignment directly yields a significantly broader range of translations than either resource alone.
- OmegaWiki offers sense definitions of this word sense in Spanish and French, which are useful for a translator fluent in one of these languages. Moreover, the Spanish sense definition from OmegaWiki can directly be used to identify the correct sense of the Spanish translation, which is not disambiguated in Wiktionary.
- Wiktionary also offers additional information not included in OmegaWiki, such as etymology, pronunciation, and derived terms.

Table 8.18 summarizes the information that becomes available through the sense alignment of Wiktionary and OmegaWiki for our example word *bass*.

While this is only meant as an illustrative example, Table 8.19 shows statistics about the most important LMF classes regarding both single resources as well as

Resource	Translation languages	Available definitions	Additional information types
Wiktionary	12	1	5
OmegaWiki	6	3	0
Combined	16	4	5

Table 8.18: Information gain through the alignment for one sense of *bass*.

their combination; these numbers thus describe the subset of UBY which was created for this application scenario. As can be seen, even with only two languages and two resources considered, a translation resource of exceptional size with over 500,000 lexical entries and senses and well over 200,000 paradigmatic relations can be obtained. Probably most important for translation applications, we also have almost 1,600,000 instances of the `Equivalent` class, which represent the translations (as discussed in Section 7.3; a breakdown into single languages can be found in Table 3.2 in Section 3.3). In Table 8.20, we can see that over 80,000 `SenseAxis` instances are available, over 25,000 of them stemming from our alignment of the two resources. Considering the around 60,000 senses in the English OmegaWiki, we have reached a fairly dense alignment of the two resources covering about half of OmegaWiki.

In summary, a combined usage of Wiktionary and OmegaWiki in UBY fulfills the following properties:

1. **Continuously updated lexical-semantic knowledge.** The frequently updated and extended knowledge in both resources can at any time be integrated into UBY as the conversion routines into the common model need no or only minor modifications in the future. This also relieves end users from the burden of adapting their applications to changes in the underlying resources as the unified output model remains stable.
2. **High coverage.** The alignments at word sense level significantly improve upon the available information in the isolated resources, which is very valuable for translation purposes.
3. **A standardized structure.** The UBY-LMF model ensures that the resource can be queried with consistent and reliable results.
4. **Interoperability.** The resource is not only in a format which is machine readable, but it is also compliant to existing ISO standards to allow for easy reuse and integration into a translation environment.

### 8.3 Chapter Summary and Contributions

In this chapter, we presented two different applications of WSA which yield improvements for particular NLP tasks.

First, we described how alignments calculated with Dijkstra-WSA can be used for clustering fine-grained GermaNet and WordNet senses. We align them to the

Resource	LexicalEntry	Sense	SenseRelation	Equivalent
Wiktionary en	335,749	421,848	22,313	694,282
OmegaWiki en	51,715	57,921	7,157	335,173
Wiktionary de	85,575	72,752	183,684	250,674
OmegaWiki de	30,967	34,691	7,165	304,590
Total	504,006	587,212	220,319	1,584,719

Table 8.19: Statistics about the combination of Wiktionary and OmegaWiki. The *Equivalent* class represents the translations found in each resource.

Resource Pair	SenseAxis	Information source
OmegaWiki en–OmegaWiki de	58,785	Voluntary editors
OmegaWiki en–Wiktionary en	25,727	Automatic alignment
Total	84,512	

Table 8.20: Alignment statistics for Wiktionary and OmegaWiki.

three collaboratively constructed sense inventories OmegaWiki, Wiktionary and Wikipedia and exploit 1:n alignments which indicate a difference in granularity in the resources. We showed that a significant improvement in word sense disambiguation accuracy is possible with this method, which is also substantially more flexible and generic than previous approaches. We also discussed the properties of the different LSRs regarding coverage and granularity in this context, and showed that combining clusterings of different resources for different parts of speech leads to performance comparable to state-of-the-art systems, which unlike our approach require external knowledge (other than the resources used for alignment) and extensive resource-specific feature engineering. Our clusterings are freely available to the research community on our website.<sup>3</sup>

Secondly, we argued that collaboratively constructed multilingual lexical resources present a valuable source of knowledge for translation applications. They are maintained by a crowd of users, thus guaranteeing highly accurate and up to date information, while at the same time being available with almost no restrictions. We also discussed the results of the alignment we produced between them in the context of translation applications, giving illustrative examples of its benefits such as the substantial increase of coverage, especially concerning available translations. In this respect, the aligned resource outperforms either of the single resources by far.

**Contribution 1** We discuss how our WSA algorithm Dijkstra-WSA can be exploited to cluster fine-grained sense inventories, and we demonstrate that such a clustering significantly improves WSD performance for German and English, using GermaNet and WordNet as sense inventories.

**Contribution 2** We explain how an alignment between multilingual resources, specifically OmegaWiki and Wiktionary, can improve a computer-aided translation environment by increasing the coverage of senses, information types and especially translations.

<sup>3</sup><https://www.ukp.tu-darmstadt.de/data/lexical-resources/>



# Chapter 9

## Conclusions

### 9.1 Summary of the Thesis

Word sense alignment is the task of identifying equivalent senses in different lexical-semantic resources. In this thesis, we discussed this task from several angles. We laid the foundation by giving an exact definition of the problem at hand and comparing it to related tasks. Then, we presented the set of resources which we considered for alignment and analyzed them in detail, presented several algorithmic approaches to WSA, discussed the unified resource UBY which served as a motivational frame for our work, and also described several applications which benefit from sense alignments.

In order to allow the reader to fully comprehend the task of WSA and see it in a greater context, we tried not only to define the problem in Chapter 2, but also outlined the common and distinctive traits with regard to related tasks in NLP and other fields. Thereby, we established that WSA has unique requirements which have to be considered when designing algorithmic approaches for this challenge. For instance, we cannot (at least not in the general case) rely on well-defined structures or instance-based matching as it is common for ontologies or database schemata, and our setting also differs from other semantic processing tasks such as WSD or semantic relatedness calculation, which rely on different assumptions about their input and output.

For the discussion of resources in Chapter 3, we distinguished between expert-built and collaboratively constructed resources, and laid a special focus on OmegaWiki, for which we presented a detailed discussion of its content and structure. We especially considered it in relation to Wiktionary, which was constructed according to a similar paradigm, and discovered that, while OmegaWiki is substantially smaller and thus has several gaps in the coverage of information types and senses, its concept-centered, well-defined and language-agnostic structure has many properties interesting for its usage in NLP applications. For instance, it has unambiguous semantic relations and translations. Following the presentation of the resources, we performed an in-depth analysis regarding the resources' suitability and mutual compatibility for WSA. In particular, we considered glosses and structures induced via relations between senses, which are the two perennial ways to explicitly and implicitly describe concepts. We found that there are many remarkable differences between resources, and based on these findings we selected several LSR pairs which

we deemed especially interesting for investigating WSA and developing generally applicable algorithmic approaches, i.e. approaches which are suitable for a wide range of resources. Accordingly, we describe the WSA datasets these pairs participate in, and apart from the ones already used in previous work, we present four novel datasets which we created in the course of our own work. Especially interesting is the German Wiktionary-Wikipedia dataset which was derived from edits made by the Wiktionary community.

In Chapter 4, we aligned Wiktionary and OmegaWiki (the first alignment between two collaboratively constructed resources) using the similarity-based approach presented in earlier work, and we achieved comparable results. We also improved the algorithm by adding a machine translation component, which we used to align WordNet and the German part of OmegaWiki. Comparative experiments with the English OmegaWiki (i.e. for the monolingual case) revealed that only few errors were introduced by the translation component, so we assume that the idea in general is valid.

Chapter 5 mainly dealt with Dijkstra-WSA, the graph-based algorithm for word sense alignment we developed. It works on the graph structure induced by LSRs (e.g. via semantic relations or links) and exploits the intuition that related senses are located in adjacent regions of the resources, but without making assumptions about the exact nature of the edges. We show that this algorithm on its own performs competitively on 6 out of 8 evaluation datasets, and in combination with the similarity-based approach presented in Chapter 4, it achieves a statistically significant improvement over all previous work on the considered datasets. Dijkstra-WSA is language-independent and does not require any external knowledge in the form of annotated training data or corpora. We estimate parameters on a development set for optimal results, but our experiments show that reasonable results can be achieved by using default parameters based on the resource sizes and structures.

In Chapter 6, we combined similarity-based and graph-based measures for WSA in a machine learning framework, and achieved a further overall alignment quality improvement in terms of F-measure for four out of eight considered datasets. On three others, we could achieve a significant improvement in alignment precision and accuracy. We investigated not only different machine learning classifiers (where Bayesian Networks showed the most robust results), but also additional machine learning features derived directly from the LSRs. However, none of these led to further improvements. We consider this to be evidence that a joint usage of global structure as well as the content of the LSRs is indeed sufficient, and also preferable over using them separately or in a simple backoff approach. This is an important result for WSA in general, as we have shown that even when limited to the most salient properties of LSR which can be found across different languages and resources, satisfactory results are possible for LSRs which are very heterogeneous. This is in strong contrast to previous work in WSA or related fields such as ontology matching, where it is common to make specific assumptions about the algorithm input to achieve good results, e.g. by exploiting the semantics of particular relations. To conclude this chapter, we also discussed different approaches to N-way alignment (i.e. the alignment of more than two resources at once), but none of these yielded satisfactory results.

We discussed the unified LSR UBY, which provides the greater context for the

work in this thesis, in Chapter 7. First, we presented UBY-LMF, the representation format that was developed in order to reflect the structure and content of many different LSRs as accurately as possible, and we also briefly explained the LMF standard it depends upon. We demonstrated how the standardization can be operationalized, using OmegaWiki as a “walkthrough example”, and in course of this we presented the most important features of UBY-LMF. This also covered the representation of the sense alignments which are at the very heart of our work. These alignments are of course also contained in the final resource UBY, whose properties we briefly discussed.

Finally, in Chapter 8, we presented two NLP applications which can benefit from our calculated alignments. First, we used them for clustering fine-grained GermaNet and WordNet senses by exploiting 1:n alignments to OmegaWiki, Wiktionary and Wikipedia. We showed that clustering senses in this way can lead to a significant improvement in word sense disambiguation accuracy on standard evaluation datasets for German and English. We also discussed how different properties of the LSRs with regard to coverage and granularity influence the results in this context, discovering for instance that similar degrees of polysemy on a particular part of speech lead to only modest improvements. Consequently, we investigated combining clusterings of different resources for different parts of speech and achieved another increase in performance. Compared to previous work, our approach is remarkable as it is flexibly and generally applicable: it does not require external knowledge or resource-specific feature engineering, and it is completely language-independent. The second application we discussed was computer-aided translation, and we argued that the collaboratively constructed multilingual LSRs OmegaWiki and Wiktionary can be valuable sources of knowledge, and specifically additional translations, for this kind of applications, as they freely provide highly accurate and up to date information. The corresponding proof-of-concept is left to future work.

In Appendix A, we will conclude this thesis by discussing our contributions from a software engineering perspective. First, we present JOWKL, a Java-based API for OmegaWiki, which for the first time allows easy programmatic access to the LSR. Following the discussion of UBY-LMF and the final resource UBY in Chapter 7, we discuss two different access ways to UBY for which we participated in their creation: a Java-based API, which allows easy programmatic access, as well as a web interface which enables users to conveniently browse the contents of all included resources as well as the sense alignment connecting them. Finally, we will explain how the standardization efforts and the API enabled the implementation of the alignment framework we used for our experiments by providing an easy and uniform way to access the resources without the need to resort to proprietary APIs or incompatible representation formats.

## 9.2 Outlook

There are many directions for future work, some of which have been pointed out in the respective chapters. Here we provide a concise summary on the issues we aim to pursue further.

First of all, as one of our goals is creating a large-scale sense-aligned LSR, a

perennial topic is the consideration of further resource pairs for alignment, in order to eventually achieve a densely aligned resource. While, arguably, not all possible alignments seem to be immediately useful,<sup>1</sup> we continue to identify those combinations which might be beneficial for language processing, and which furthermore have interesting properties which motivate further investigations with regard to the algorithmic approaches we developed. This is in line with our overarching goal of creating a flexible and universally usable alignment approach and involves closer examination of resources which proved challenging to align (such as VerbNet, cf. Section 3.5.1), investigation of resources which have been disregarded thus far (especially IMSLex) and also coverage of new resources which will be integrated into future releases of UBY. For these, their content and structure have to be analyzed (as for the other LSRs, see Section 3.4) with regard to their applicability within the similarity- and graph-based frameworks, and new evaluation datasets need to be created, which in turn will further broaden the foundation for WSA research.

This especially includes further experiments on cross-lingual alignments; while we presented a case study for this in the present thesis (Section 4.4), there are many other configurations and language pairs to be covered to make the presented approaches universally applicable. For instance, an easily implementable idea for Dijkstra-WSA would be to apply machine translation in the candidate extraction step and leave the (inherently language-independent) graph algorithm unchanged.

Considering the creation of new datasets, we also want to further examine annotations made by the crowd for the collaboratively constructed resources. We created and used a German dataset for Wiktionary–Wikipedia alignment (see Section 3.5.2) which is the first of its kind, but we want to investigate in more detail to which extent these alignments are reliable, what steps are necessary to improve the dataset’s size and quality, and how negative examples (i.e. non-alignments) can be more reliably derived. We also plan to investigate if comparable datasets could be created for other Wiktionary language editions.

Regarding actual extensions of our algorithmic ideas, one of the important issues to tackle is the investigation of more elaborate (gloss) similarity measures, which are required by the approaches we presented in Chapters 4 and 6. While cosine and PPR similarity measures proved effective for our purposes, text similarity is a very active field of research in its own right, and it would be interesting to see how different (or compositional) similarity measures, for instance those contained in the framework DKPro-Similarity (Bär et al., 2013), would influence the results. The integration of established techniques from IR such as lexical expansions (Cholakov et al., 2014a) also seems applicable in this context. Recently, Pilehvar and Navigli (2014) achieve good results by (among other things) utilizing a modified PPR measure which operates on custom graph representations for each LSR. This is also a possible solution for using this gloss similarity measure for languages other than English, for instance by using GermaNet as underlying resource. Caselli et al. (2013) propose to use “shallow frame structures” (akin to the POS patterns we investigated as a feature in Section 6.3.1) to express similar syntactic usage of word senses – this line of research seems promising especially for the syntax-focused resources like

---

<sup>1</sup>For instance, there is no conceivable immediate incentive to align the English Wikipedia with the German syntax-focused resource IMSLex; however, the perpetual evaluation of possible application scenarios might prove us wrong in time.



VerbNet and IMSLex.

For Dijkstra-WSA, the main direction for future work is to increase recall while keeping high precision. A straightforward idea would be to apply the algorithm iteratively. As a considerable number of cross-resource edges are added to the graph during a single run, revisiting the still unaligned nodes might be worthwhile as new paths have become available. Another possible way would be to not only link monosemous lexemes during the graph construction to increase edge density, but also to create edges for polysemous ones. Laparra et al. (2010) discuss a possibility to do this with high precision. The main idea is to focus on lexemes with a low degree of polysemy and align if one of the possible senses is clearly more similar to the source sense than the other(s). If recall is still low, more polysemous lexemes can be examined. Pilehvar and Navigli (2014) adopt and extend this idea of polysemous linking to further improve their WSA approach, in combination with the novel gloss similarity measure discussed above. They build upon our work and achieve even better results, which we take as a hint that our work provides a valuable addition to the body of WSA research, inspiring further investigation and improvement of methods in the field.

Another issue we aim to address is the limitation of the graph size by the lexeme frequency. Although proving effective for our purposes, it can be reformulated using more sophisticated notions based on, for instance, TF/IDF. A weighting of edges (e.g. based on gloss similarities) has not been considered in our work, but would be easily applicable to the existing framework. The combination of graph distances and similarities has already proven effective for the machine learning approach we presented, and it would be interesting to see how an even closer interweaving of these notions might be beneficial. Following our initial experiments on N-way alignment, which have not yet yielded satisfactory results, we also want to investigate whether integration of joint knowledge from several LSRs into the machine learning approach might be helpful. For instance, the information that two senses in resources *A* and *B* share a strong resemblance to senses in another resource *C* could be expressed by additional features. An even more elaborate idea would be to move away from the relatively simple shortest-path algorithm and investigate entirely different graph-based algorithms, e.g. for matching nodes in bipartite graphs.

An idea which stems from the close relationship between WSD and WSA (which we outlined in Section 2.4) is to investigate the applicability of Dijkstra-WSA for the disambiguation of senses in texts. There are various possibilities to represent documents (or document collections) as graphs, e.g. by exploiting cooccurrences, so that it would be interesting to see if in this way a flexible and accurate WSD algorithm could be designed solely based on this structural knowledge. This work would be in line with previous WSD work, which later was applied to WSA (Laparra et al., 2010).

Apart from the work directly concerning the alignments, we will of course also continue the development of infrastructure-related aspects of UBY such as the API and the web interface. For instance, we plan to extend the UI to allow editing of the alignment information by the users. The rationale behind this is that there will always be errors resulting from automatic alignment, no matter how precise the algorithm is, so that a convenient editing interface would help to improve the quality of the underlying resource. Another goal is to enhance the visualization of alignments

across multiple resources. Right now, we use pairwise alignments between resources to create sense clusters, but as we plan to add more sense alignments to UBY in the future, the appropriate visual resolution of contradictory or invalid alignments will become necessary.

In the long run, we strive to move beyond the notion of UBY just being a union of other resources, i.e. we plan to develop it into a resource which can itself be edited and expanded. While it should be reasonable effort to implement this in the API or a GUI for single users which have their own local copy, we strongly believe that for UBY to grow and improve significantly it has to be made available to the community for collaborative editing. Wiktionary and Wikipedia show that such an approach is feasible, and we think that the release versions of UBY, containing the resources and alignments described in this thesis as a starting point, accompanied by a powerful user interface, can attract a significant amount of attention and user contributions. We are aware that for this to work additional issues like user management, versioning, hosting etc. have to be considered, but the existing examples and the lessons learned from them will prove useful in the long-term establishment of our own platform.

Regarding the actual usage of UBY (and especially the sense alignments), we are actively investigating new possible applications where, most likely, isolated resources have been used in the past. However, for this discussion of future work, we want to focus on the applications we presented in this thesis.

For the clustering of word senses, one task we intend to investigate in the future is an evaluation on different datasets, such as the MASC corpus (Passonneau et al., 2012) and the forthcoming extension to the TüBa-D/Z corpus (Henrich et al., 2013). We used WebCAGe and Senseval here to ensure comparability to previous work, but it would be insightful to see to what extent our approach is beneficial for WSD on different text types.

Another important goal is clustering against resources with a different nature than the ones we used here. LSRs such as VerbNet (Kipper Schuler, 2005), FrameNet (Ruppenhofer et al., 2010) and IMSLex (Fitschen, 2004) focus on different information types such as syntactic properties, so it would be interesting to see how the alignment affects the clusterings. While Dijkstra-WSA might be applicable for some of these resources, we already pointed out that for other resources such as VerbNet new approaches will have to be investigated.

As Dijkstra-WSA is, nevertheless, applicable to a wide range of LSRs, we would also like to investigate clustering LSRs other than GermaNet and WordNet, which are by far not the only ones with a tendency towards microdistinction of senses (Jorgensen, 1990); this is also in line with the planned investigation of further resource pairs. Not only might this improve performance when these sense inventories are used for WSD, but it might also help in the curation of these resources by identifying questionable sense distinctions. This seems especially interesting for Wiktionary and OmegaWiki, which have quite different sense granularities but whose collaborative construction model allows for quick and easy revision of entries.

Regarding improvements to the clustering approach itself, we would like to evaluate to what extent the clusters we create respect the existing taxonomic structure of WordNet and GermaNet, and, following the work by Snow et al. (2007), we want to investigate how violations of the taxonomy can be addressed in the algorithmic

approach. Moving away from the examination of single clusters and widening the scope to the global structure of the resources seems a natural extension of the graph-based approach, and might also be useful for uncovering errors and inconsistencies, not only on the sense level but also on the structural level.

Given the flexibility of our approach, a mid-term goal is to publish the clustering component as part of our alignment framework, i.e. as a free/open source library, to enable anyone to easily produce high-quality clusterings of WordNet and other resources for their own applications.

For further work on multilingual applications, we will also consider the other alignments which have been integrated into UBY in the meantime. The multilingual Wikipedia would be the next logical choice to integrate at this point, while other resources might be beneficial as well. For instance, grammatical information contained in FrameNet might also be useful for translation applications as finding a grammatically fitting translation is desirable. Another goal for future work is to include more language editions of Wiktionary and OmegaWiki and, more generally, an increased number of cross-lingual alignments in the unified resource.

A crucial point for further research is of course the actual usage of UBY in translation applications. The integration into a computer-aided translation environment or an SMT system would be particularly interesting. For this, we would be interested in collaborating with researchers from the (machine) translation community in order to assess the usefulness of aligned resources, and also to discover aspects in which further improvement is necessary, especially regarding the coverage and precision.



# Bibliography

- Adafre, S. F. and de Rijke, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of the EACL '06 Workshop 'New Text: Wikis and Blogs and Other Dynamic Text Sources'*, pages 62–69, Trento, Italy.
- Agirre, E., De Lacalle, O. L., and Soroa, A. (2009). Knowledge-based WSD on Specific Domains: Performing Better Than Generic Supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1501–1506, Pasadena, California, USA.
- Agirre, E. and Lopez de Lacalle, O. (2003). Clustering WordNet word senses. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 11–18, Borovets, Bulgaria.
- Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 33–41, Athens, Greece.
- Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38(1):135–187.
- Arvind, V., Köbler, J., Kuhnert, S., and Vasudev, Y. (2012). Approximate Graph Isomorphism. In Rován, B., Sassone, V., and Widmayer, P., editors, *Mathematical Foundations of Computer Science 2012*, volume 7464 of *Lecture Notes in Computer Science*, pages 100–111. Springer Berlin Heidelberg.
- Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The Meaning Multilingual Central Repository. In *Proceedings of the second international WordNet Conference (GWC 2004)*, pages 23–30, Brno, Czech Republic.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 86–90, Montreal, Canada.
- Banerjee, S. and Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the Third International Con-*

- ference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 136–145, London, UK.
- Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 2670–2676, Hyderabad, India.
- Bär, D. (2013). *A Composite Model for Computing Similarity Between Texts*. Phd thesis, Technische Universität Darmstadt, Darmstadt, Germany.
- Bär, D., Zesch, T., and Gurevych, I. (2012). Text Reuse Detection Using a Composition of Text Similarity Measures. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 167–184, Mumbai, India.
- Bär, D., Zesch, T., and Gurevych, I. (2013). DKPro Similarity: An Open Source Framework for Text Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 121–126, Sofia, Bulgaria. Association for Computational Linguistics.
- Bergenholtz, H., Nielsen, S., and Tarp, S. (2009). *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Linguistic Insights: Studies in Language and Communication. P. Lang.
- Berlin, J. and Motro, A. (2002). Database schema matching using machine learning with feature selection. In *Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE 2002)*, pages 452–466, London, UK. Springer-Verlag.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.
- Bhagwani, S., Satapathy, S., and Karnick, H. (2013). Merging word senses. In *Proceedings of the 8th Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-8)*, pages 11–19, Seattle, USA.
- Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge Approaches to Linguistics. Cambridge University Press.
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia: A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165.
- Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1352–1362, Sofia, Bulgaria.

- Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valletta, Malta.
- Buitelaar, P. (2000). Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *Proceedings of the NAACL-ANLP 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*, pages 14–19, Seattle, USA.
- Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards Linguistically Grounded Ontologies. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., and Simperl, E., editors, *The Semantic Web: Research and Applications*, pages 111–125, Berlin Heidelberg. Springer-Verlag.
- Burch, T. and Rapp, A. (2007). Das Wörterbuch-Netz: Verfahren - Methoden - Perspektiven. In *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006*, Historisches Forum 10, Teilband I, pages 607–627. Berlin: Humboldt-Universität zu Berlin.
- Burgun, A. and Bodenreider, O. (2001). Comparing Terms, Concepts and Semantic Classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, pages 77–82, Pittsburgh, PA, USA.
- Carl, M., Kay, M., and Jensen, K. (2010). Long-distance Revisions in Drafting and Post-editing. *Research in Computing Science – Special Issue: Natural Language Processing and its Applications*, 46:193–204.
- Caselli, T., Vieu, L., Carlo, S., and Vetere, G. (2013). Aligning Verb Senses in Two Italian Lexical Semantic Resources. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 33–41, Trento, Italy.
- Chen, J. N. and Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics*, 24(1):61–95.
- Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). *Linking linguistic resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer, Heidelberg.
- Chklovski, T. and Mihalcea, R. (2003). Exploiting agreement and disagreement of human annotators for word sense disambiguation. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 105–112, Borovets, Bulgaria.

- Cholakov, K., Biemann, C., Eckle-Kohler, J., and Gurevych, I. (2014a). Lexical Substitution Dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluations (LREC 2014)*, pages 1406–1411.
- Cholakov, K., Eckle-Kohler, J., and Gurevych, I. (2014b). Automated Verb Sense Labelling Based on Linked Lexical Resources. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 68–77, Gothenburg, Sweden.
- Cohen, W., Ravikumar, P., and Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *Proceedings of the KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 73–78, Washington DC, USA.
- Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8, Philadelphia, USA.
- Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press, Cambridge, UK.
- De Melo, G. and Weikum, G. (2008a). A Machine Learning Approach to Building Aligned Wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources*, pages 163–170, Hong Kong.
- De Melo, G. and Weikum, G. (2008b). Language as a Foundation of the Semantic Web. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, volume 401 of *CEUR WS*, Karlsruhe, Germany. CEUR.
- De Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, pages 513–522, Hong Kong, China.
- De Melo, G. and Weikum, G. (2010). Providing Multilingual, Multimodal Answers to Lexical Database Queries. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, pages 348–355, Valetta, Malta.
- Decadt, B., Hoste, V., Daelemans, W., and van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112, Barcelona, Spain.
- Declerck, T., Mörth, K., and Lendvai, P. (2012). Accessing and standardizing Wiktionary lexical entries for the translation of labels in Cultural Heritage taxonomies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2511–2514, Istanbul, Turkey.



- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- Dhamankar, R., Lee, Y., Doan, A., Halevy, A., and Domingos, P. (2004). iMAP: discovering complex semantic matches between database schemas. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pages 383–394, Paris, France.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.
- Doan, A., Madhavan, J., Domingos, P., and Halevy, A. (2003). Ontology matching: A machine learning approach. In *Handbook on Ontologies in Information Systems*, pages 397–416. Springer.
- Dolan, W. B. (1994). Word sense ambiguation: Clustering related senses. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, volume 2, pages 712–716, Kyoto, Japan.
- Eckle-Kohler, J. and Gurevych, I. (2012). Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability across languages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 550–560, Avignon, France.
- Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., and Meyer, C. M. (2012). UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 275–282, Istanbul, Turkey.
- Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. (2014). lemonUby - a large, inter-linked, syntactically-rich lexical resource for Ontologies. *Semantic Web Journal*, page (to appear).
- Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J. P., Cimiano, P., and Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.
- Engelberg, S. and Lemnitzer, L., editors (2001). *Lexikographie und Wörterbuchbenutzung*, volume 14 of *Einführungen*. Stauffenburg, Tübingen.
- Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). An Approach for Extracting Bilingual Terminology from Wikipedia. In *Database Systems for Advanced Applications*, volume 4947 of *Lecture Notes in Computer Science*, pages 380–392. Berlin/Heidelberg: Springer.
- Euzenat, J. and Shvaiko, P. (2013). *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition.

- Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.
- Ferrandez, O., Ellsworth, M., Munoz, R., and Baker, C. F. (2010). Aligning FrameNet and WordNet based on Semantic Neighborhoods. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 310–314, Valletta, Malta.
- Fillmore, C. J. (1982). *Frame Semantics*, pages 111–137. Hanshin Publishing Company, Frankfurt a. M.
- Fitschen, A. (2004). *Ein computerlinguistisches Lexikon als komplexes System*. PhD thesis, Universität Stuttgart, Stuttgart, Germany.
- Flati, T. and Navigli, R. (2012). The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary. *Journal of Artificial Intelligence Research (JAIR)*, 43:135–171.
- Flati, T. and Navigli, R. (2013). SPred: Large-scale Harvesting of Semantic Predicates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1222–1232, Sofia, Bulgaria.
- Flati, T., Vannella, D., Pasini, T., and Navigli, R. (2014). Two is bigger (and better) than one: The Wikipedia bitaxonomy project. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 945–955, Baltimore, USA.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 233–236, Genoa, Italy.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70.
- Francopoulo, G. and George, M. (2013). Model Description. In Francopoulo, G., editor, *LMF: Lexical Markup Framework*, Computer Engineering and IT, chapter 2, pages 19–40. London: Wiley-ISTE.
- Gaillard, B., Boualem, M., and Collin, O. (2010). Query translation using Wikipedia-based resources for analysis and disambiguation. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*, Saint-Raphael, France.
- Garoufi, K., Zesch, T., Gurevych, I., et al. (2008). Graph-theoretic analysis of collaborative knowledge bases in natural language processing. In *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, Karlsruhe, Germany.

- Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.
- Giunchiglia, F., Shvaiko, P., and Yatskevich, M. (2004). S-Match: an Algorithm and an Implementation of Semantic Matching. In Bussler, C., Davies, J., Fensel, D., and Studer, R., editors, *The Semantic Web: Research and Applications*, volume 3053 of *Lecture Notes in Computer Science*, pages 61–75. Springer Berlin Heidelberg.
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, Avignon, France.
- Gurevych, I. and Kim, J., editors (2012). *The People’s Web Meets NLP: Collaboratively Constructed Language Resources*. Theory and Applications of Natural Language Processing. Springer.
- Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., and Zesch, T. (2007). Darmstadt Knowledge Processing Repository Based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.
- Hamp, B. and Feldweg, H. (1997). GermaNet – A lexical-semantic net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.
- Hartmann, S. and Gurevych, I. (2013). FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, volume 1, pages 1363–1373, Sofia, Bulgaria.
- Hellmann, S., Brekle, J., and Auer, S. (2013). Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud. In Takeda, H., Qu, Y., Mizoguchi, R., and Kitamura, Y., editors, *Semantic Technology*, volume 7774 of *Lecture Notes in Computer Science*, pages 191–206. Springer Berlin Heidelberg.
- Henrich, V. and Hinrichs, E. (2010). Standardizing wordnets in the ISO standard LMF: WordNet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 456–464, Beijing, China.

- Henrich, V. and Hinrichs, E. (2012). A Comparative Evaluation of Word Sense Disambiguation Algorithms for German. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 576–583, Istanbul, Turkey.
- Henrich, V., Hinrichs, E., and Barkey, R. (2013). Extending the TüBa-D/Z treebank with GermaNet sense annotation. In Gurevych, I., Biemann, C., and Zesch, T., editors, *Proceedings of the 25th Conference of the German Society for Computational Linguistics (GSCL 2013)*, volume 8105 of *Lecture Notes in Artificial Intelligence*, pages 89–96. Springer.
- Henrich, V., Hinrichs, E., and Vodolazova, T. (2011). Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. In *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, pages 126–130, Poznan, Poland.
- Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). WebCAGe – A Web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 387–396, Avignon, France.
- Herbert, B., Szarvas, G., and Gurevych, I. (2011). Combining query translation techniques to improve cross-language information retrieval. In *Proceedings of the 33rd European Conference on Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 712–715. Springer, Berlin/Heidelberg.
- Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27. (Introduction to the special issue “Artificial Intelligence, Wikipedia and Semi-Structured Resources”).
- Hripcsak, G. and Rothschild, A. S. (2005). Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.
- Ide, N. (2006). Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Gelbukh, A., editor, *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2006)*, volume 3878 of *Lecture Notes in Computer Science*, pages 13–27. Springer.
- Ide, N. and Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China.
- Ide, N. and Wilks, Y. (2006). Making Sense About Sense. In Agirre, E. and Edmonds, P., editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*, chapter 3. Springer.
- ISO24613 (2008). *Language resource management – Lexical markup framework (LMF)*. Number ISO 24613:2008. Geneva, International Organization for Standardization.

- ISO7098 (1991). *Information and documentation – Romanization of Chinese*. Number ISO 7098:1991. Geneva, International Organization for Standardization.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.
- Johansson, R. and Nugues, P. (2007). Using WordNet to extend FrameNet coverage. In *Building Frame Semantics Resources for Scandinavian and Baltic Languages*, pages 27–30. Department of Computer Science, Lund University.
- Jorgensen, J. C. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.
- Kang, J. and Naughton, J. F. (2003). On schema matching with opaque column names and data values. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD 2003)*, pages 205–216, San Diego, California.
- Kilgarriff, A. (2010). A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of the NAACL-HLT 2010 Demonstration Session*, pages 21–24, Los Angeles, CA, USA.
- Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extending VerbNet with Novel Verb Classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1027–1032, Genoa, Italy.
- Kipper Schuler, K. (2005). *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.
- Kirschner, C. (2012). Kombination mehrerer lexikalisch-semantischer Ressourcen durch multiple Alignments von Wortbedeutungen. Master thesis, Technische Universität Darmstadt.
- Klein, W. and Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). *Lexicographica*, 26:79–96.
- Knight, K. and Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI 1994)*, pages 773–778, Menlo Park, CA, USA.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, page 79–86, Phuket, Thailand.
- Koehn, P. (2009). A Process Study of Computer Aided Translation. *Machine Translation*, 23(4):241–263.
- Kohomban, U. S. and Lee, W. S. (2005). Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 34–41, Ann Arbor, USA.

- Krizhanovsky, A. (2012). A quantitative analysis of the English lexicon in Wiktionaries and WordNet. *International Journal of Intelligent Information Technologies*, 8(4):13–22.
- Kwong, O. Y. (1998). Aligning WordNet with Additional Lexical Resources. In Harabagiu, S., editor, *Proceedings of the COLING-ACL '98 Workshop 'Usage of WordNet in Natural Language Processing Systems'*, pages 73–79, Montreal, QC, Canada.
- Laparra, E., Rigau, G., and Cuadros, M. (2010). Exploring the integration of WordNet and FrameNet. In *Proceedings of the 5th Global WordNet Conference (GWC 2010)*, Mumbai, India.
- Le, B. T., Dieng-Kuntz, R., and Gandon, F. (2004). On ontology matching problems - for building a corporate semantic web in a multi-communities organization. In *Proceedings of the 4th International Conference on Enterprise Information Systems (ICEIS 2004)*, pages 236–243.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, pages 24–26, Toronto, Canada.
- Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, USA.
- Lew, R. (2011). Online dictionaries of English. In Fuertes-Olivera, P. A. and Bergenholtz, H., editors, *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, pages 230–250. London/New York: Continuum.
- Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the Semantic Web*, pages 251–263. Springer.
- Màrquez, L., Exsudero, G., Martínez, D., and Rigau, G. (2006). Supervised Corpus-Based Methods for WSD. In *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, pages 167–216. Springer.
- Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 1:151–164.
- Matuschek, M. and Gurevych, I. (2014). High performance word sense alignment by joint modeling of sense distance and gloss similarity. In *Proceedings of the the 25th International Conference on Computational Linguistics (COLING 2014)*, page (to appear), Dublin, Ireland.

- Matuschek, M., Meyer, C. M., and Gurevych, I. (2013). Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Computer-Aided Translation. *Translation: Computation, Corpora, Cognition. Special Issue on "Language Technology for a Multilingual Europe"*, 3(1):87–118.
- Matuschek, M., Miller, T., and Gurevych, I. (2014). A Language-independent Sense Clustering Approach for Enhanced {WSD}. In Ruppert, J. and Faaß, G., editors, *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)*, pages 11–21, Hildesheim. Universitätsverlag Hildesheim.
- McCarthy, D. (2006). Relating WordNet senses for word sense disambiguation. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 17–24, Trento, Italy.
- McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., and Cimiano, P. (2011a). Combining Statistical and Semantic Approaches to the Translation of Ontologies and Taxonomies. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-5*, pages 116–125, Portland, Oregon. Association for Computational Linguistics.
- McCrae, J., Montiel-Ponsoda, E., and Cimiano, P. (2012). Integrating WordNet and Wiktionary with Lemon. In *Linked Data in Linguistics*, pages 25–34. Berlin/Heidelberg: Springer.
- McCrae, J., Spohr, D., and Cimiano, P. (2011b). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *The Semantic Web: Research and Applications*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer Berlin / Heidelberg.
- McFate, C. J. and Forbus, K. D. (2011). NULEX: an open-license broad coverage lexicon. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 363–367, Portland, OR, USA.
- Medelyan, O., Legg, C., Milne, D., and Witten, I. H. (2009). Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754.
- Meyer, C. M. (2013). *Wiktionary: The Metalexigraphic and the Natural Language Processing Perspective*. PhD thesis, Technische Universität Darmstadt, <http://tuprints.ulb.tu-darmstadt.de/3654/>.
- Meyer, C. M. and Gurevych, I. (2010). Worth its Weight in Gold or Yet Another Resource? A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Gelbukh, A., editor, *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 6008 of *Lecture Notes in Computer Science*, pages 38–49. Berlin/Heidelberg: Springer.
- Meyer, C. M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *Proceedings*

of the 5th International Joint Conference on Natural Language Processing (*IJCNLP 2011*), pages 883–892, Chiang Mai, Thailand.

- Meyer, C. M. and Gurevych, I. (2012a). OntoWiktionary: Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In Pazienza, M. T. and Stellato, A., editors, *Semi-Automatic Ontology Development: Processes and Resources*, chapter 6, pages 131–161. IGI Global, Hershey, PA, USA.
- Meyer, C. M. and Gurevych, I. (2012b). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford University Press.
- Mihalcea, R. (2007). Using Wikipedia for automatic word sense disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, pages 196–203, Rochester, NY, USA.
- Mihalcea, R. and Faruque, E. (2004). SenseLearner: Minimally supervised word sense disambiguation for all words in open text. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 155–158.
- Mihalcea, R. and Moldovan, D. I. (2001a). Automatic generation of a coarse grained WordNet. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 454–459, Pittsburgh, USA.
- Mihalcea, R. and Moldovan, D. I. (2001b). eXtended WordNet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100, Pittsburgh, PA, USA.
- Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005). SKOS Core: Simple Knowledge Organisation for the Web. In *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice*, DCMI '05, pages 1:1–1:9. Dublin Core Metadata Initiative.
- Miller, T., Erbs, N., Zorn, H.-P., Zesch, T., and Gurevych, I. (2013). DKPro WSD: A generalized UIMA-based framework for word sense disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 37–42, Sofia, Bulgaria.
- Miller, T. and Gurevych, I. (2014). WordNet–Wikipedia–Wiktionary: Construction of a Three-way Alignment. In *Proceedings of the 9th International Conference on Language Resources and Evaluations (LREC 2014)*, pages 2094–2100, Reykjavik, Iceland.
- Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, Chicago, IL, USA.



- Mörth, K., Declerck, T., Lendvai, P., and Váradi, T. (2011). Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. In *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web*, pages 80–85, Bonn, Germany.
- Müller, C. and Gurevych, I. (2009). Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, , September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 219–226. Springer, Berlin/Heidelberg.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press.
- Navigli, R. (2006). Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING 2006)*, pages 105–112, Sydney, Australia.
- Navigli, R. (2009a). Using Cycles and Quasi-Cycles to Disambiguate Dictionary Glosses. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 594–602, Athens, Greece.
- Navigli, R. (2009b). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69.
- Navigli, R. and Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.
- Navigli, R. and Ponzetto, S. P. (2012b). BabelNetXplorer: a platform for multilingual lexical knowledge base access and exploration. In *Proceedings of the 21st international conference on World Wide Web (WWW 2012)*, pages 393–396, Lyon, France.
- Navigli, R. and Ponzetto, S. P. (2012c). BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2012)*, Toronto, Canada.
- Navigli, R. and Ponzetto, S. P. (2012d). Joining forces pays off: Multilingual Joint Word Sense Disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 1399–1410, Jeju, Korea.
- Navigli, R. and Ponzetto, S. P. (2012e). Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 67–72, Jeju, Korea.

- Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(7):1075–1086.
- Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, 70(5):56–65.
- Niemann, E. and Gurevych, I. (2011). The People’s Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS 2011)*, pages 205–214, Oxford, UK.
- Padró, M., Bel, N., and Neculescu, S. (2011). Towards the Automatic Merging of Lexical Resources: Automatic Mapping. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*, pages 296–301, Hissar, Bulgaria.
- Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex-09)*, pages 9–15, Pisa, Italy.
- Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different sense granularities for different applications. In Porzel, R., editor, *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, USA.
- Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.
- Passonneau, R. J., Baker, C. F., Fellbaum, C., and Ide, N. (2012). The MASC word sense corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3025–3030, Istanbul, Turkey.
- Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.
- Peters, W., Peters, I., and Vossen, P. (1998). Automatic sense clustering in EuroWordNet. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, pages 409–416, Granada, Spain.
- Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International WordNet Conference*, pages 293–302, Mysore, India.
- Pilehvar, M. T. and Navigli, R. (2014). A Robust Approach to Aligning Heterogeneous Lexical Resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 468–475, Baltimore, USA.

- Ponzetto, S. P. and Navigli, R. (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 2083–2088, Pasadena, CA, USA.
- Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1522–1531, Uppsala, Sweden.
- Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval: 30th European Conference on IR Research*, volume 4956 of *Lecture Notes in Computer Science*, pages 522–530, Berlin/Heidelberg. Springer.
- Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing (ICSC 2007)*, pages 517–526, Washington, DC, USA.
- Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30.
- Reed, S. L. and Lenat, D. B. (2002). Mapping Ontologies into CYC. In *Proceedings of the AAAI Workshop on Ontologies and the Semantic Web*, pages 1–6, Edmonton, AB, Canada.
- Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference for Artificial Intelligence (IJCAI 1995)*, pages 448–453, Montreal, Canada.
- Resnik, P. and Yarowsky, D. (2000). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.
- Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence: Proceedings of the Third International Atlantic Web Intelligence Conference*, volume 3528 of *Lecture Notes in Computer Science*, pages 380–386. Berlin/Heidelberg: Springer.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.
- Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

- Schöning, U. (1988). Graph isomorphism is in the low hierarchy. *Journal of Computer and System Sciences*, 37(3):312–323.
- Serasset, G. (2012). DBnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2466–2472, Istanbul, Turkey.
- Shi, L. and Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference*, volume 3406 of *Lecture Notes in Computer Science*, pages 100–111. Berlin/Heidelberg: Springer.
- Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.
- Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1005–1014, Prague, Czech Republic.
- Snyder, B. and Palmer, M. (2004). The English all-words task. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 41–43, Barcelona, Spain.
- Soanes, C. and Stevenson, A., editors (2003). *Oxford Dictionary of English*. Oxford University Press.
- Somers, H. (2003). Translation memory systems. In *Computers and Translation: A translator’s guide*, volume 35 of *Benjamins Translation Library*, chapter 3, pages 31–47. Amsterdam: John Benjamins Publishing.
- Soria, C., Monachini, M., and Vossen, P. (2009). WordNet-LMF: fleshing out a standardized format for WordNet interoperability. In *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pages 139–146, Palo Alto, California, USA.
- Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufi, D., Koeva, S., Totkov, G., Dutoit, D., and Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the First International WordNet Conference*, pages 12–14, Mysore, India.
- Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). YAGO: a Core of Semantic Knowledge. In *Proceedings of the 16th international conference on World Wide Web (WWW 2007)*, pages 697–706, Banff, Alberta, Canada.
- Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Anchor Books.
- Tomuro, N. (2001). Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 1010–1017, Pittsburgh, USA.

- Toral, A., Bracale, S., Monachini, M., and Soria, C. (2010). Rejuvenating the Italian WordNet: Upgrading, Standarising, Extending. In *Proceedings of the 5th Global WordNet Conference*, pages 260–266, Mumbai, India.
- Toral, A., Ferrandez, O., Agirre, E., and Munoz, R. (2009). A study on linking Wikipedia categories to WordNet synsets using text similarity. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pages 449–454, Borovets, Bulgaria.
- Toral, A., Muñoz, R., and Monachini, M. (2008). Named Entity WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 741–747, Marrakech, Morocco.
- Veale, T., Seco, N., and Hayes, J. (2004). In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1333–1338, Geneva, Switzerland.
- Vossen, P., editor (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.
- Yatskevich, M. and Giunchiglia, F. (2004). Element level semantic matching using WordNet. In *Meaning Coordination and Negotiation Workshop at 3rd International Semantic Web Conference (ISWC 2004)*, pages 37–48, Hiroshima, Japan.
- Yuret, D. (2004). Some experiments with a naive Bayes WSD system. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 265–268, Barcelona, Spain.
- Zaenen, A., Karttunen, L., and Crouch, R. (2005). Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, USA.
- Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen / Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 197–205.
- Zesch, T., Müller, C., and Gurevych, I. (2008). Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI 2008)*, pages 861–867, Chicago, IL, USA.



# Appendix A

## Implemented Software

### A.1 Java OmegaWiki Library (JOWKL)

As we pointed out in Section 3.3, access to OmegaWiki was originally only possible via its web interface, or by downloading it as database dump<sup>1</sup> and resorting to using plain SQL.

To make integration into NLP applications possible, and especially as a preparation for integration into UBY, we decided to create a more convenient way to access

---

<sup>1</sup>[http://www.omegawiki.org/Help:Downloading\\_the\\_data#SQL\\_Database\\_dump](http://www.omegawiki.org/Help:Downloading_the_data#SQL_Database_dump)

```
String ow_host = "localhost";
String ow_db = "OmegaWikiDB";
String ow_user = "user";
String ow_pass = "pwd";
String db_driver = "com.mysql.jdbc.Driver";
String db_vendor = "mysql";
int ow_language= OWLanguage.English;
DatabaseConfiguration dbConfig_ow =
    new DatabaseConfiguration(
        ow_host,
        ow_db,
        db_driver,
        db_vendor,
        ow_user,
        ow_pass,
        ow_language);
OmegaWiki ow = new OmegaWiki(dbConfig_ow);
```

Figure A.1: Establishing a connection to the OmegaWiki DB.

it programmatically. The result is JOWKL<sup>2</sup> (Java OmegaWiki Library), a free,<sup>3</sup> Java-based application programming interface that enables access to all information in OmegaWiki. The core development goal is to enable fast and efficient access to all of the information available in an OmegaWiki database dump without the need for preprocessing it. The only requirement is to download and import it into a local database.

The main (and straightforward) principle of the API is to directly reflect the concepts (represented as tables) in the OmegaWiki database as Java classes, and allow access to the contained information via methods which encapsulate appropriate SQL statements. On top of that, we provide a range of convenience methods for aggregating information which is spread out over several database tables. Here, we provide some code examples to illustrate how the API is set up, and how typical requests are handled.

First, a connection to the previously imported database needs to be established, which is in turn used to create an `OmegaWiki` object (see Figure A.1).

Now, we can easily retrieve all senses (i.e. defined meanings) for the word *table*, where `ow_language` encodes the language of the query word according to the ISO-639 language code:

```
Set<DefinedMeaning> meanings =
    ow.getDefinedMeaningByWord("table", ow_language);
```

Making use of the `DefinedMeaning` objects, we can now easily access various information snippets:

**Definitions in arbitrary languages.** The desired translation language can again be provided via a parameter.

```
Set<TranslatedContent> glosses = dm.getGlosses(ow_language);
for (TranslatedContent tc : glosses)
{
    System.out.println("Definiton: "+tc.getGloss());
}
```

**Translations into arbitrary languages.** Here, we also print the name of the language translated to. Note that, as explained earlier, synonyms are treated as translations in the same language, which is why the corresponding object is dubbed `SynTrans`.

---

<sup>2</sup><http://code.google.com/p/jowkl/>

<sup>3</sup>JOWKL is licensed under the Apache Software License (ASL) version 2.



```

Set<SynTrans> translations = dm.getSynTranases();
for (SynTrans st :translations)
{
    System.out.println(
        "Language: "+
        OWLanguage.getName(st.getSyntrans().getLanguageId()+
        " Translation: "+
        st.getSyntrans().getSpelling());
}

```

**Relations to other concepts.** In this case, the targets of the links are encoded as `Integer` value, as the internal identifiers are numerical. The type of the relation between concepts (such as hyponymy) is also available.

```

Map<DefinedMeaning,Integer> links = dm.getDefinedMeaningLinksAll();
for (DefinedMeaning dm_target : links.keySet())
{
    System.out.println(
        DefinedMeaningLinkType.getName(links.get(dm_target))+
        " relation with target "+
        dm_target.getSpelling());
}
}

```

Many more access methods have been implemented. Further information can be obtained from the API documentation on the Google Code page.<sup>4</sup>

## A.2 Access to UBY

### A.2.1 UBY-API

As pointed out in Chapter 7, easy access to information available in sense-aligned LSRs is crucial for their acceptance and use in NLP. While single LSRs and their APIs are reasonably well understood, researchers face the problem of using them in an orchestrated manner. Thus, for convenient access to UBY, we implemented a Java-based API built around the Hibernate framework, which is the foundation of the UBY database (see Section 7.4). Our main design principle is to keep the access to the resource as simple as possible, despite the rich and complex structure of UBY. To this end, we directly represent instances of the UBY-LMF model as Java objects, providing methods for direct access to their attributes and related objects. On top of this, we add a large number of convenience methods to aggregate information which is related, but spread out over several classes: this is internally realized via

---

<sup>4</sup><https://code.google.com/p/jowkl/>

joint database tables which are temporarily stored in memory after the first access. For instance, we can directly iterate over all lexical entries (and their senses) with a particular part of speech.

Another important design aspect is to ensure that the functionality of the individual, resource-specific APIs or user interfaces is mirrored in the UBY-API. This enables porting legacy applications to the new resource. As an example, see the corresponding UBY-API operations for the most important operations in the WordNet API in Table A.1.

<b>WordNet function</b>	<b>UBY function</b>
<b>Dictionary</b> getIndexWord(pos, lemma)	<b>UBY</b> getLexicalEntries(pos, lemma)
<b>IndexWord</b> getLemma()	<b>LexicalEntry</b> getLemmaForm()
<b>Synset</b> getGloss() getWords()	<b>Synset</b> getDefinitionText() getSenses()
<b>Pointer</b> getType()	<b>SynsetRelation</b> getRelName()
<b>Word</b> getPointers()	<b>Sense</b> getSenseRelations()

Table A.1: Some equivalent operations in the WordNet-API and the UBY-API.

A notable aspect of importing resources into UBY is that the naming conventions change, while the content remains the same. For instance, an `IndexWord` in WordNet becomes a `LexicalEntry` in UBY. We believe that this harmonization of the terminology leads to more consistent class names and in turn to a more intuitive understanding and usage of the single resources by end users. More importantly, it is fundamental to the structural interoperability of resources.

While it is possible to limit access to single resources by a parameter and thus mimic the behavior of the legacy APIs (e.g. only retrieve synsets and their relations from WordNet, see Figure A.2), the true benefit of the UBY-API becomes

```

Iterator<Synset> sIt = uby.getSynsetIterator(WordNet);
while(sIt.hasNext()){
    Synset s = synsetIterator.next();
    System.out.println("Synset: "+s.getId());
    for(SynsetRelation rel : s.getSynsetRelations()){
        System.out.println("Relation: "+rel.getRelType()+
            " "+rel.getSource().getId()+" "+rel.getTarget().getId());
    }
}

```

Figure A.2: Accessing WordNet in the UBY-API.

```

for(LexicalEntry le : uby.getLexicalEntries("go", EPartOfSpeech.verb)){
    System.out.println("Lemma: "+le.getLemmaForm());
    for(Sense s : le.getSenses()){
        System.out.println("- Sense: "+s.getId());
        Synset ss = s.getSynset();
        System.out.println ("-- Synset: "+synset.getId());
    }
}
}

```

Figure A.3: Accessing knowledge from multiple LSRs in the UBY-API.

visible when no such constraints are applied. In this case, all imported resources are queried to get one combined result, while retaining the source of the respective information. Figure A.3 demonstrates an example of how to retrieve all senses and the corresponding synsets of *go*. On top of this, the information about existing sense alignments across resources can be accessed via **SenseAxis** relations, so that the returned combined result covers not only the lexical, but also the sense level. This yields the structural and semantic interoperability of multiple resources which was the core objective of UBY.

## A.2.2 UBY Web Interface

While easy programmatic access to unified resources is crucial for employing them in NLP tasks (for instance, as we will explain in Appendix A.2.3, in WSA), the initial step of determining their added value for particular tasks is a challenge in itself, because it is not intuitively clear what kind of information is available in what resource and how it can be related and exploited by human users and machines. In other words, what is also required are tools for qualitative and exploratory examination of the aligned resources. Thus, we created a web interface which directly allows accessing the UBY database in a browser without the need to download and install the database or API.

Its two main features are as follows:

- A graph-based visualization of sense alignments between the LSRs integrated in UBY. Different senses of the same lemma which are aligned across LSRs are grouped. This allows intuitively exploring and assessing the individual senses across resource boundaries.
- A textual view for uniformly examining lexical information in detail. For a given lemma, all senses available in UBY can be retrieved and the information attached to them can be inspected in detail. Additionally, the user can compare any two senses in a detailed contrasting view. For aligned senses, this enables the immediate discovery and examination of complementary lexical information from different LSRs.

Thus, we believe that the UBY web interface can become a valuable tool for

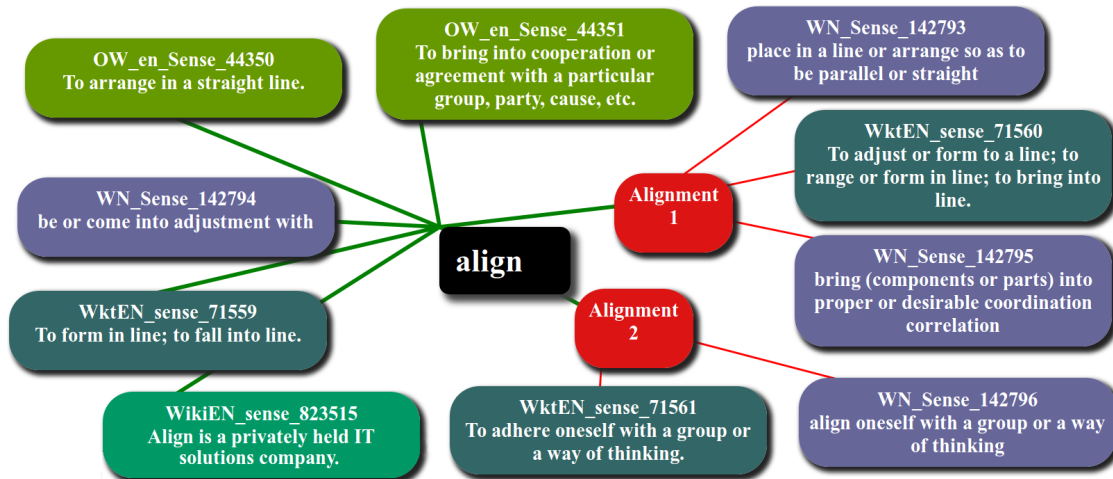


Figure A.4: In the visual view, transitive sense alignments (i.e. alignments that connect more than two senses) are expressed via alignment nodes. Sense nodes are color-coded by resource, and a click on a sense ID opens the detailed textual view.

NLP researchers using sense-aligned LSRs in their particular tasks, as well as for researchers directly working on the improvement of LSRs and their alignments.

## Technical basis

### Visual view

The natural entry point to the visual view is the search box for a lemma,<sup>5</sup> and the result is a graph, with the query lemma as the central node and the retrieved senses as nodes attached to it (see Figure A.4). The sense nodes are coloured according to the source LSRs in order to visualize which LSRs contain senses for a particular lemma. Optionally, single LSRs can be hidden based on a selection list, and to keep the view compact, the definition is only displayed when a node is clicked.

The sense alignments between LSRs available in UBY are represented by *alignment nodes*, which are displayed as hubs connecting aligned senses. For generating the alignment nodes, we cluster senses based on their pairwise alignments and include all senses which are directly or transitively aligned. Thus, the visual view provides a direct visualization of which and how many senses from different LSRs are connected in this way, and how many unique sense clusters for a lemma exist in UBY. This enables the user to intuitively assess the possible information gain provided by the sense alignments, as well as their validity.

In Figure A.4, we show the grouping of senses for the verb *align*; their definitions confirm that the alignments are plausible. If a user wants to inspect a specific sense in more detail, a click on the link within a sense node opens the textual view described below.

The screenshot shows the UBY interface for the lemma 'align'. On the left, there are two panels: 'Text Browser Visual Browser' and 'Resources (2)'. The 'Resources' panel lists various LSRs like FrameNet, OmegaWikide, OmegaWikien, VerbNet, WikipediaDE, Wikipedia, WiktionaryDE, WiktionaryEN, and WordNet. The main area displays a list of senses (1) with their definitions and LSRs. A 'Sense Comparison View' panel (4) is at the bottom left, containing a 'Compare' button. On the right, a detail panel (3) for the sense 'align' (WN\_Sense\_142795) provides lexical, semantic, and syntactic information, as well as sense alignments (5).

Figure A.5: Standard search result. (1) List of senses with definitions from selected LSRs. (2) LSR selection panel. (3) Detail panel with more information about a selected sense. (4) “Drag and drop” panel for directly comparing two senses. (5) Sense alignments for the selected sense.

The detailed sense view for 'align' (WN\_Sense\_142793) is shown. It includes a link back to the search results page. The information is organized into several sections:
 

- Lexical Information:**
  - Sense Examples:** 1. align the car with the curb, 2. align the sheets of paper on the table
  - Semantic Labels:** 1. verb.change
- Syntactic Information:** 1. subject\_nounPhrase align complement\_nounPhrase
- Semantic Information:**
  - Semantic Roles:** 1. somebody, 2. something
  - Semantic Argument >> Semantic Labels:** 1. somebody, 2. something
  - Synset:** WN\_Synset\_84376. Definition: place in a line or arrange so as to be parallel or straight
  - Sense Relation:** 1. Relation type: antonym. Source Sense: WN\_Sense\_142793 Destination Sense: WN\_Sense\_109405
- Sense Alignment:** 1. WktEN\_sense\_71560

Figure A.6: The detailed sense view offers all available information, including links to other senses via alignments or semantic relations.

## Textual view

While the query mechanism for the textual view is the same as for the visual view, in this case the interface returns a list of senses (see (1) in Figure A.5), including definitions, available for this lemma either in all LSRs, or only those selected by the user (2). The retrieved senses are grouped by LSR, and a maximum of two senses per LSR is displayed initially with the option of expanding this to the full list, in order to keep the overview clean and compact. If no result is found for a resource due to coverage gaps this is also explicitly indicated. Additionally, the LSRs are

<sup>5</sup>Filtering by POS is to be included in a future release.

<b>align</b> (verb) WN_Sense_142793		<b>align</b> (verb) WktEN_sense_71560	
place in a line or arrange so as to be parallel or straight	(1)	To adjust or form to a line; to range or form in line; to bring into line.	(1)
<b>Lexical Information:</b>	(2)	<b>Lexical Information:</b>	(2)
<b>Sense Examples:</b> 1. align the car with the curb 2. align the sheets of paper on the table		<b>Semantic Labels:</b> 1. transitive	
<b>Semantic Labels:</b> 1. verb.change			
<b>Semantic Information:</b>	(3)	<b>Sense Alignment:</b>	(4)
<b>Synonym:</b> 1. aline ( <i>Compare</i> ) 2. adjust ( <i>Compare</i> ) 3. line up ( <i>Compare</i> )		1. WN_Sense_130408 ( <i>Compare</i> ) 2. WN_Sense_142793 ( <i>Compare</i> ) 3. WN_Sense_47418 ( <i>Compare</i> ) 4. WN_Sense_74413 ( <i>Compare</i> ) 5. WN_Sense_140939 ( <i>Compare</i> ) 6. WN_Sense_142795 ( <i>Compare</i> ) 7. WN_Sense_49861 ( <i>Compare</i> )	
<b>Sense Relation:</b> 1. Relation Name: antonym Destination Sense: WN_Sense_109405			
<b>Sense Alignment:</b>	(4)		
1. WktEN_sense_71560 ( <i>Compare</i> )			

Figure A.7: In the sense comparison view, detailed information for two arbitrary senses can be inspected. Below the definition for each sense (1), lexical (2) and semantic (3) information is listed if available. Note the alignment sections (4) which contain links to the aligned senses, as well as links to open this comparison view for another pair of senses immediately.

colour-coded like in the visual view.

By clicking on a sense entry, an expanded view is opened on the right-hand side (3) to show more detailed information attached to the sense (e.g. sense examples). As with the initial query result, the single sections are shortened when there are more than two results, with the option of expanding them. Optionally, a full screen view can be opened which allows the user to explore even more information associated with the selected sense. In this detailed view of a sense, it is also possible to navigate to other senses by following the hyperlinks, e.g. for following sense alignments across LSRs or semantic relations within an LSR (see Figure A.6).

For comparing the information attached to two senses in parallel, we integrated the option to open a comparison view. For this, the user can directly drag and drop two senses to a designated area of the UI to compare them ((4) in Figure A.5), or click the *Compare* link which appears next to regular links to other senses.

The advantage of the comparison view is illustrated in Figure A.7: as the information is presented in a uniform way (due to the standard-compliant representation of UBY), a user can easily compare the information available from different LSRs without having to use different tools, terminologies, and UIs. In particular, for senses that are aligned across LSRs, the user can immediately detect complementary information, e.g. if a Wiktionary sense does not have sense examples but the aligned WordNet sense does, this additional information becomes directly accessible. As it is also possible to compare senses within a single resource, e.g. to examine differences between senses of the same lemma, this comparison mechanism also constitutes a significant advantage over existing UIs. These usually allow only one entry to be inspected at a time, which makes comparing entries cumbersome. To our knowledge, such a contrasting view of two *word senses* has not been offered by any resource or UI so far. In combination with the visual view, this feature enables an in-depth

assessment of the resources and alignments.

### A.2.3 Building a WSA Framework with UBY

As we laid out in Chapter 7, UBY was designed in order to facilitate orchestrated usage of resources in NLP applications, with combined knowledge via sense alignments being one of its major assets. Creating these alignments was the main focus of our work, and after presenting the details of our efforts in the course of this thesis, it is important to point out that the standardization effort described earlier, along with the UBY-API, was indispensable for our alignment research. Both the similarity-based and the graph-based frameworks we developed (and which were later combined) are based on UBY. In other words, the creation of UBY was a direct prerequisite for the progress we made in WSA, and thus, in a sense, UBY fueled its own development and growth.

While the WSA research would, in principle, be possible without a unified representation (as is shown by the excellent previous work in this field), the unified representation of LSRs offers the following advantages:

- Lexemes contained in each resource (i.e., `LexicalEntry` objects) are available and comparable across resources, which makes the process of candidate extraction for a given sense easy: we just have to query for all lexemes with the same lemma-POS combination and return their associated senses. The very same mechanism is also used for determining if a lexeme is monosemous; this is a requirement for the monosemous linking (Section 5.3.1) as well as the calculation of trivial alignments (Section 5.3.2). If the returned set for a specific lexeme and resource contains only one object, monosemy is guaranteed.
- Once the candidate senses are available, the sense descriptions (glosses) are directly accessible as well, making it easy to pass them on to the similarity calculation component which is based on the DKPro-Similarity framework (Bär et al., 2013). The resulting similarity values can be passed on to WEKA which is responsible for the machine learning part of the framework. DKPro-Core (Gurevych et al., 2007) is used for lemmatization and POS tagging of the glosses, which is also required for the monosemous linking. The candidates, along with their glosses, are also directly ready for usage in a human annotation task, i.e. to create gold standards.
- The core of Dijkstra-WSA are, of course, the nodes and edges. Here, the UBY-API allows easy construction of the graphs by iterating over all `Sense` or `Synset` objects of a given resource to use them as nodes, and at the same time extract the `SenseRelations` and `SynsetRelations` which serve as edges, in addition to the already mentioned monosemous links. UBY also directly provides the uniform and unique identifiers (also across resource boundaries) which are necessary for such a graph.
- To complete the “life cycle”, the output of the framework are, of course, the alignments, which are represented as pairs of IDs, or more specifically, UBY `Sense` or `Synset` identifiers. These can be used for instantly creating `SenseAxis` instances which are fed back into the UBY database.

In summary, our framework makes the creation of alignments between arbitrary resources as easy as possible, as long as they have been converted to the UBY format. The user is released from the burden of handling different APIs, file formats, identifiers, etc. Moreover, the approaches we describe here only require knowledge from the LSRs themselves, i.e. we refrain from using corpora, other resources etc. The only “external” knowledge is needed for basic preprocessing and similarity calculation tasks. As such, the framework can be considered self-contained, and is thus flexibly applicable to a variety of resources and languages as demonstrated by our experiments.

Following our general paradigm of making the resources and tools for creating them freely available (see Section 7.5), the framework was published under a non-restrictive license along with the UBY-API and converters.<sup>6</sup>

### A.3 Chapter Summary and Contributions

In this appendix, we present our contributions to WSA from a software development perspective. This includes Java-based APIs for OmegaWiki and UBY, as well as a web interface which further facilitates access to UBY by allowing exploratory examination of the resource without any initial effort.

In summary, our contributions in this chapter are:

**Contribution 1** We discuss JOWKL, a Java-based API which for the first time allows easy programmatic access to OmegaWiki.

**Contribution 2** We present the Java-based API which was developed for accessing UBY and which allows easy access to the standardized resources, as well as to the alignments between them.

**Contribution 3** We discuss the web interface to UBY. The interface combines a novel, intuitively understandable graph view for sense clusters with a textual browser that allows exploring the offered information in greater detail, with the option of comparing senses from different resources.

**Contribution 4** We discuss the implementation of our WSA framework which was built on the UBY-API and which was used for the experiments presented in the course of this thesis.

---

<sup>6</sup><http://code.google.com/p/uby/>