Using *lemonUby* for data integration on the Multilingual Semantic Web

Judith Eckle-Kohler, Iryna Gurevych, John McCrae and Christian Chiarcos

Abstract This chapter addresses lexicon-based methods for data integration on the Multilingual Semantic Web. It presents use cases of a new lexical resource integrated in the Semantic Web called *lemonUby* which combines data from the standardized lexical resource UBY with the principled model, *lemon*, for representing lexical data on the Web. This resource provides not only rich information about many lexical entries in two languages, it is also significantly linked both within its component resources and to other lexical resources and terminology repositories on the Web. This chapter describes further linkings of *lemonUby*, such as the linking with the MASC corpus. A particular focus are methods for cross-lingual linking of verb senses on the fly. We present results for the cross-lingual linking of the English VerbNet and two German lexica rich in verbs.

1 Motivation

Lexical resources in the multilingual Linked Open Data cloud play an important role for the development of the Multilingual Semantic Web, because they can be used for cross-lingual linking and localization of Web content. While many lexical

Judith Eckle-Kohler
Ubiquitous Knowledge Processing Lab (UKP-TUDA), Technische Universität Darmstadt, Germany, www.ukp.tu-darmstadt.de

Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA), Technische Universität Darmstadt and Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research and Educational Information, Germany, www.ukp.tu-darmstadt.de,

John McCrae
CITEC, Universität Bielefeld, Germany, jmccrae@cit-ec.uni-bielefeld.de

Christian Chiarcos
Information Sciences Institute, University of Southern California, USA, chiarcos@isi.edu

resources, especially terminological resources, are already available on the Web, syntactic descriptions, in particular of verbs, are largely missing so far. In the context of lexicalizing relational knowledge (such as, e.g., $like(Experiencer;Theme)$) in multiple languages, resources providing fine-grained syntactic information on verbs are required on a large scale and for many languages. Verbs are often used to express relations, e.g. the relation $like(Experiencer,Theme)$ can be lexicalized syntactically as $NP\ likes\ NP$.

2 Approaches to data integration

Recently, the language resource community has begun to explore the opportunities offered by the Semantic Web, lead by the formation of the Linguistic Linked Open Data (LLOD) cloud and an increasing interest in making use of Linked Open Data principles in the context of Natural Language Processing (NLP) and Linguistics [Chiarcos et al(2012)Chiarcos, Nordhoff, and Hellmann]. The use of RDF supports data integration and offers a large body of tools for accessing this data. Furthermore, the linked data approach gives rise to novel research questions in the context of language resources and their application.

For lexical resources, data integration has been in the focus of interest for many years, resulting in numerous mappings and linkings of lexica, as well as standards for representing lexical resources, such as the ISO 24613:2008 Lexical Markup Framework (LMF) [Francopoulo et al(2006)Francopoulo, Bel, George, Calzolari, Monachini, Pet, and Soria]. In this context, the LLOD cloud can be considered as a new data integration platform, enabling linkings not only between lexical resources, but also between lexical resources and other language resources, such as terminology resources and corpora.

lemon, a lexicon model for representing and sharing ontology lexica, has been proposed as a common interchange format for lexical resources on the Semantic Web[McCrae et al(2012)McCrae, Aguado-de Cea, Buitelaar, Cimiano, Declerck, Gómez-Pérez, Gracia, Hollink, Montie
Making use of a common interchange format is important, to integrate resources such as FrameNet and WordNet, which have been characterised as complementary resources [Baker and Fellbaum(2009)]. The RDF version of FrameNet currently available does not adhere to an interchange format such as *lemon*, but is specific to the underlying data model of FrameNet.

Independently from linked data principles and Semantic Web technology, the large-scale lexical-semantic resource UBY [Gurevych et al(2012)Gurevych, Eckle-Kohler, Hartmann, Matuschek, Meyer
has been developed.[1] UBY is based on LMF and has currently integrated 10 lexical resources in English and German. A subset of these resources is interlinked at the word sense level.

Recently, a selection of UBY lexica have been converted to the *lemon* format, resulting in the large resource *lemonUby*. This resource contains interoperable and interlinked versions of WordNet [Fellbaum(1998)], FrameNet [Baker et al(1998)Baker, Fillmore, and Lowe],

[1] `http://www.ukp.tu-darmstadt.de/uby/`

VerbNet [Kipper et al(2008)Kipper, Korhonen, Ryant, and Palmer], English and German Wiktionary[2], and the English and German entries of OmegaWiki.[3] *lemonUby* has been linked to other lexical resources (e.g. the WordNet versions 2.0 and 3.0 in the LLOD) and to terminology resources in the LLOD cloud. The linking to terminology resources comprises a linking of linguistic terminology used in *lemonUby* to ISOCat[4], the implementation of the ISO 12620:2009 Data Category Registry, as well as a linking to the Ontologies of Linguistic Annotation (OLiA,[Chiarcos(2008), Chiarcos(2012)]).

3 *lemon* and UBY-LMF

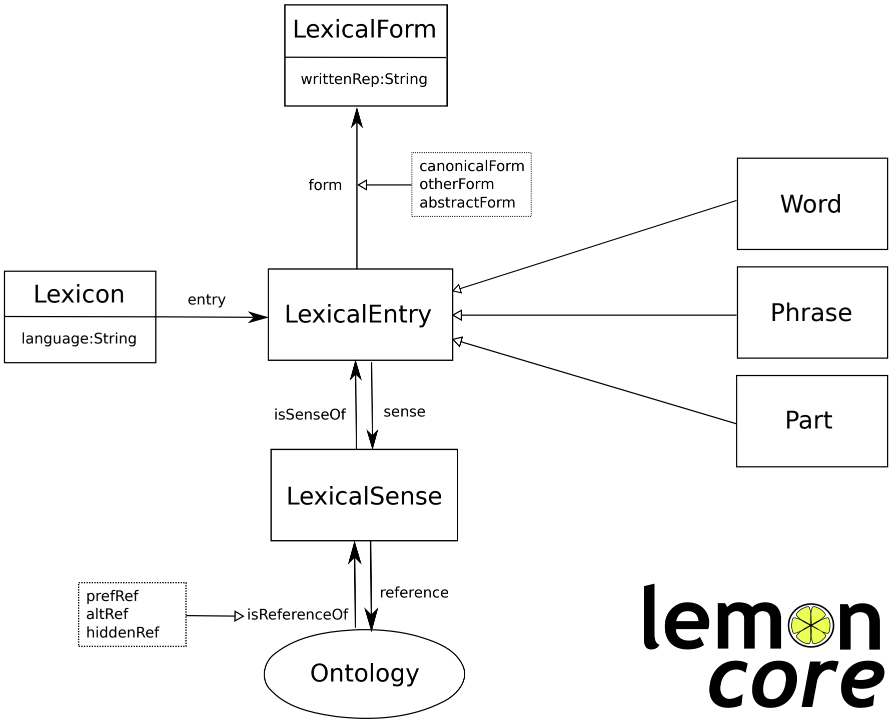

Fig. 1 The core of the *lemon* model

[2] http://www.wiktionary.org

[3] http://www.omegawiki.org

[4] http://www.isocat.org/

The *lemon* model[McCrae et al(2012)McCrae, Aguado-de Cea, Buitelaar, Cimiano, Declerck, Gómez-Pérez, Gracia, is a lexicon model for representing and sharing ontology lexica. It supports publishing lexical-semantic resources as linked data on the basis of the following principles:

LMF-based: To allow easy conversion from non-linked data resources.
RDF-native: Publishing as linked data, with RDFS and OWL used to describe the semantics of the model.
Modular: Separation of lexicon and ontology layers, so that *lemon* lexica can be linked to existing ontologies in the linked data cloud.
Externally defined data categories: Linking to data categories in annotation terminology repositories, rather than being limited to a specific part-of-speech tag set.
Principle of least power: The smaller the model and the less expressive the language, the wider its adoption and the higher the reusability of the data[Shadbolt et al(2006)Shadbolt, Hall, and Berners

This *lemon* core model is illustrated in Fig. 1, which defines the basic elements used by all lexica published as linked data. In addition to this, there are a number of modules used to model linguistic description, syntax, morphology and relationships between lexica.[5]

In UBY, interoperability is achieved by standardizing lexical resources according to UBY-LMF [Eckle-Kohler et al(2012)Eckle-Kohler, Gurevych, Hartmann, Matuschek, and Meyer, Eckle-Kohler et al(2013)Eckle-Kohler, Gurevych, Hartmann, Matuschek, and Meyer], a lexicon model which is a full-fledged instantiation of the ISO standard LMF, specifically for NLP.

In comparison to the *lemon* lexicon model, UBY-LMF is similar in two aspects: first it is LMF-based, and second, it uses externally defined data categories from ISOCat.[6] In contrast to *lemon*, however, UBY-LMF is based on two quite different principles:

Principle of Adoption: UBY-LMF has been designed to fully cover a wide range of heterogeneous lexical resources without information loss.
Independence of implementation: UBY-LMF is independent of any particular implementation. There are many ways to implement an LMF lexicon model [Francopoulo et al(2007)Francopoulo, Bel, C including RDF.

We performed a mapping of UBY-LMF to *lemon* which allows conversion of lexical resources in UBY-LMF format to *lemon* format.

Although both UBY-LMF and *lemon* are based on LMF, the mapping revealed substantial differences. These are mainly due to the fact that *lemon* is a model for ontology lexica where the lexicon and ontology layers are kept separate. Thus, sense representations in *lemon* primarily consist of references to the associated ontology where a rich and domain-specific sense definition is provided. The development of UBY-LMF, on the other hand, has been driven by the requirement to cover a

[5] More details on the model and descriptions of the modules can be found at `http://lemon-model.net`

[6] `http://www.isocat.org/rest/dcs/484`

large variety of lexical information types, which ranges from morphology and lexical syntax to lexical semantics and the mapping between syntactic and semantic arguments. Thus, the resulting lexicon model makes use of very fine-grained sense specifications which are often grounded in linguistic theories.

4 Linking *lemonUby* with corpora

The OLiA ontologies (and the terminology repositories it is linked with) can be used to represent, compare and integrate linguistic annotations in corpora on the basis of formal concepts rather than arbitrary strings [Chiarcos(2010)]. In a Linked Data context, they explicitly allow to compare the linguistic categories used in *lemonUby* with the morphosyntactic annotations in linguistic corpora, if these are represented in RDF. An RDF version of the MASC corpus [Ide et al(2008)Ide, Baker, Fellbaum, Fillmore, and Passonneau] has been produced, a resource that also provides FrameNet and WordNet annotations, and whose annotations can thus be directly compared with and combined with UBY resources. We performed a linking of the FrameNet 1.5 version contained in *lemonUby* with the FrameNet annotations in the MASC corpus. As MASC provides FrameNet sense annotations for different text genres and domains, this linking can be used to enrich FrameNet senses in *lemonUby* by genre and domain information.

5 Cross-lingual linking of verb senses

The standardized format for English and German subcategorization frames defined in UBY-LMF can be exploited for the linking of verb senses across English and German. We show in detail how such a linking based on subcategorization frame information is performed on the fly and point out important properties of the underlying representation of subcategorization frames related to this linking.

As an example, the cross-lingual linking of VerbNet and two German lexicons, i.e., the German wordnet GermaNet [Kunze and Lemnitzer(2002)] and the large syntactic subcategorization lexicon IMSLex [Eckle-Kohler(1999)] is described and evaluated. This is a particularly interesting linking, because it demonstrates how large lexica with a focus on particular information types in one language (e.g., syntactic subcategorization frames in IMSLex) can be enriched by complementary information from lexica in other languages (e.g., VerbNet). Both GermaNet and IMSLex contain detailed subcategorization information for verbs, but they do not contain information on semantic roles or selectional preferences for the arguments of verbs. Yet, they are the only large German lexica[7] containing fine-grained subcategorization information which are freely available for research purposes. The cross-lingual linking of these German lexicons and VerbNet provides access not only to

[7] GermaNet provides information for 8626 verb lemmas, IMSLex for 10879 verb lemmas.

information on semantic roles or selectional preferences, but also to semantic role information from FrameNet via the VerbNet–FrameNet linking in UBY.

Finally, we compare the cross-lingual linking of verb senses using original UBY lexicons (based on UBY-LMF) with the linking of verb senses which is based on the mapping of UBY-LMF to *lemon*.

6 Conclusion

This chapter describes methods for multilingual data integration on the Semantic Web which rely crucially on semantically interoperable language resources, i.e., language resources which are linked to reference terminology repositories, such as OLiA or ISOCat. The use cases of *lemonUby* described in this chapter give a good indication of the potential benefits of linking language resources on the Web on a large scale.

Acknowledgements The work on *lemon* was developed in the context of the Monnet project, which is funded by the European Union FP7 program under grant number 248458, the CITEC excellence initiative funded by the DFG (Deutsche Forschungsgemeinschaft). The work on UBY has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. The research of Christian Chiarcos has been supported by the German Academic Exchange Service (DAAD) through a PostDoc fellowship at the University of Southern California. We thank Silvana Hartmann, Michael Matuschek and Christian M. Meyer for their contributions to this work.

References

[Baker and Fellbaum(2009)] Baker CF, Fellbaum C (2009) WordNet and FrameNet as complementary resources for annotation. In: Proceedings of the Third Linguistic Annotation Workshop, Suntec, Singapore, ACL-IJCNLP '09, pp 125–129

[Baker et al(1998)Baker, Fillmore, and Lowe] Baker CF, Fillmore CJ, Lowe JB (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'98), Montreal, Canada, pp 86–90

[Chiarcos(2008)] Chiarcos C (2008) An ontology of linguistic annotations. LDV Forum 23(1):1–16

[Chiarcos(2010)] Chiarcos C (2010) Towards robust multi-tool tagging. an owl/dl-based approach. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp 659–670

[Chiarcos(2012)] Chiarcos C (2012) Ontologies of linguistic annotation: Survey and perspectives. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp 303–310

[Chiarcos et al(2012)Chiarcos, Nordhoff, and Hellmann] Chiarcos C, Nordhoff S, Hellmann S (eds) (2012) Linked Data in Linguistics. Representing Language Data and Metadata. Springer, Heidelberg

[Eckle-Kohler(1999)] Eckle-Kohler J (1999) Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora. Logos-Verlag, Berlin, Germany, phDThesis

[Eckle-Kohler et al(2012)Eckle-Kohler, Gurevych, Hartmann, Matuschek, and Meyer] Eckle-Kohler J, Gurevych I, Hartmann S, Matuschek M, Meyer CM (2012) UBY-LMF — A Uniform Format for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp 275–282

[Eckle-Kohler et al(2013)Eckle-Kohler, Gurevych, Hartmann, Matuschek, and Meyer] Eckle-Kohler J, Gurevych I, Hartmann S, Matuschek M, Meyer CM (2013) UBY-LMF - exploring the boundaries of language-independent lexicon models. In: Francopoulo G (ed) LMF: Lexical Markup Framework, theory and practice, ISTE - Wiley, London, UK, p (to appear)

[Fellbaum(1998)] Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA

[Francopoulo et al(2006)Francopoulo, Bel, George, Calzolari, Monachini, Pet, and Soria] Francopoulo G, Bel N, George M, Calzolari N, Monachini M, Pet M, Soria C (2006) Lexical Markup Framework (LMF). In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, pp 233–236

[Francopoulo et al(2007)Francopoulo, Bel, George, Calzolari, Monachini, Pet, and Soria] Francopoulo G, Bel N, George M, Calzolari N, Monachini M, Pet M, Soria C (2007) Lexical markup framework: ISO standard for semantic information in NLP lexicons. In: Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV

[Gurevych et al(2012)Gurevych, Eckle-Kohler, Hartmann, Matuschek, Meyer, and Wirth] Gurevych I, Eckle-Kohler J, Hartmann S, Matuschek M, Meyer CM, Wirth C (2012) UBY - A Large-Scale Unified Lexical-Semantic Resource. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France, pp 580–590

[Ide et al(2008)Ide, Baker, Fellbaum, Fillmore, and Passonneau] Ide N, Baker C, Fellbaum C, Fillmore C, Passonneau R (2008) MASC: The Manually Annotated Sub-Corpus of American English. In: Proc. 6th Language Resources and Evaluation Conference (LREC 2008), Marrakesh, Morocco

[Kipper et al(2008)Kipper, Korhonen, Ryant, and Palmer] Kipper K, Korhonen A, Ryant N, Palmer M (2008) A Large-scale Classification of English Verbs. Language Resources and Evaluation 42:21–40

[Kunze and Lemnitzer(2002)] Kunze C, Lemnitzer L (2002) GermaNet — representation, visualization, application. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Canary Islands, Spain, pp 1485–1491

[McCrae et al(2012)McCrae, Aguado-de Cea, Buitelaar, Cimiano, Declerck, Gómez-Pérez, Gracia, Hollink, Montiel-Ponsoda, Spohr, and W McCrae J, Aguado-de Cea G, Buitelaar P, Cimiano P, Declerck T, Gómez-Pérez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D, Wunner T (2012) Interchanging lexical resources on the semantic web. Language Resources and Evaluation 46:701–719, DOI 10.1007/s10579-012-9182-3

[Shadbolt et al(2006)Shadbolt, Hall, and Berners-Lee] Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. IEEE intelligent systems 21(3):96–101