Inducing the Cross-Disciplinary Usage of Morphological Language Data Through Semantic Modelling

DISSERTATION

zur Erlangung der Würde einer Doktorin der Philosophie

vorgelegt der Philosophisch-Historischen Fakultät der Universität Basel

von

Bettina Klimek, M.A.

aus

Gera

Leipzig 2020
Hirsch Printmedien

Genehmigt von der Philosophisch-Historischen Fakultät der Universität Basel, auf Antrag von Prof. Dr. Gerhard Lauer (Digital Humanities Lab), Prof. Dr. Lukas Rosenthaler (Digital Humanities Lab) und in Zusammenarbeit mit Dr.-Ing. Sebastian Hellmann (Institut für Angewandte Informatik, Leipzig) und Dr. rer. nat. Marco Büchler (Institut für Angewandte Informatik, Leipzig).

Basel, den 2. Dezember 2020
Der Dekan Prof. Dr. Ralph Ubl

Dies ist eine kumulative Dissertation und beinhaltet die folgenden Einzelbeiträge:
(This is a cumulative dissertation comprising the following academic articles:)

Bettina Klimek, Markus Ackermann, Amit Kirschenbaum, and Sebastian Hellmann, 2017. "Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge." In Rehm, G. and Declerck, T. (Eds.): *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*. Springer International Publishing, pp. 130-145.

Bettina Klimek, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann, 2020. "MMoOn Core – The Multilingual Morpheme Ontology." In Hitzler, P. and Janowicz, K. (Eds.): *Semantic Web*. IOS Pre-Press, pp. 1-30.

Bettina Klimek, 2017. "Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models." In McCrae, J. P. et al. (Eds.): *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*. CEUR Workshop Proceedings 1899, pp. 68-83.

Bettina Klimek, John P. McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber, and Christian Chiarcos, 2019. "Challenges for the Representation of Morphology in Ontology Lexicons." In Kosem, I. et al. (Eds.): *Electronic Lexicography in the 21st Century (eLex 2019): Smart Lexicography*. Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 570-591.

Bettina Klimek, Natanael Arndt, Sebastian Krause, and Timotheus Arndt, 2016. "Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory." In Calzolari, N. et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, pp. 892-899.

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, and Uwe Quasthoff, 2018. "Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment." In Calzolari, N. et al. (Eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. ELRA, pp. 4372-4378.

To my daughter Luana Maya,
the best I will have ever created.

Acknowledgements

This thesis is the final result of a long scientific journey which could not have been achieved without the indispensable presence of numerous individuals. I, therefore, want to pay tribute to Prof. Dr. Gerhard Lauer and Prof. Dr. Lukas Rosenthaler for the supervision of my thesis in its final phase. My great appreciation is expressed for Prof. Dr. Gerhard Heyer who granted me guidance and supervision at the University of Leipzig. I would also like to thank Dr. Sebastian Hellmann for initiating and always supporting my doctorate. Because of him, I was fortunate to be part of the KILT and AKSW research groups of the Institute for Applied Informatics in Leipzig which I like to thank for their collaboration, inspiration, feedback and ongoing technical assistance. I would like to especially thank my former colleagues Markus Ackermann and Martin Brümmer for guiding me patiently through the depth of Linked Data and also for their openness to let me introduce them to the field of morphology in return.

Moreover, I like to give thanks to all the co-authors who worked together with me on the publications that emerged during the pursue of my studies. Many of them are members of the Linguistic Linked Open Data research community which constitutes the unification of diverse language-data-oriented researchers that inspired me to conduct this interdisciplinary work. I am grateful for all the outstanding scientists I met within this group on multiple occasions who offered their critical feedback, constant assurance on the necessity of my work, motivation as well as their trust in me.

I would like to thank Dr. Christian Chiarcos, Dr. John P. McCrae and Dr. Jorge Gracia for leading by example and having contributed to my growth as a scientist by sharing their expertise, experiences and valuable advice during various collaborations. My special thanks goes to my fellow PhD students Julia Bosque-Gil, Maxim Ionov and Christian Fäth for their constant encouragement, assistance and much appreciated companionship during this shared experience of becoming a resilient cross-disciplinary researcher. Further, I would like to thank Dr. Monika Rind-Pawlowski for never getting tired to firmly remind me of my roots as a linguist whenever I was in danger to lose the balance in my interdisciplinary research endeavours.

In addition to these people who directly contributed to achieving my research outcomes, there are many others who promoted me in no less sig-

nificant ways. Therefore, I would like to thank my parents, my brother and sister, Christian and Heike, and the rest of my family for always believing in me and taking care of my daughter so many times which allowed me to travel to the various conferences and events in the course of my work. Finally, my sincere gratitude goes to my longstanding friends Sandra Prätor, Dr. Tina Schmeißner and Kathleen Grimm who accompanied me during this thesis with their unceasing encouragement and unconditional availability whenever highly needed. I could not have wished for better travel mates on this journey.

The research activities included in this dissertation were partially funded by grants from the EU projects LIDER (GA-610782) and ALIGNED (GA-644055), the Smart Data Web BMWi project (GA-01MD15010B) as well as the PLASS project (01MD19003D).

Summary

Despite the enormous technological advancements in the area of data creation and management the vast majority of language data still exists as digital single-use artefacts that are inaccessible for further research efforts. At the same time the advent of digitisation in science increased the possibilities for knowledge acquisition through the computational application of linguistic information for various disciplines.

The purpose of this thesis, therefore, is to create the preconditions that enable the cross-disciplinary usage of morphological language data as a sub-area of linguistic data in order to induce a shared reusability for every research area that relies on such data. This involves the provision of morphological data on the Web under an open license and needs to take the prevalent diversity of data compilation into account. Various representation standards emerged across single disciplines which lead to heterogeneous data that differs with regard to complexity, scope and data formats. This situation requires a unifying foundation enabling direct reusability.

As a solution to fill the gap of missing open data and to overcome the presence of isolated datasets a semantic data modelling approach is applied. Being rooted in the Linked Open Data (LOD) paradigm it pursues the creation of data as uniquely identifiable resources that are realised as URIs, accessible on the Web, available under an open license, interlinked with other resources, and adhere to Linked Data representation standards such as the RDF format. Each resource then contributes to the LOD cloud in which they are all interconnected. This unification results from ontologically shared bases that formally define the classification of resources and their relation to other resources in a semantically interoperable manner. Subsequently, the possibility of creating semantically structured data has sparked the formation of the Linguistic Linked Open Data (LLOD) research community and LOD sub-cloud containing primarily language resources. Over the last decade, ontologies emerged mainly for the domain of lexical language data which lead to a significant increase in Linked Data-based linguistic datasets. However, an equivalent model for morphological data is still missing, leading to a lack of this type of language data within the LLOD cloud.

This thesis presents six publications that are concerned with the peculiarities of morphological data and the exploration of their semantic representation as an enabler of cross-disciplinary reuse. The Multilingual

Morpheme Ontology (MMoOn Core) as well as an architectural framework for morphemic dataset creation as RDF resources are proposed as the first comprehensive domain representation model adhering to the LOD paradigm. It will be shown that MMoOn Core permits the joint representation of heterogeneous data sources such as interlinear glossed texts, inflection tables, the outputs of morphological analysers, lists of morphemic glosses or word-formation rules which are all equally labelled as "morphological data" across different research areas. Evidence for the applicability and adequacy of the semantic modelling entailed by the MMoOn Core ontology is provided by two datasets that were transformed from tabular data into RDF: the Hebrew Morpheme Inventory and Xhosa RDF dataset. Both further demonstrate how their integration into the LLOD cloud – by interlinking them with external language resources – yields insights that could not be obtained from the initial source data.

Altogether the research conducted in this thesis establishes the foundation for an interoperable data exchange and the enrichment of morphological language data. It strives to achieve the broader goal of advancing language data-driven research by overcoming data barriers and discipline boundaries.

Contents

Chapter 1

Introduction

1.1 Motivation and Background

The rise of the Digital Age introduced an ongoing transformation of knowledge acquisition that inevitably affected all scientific disciplines. Due to the unprecedented availability and amount of digitised data, today's research landscape is progressively active in interrelating the data and results of formerly unrelated disciplines which, thus, led to the emergence of new research areas. This development especially influenced the sciences and research fields that produce or rely on language data.

Evidence for this can be drawn from the evolution of the field of lexicography (Nielsen, 2017). Lexical data as a type of language data was originally compiled by linguists long before the invention of computers, i.e. the first monolingual dictionaries dating to the late 16th century designed for educational purposes (cf. Osselton, 1990, p.1944) and Sumerian lists dating back to the third millennium BC (cf. Boisson et al., 1991, p. 263). However, by taking up the technological possibilities, print dictionaries advanced to lexical databases and electronic word nets. These were not only affecting the research direction of (e-)lexicography itself but also gave rise to increasingly accurate language processing systems provided by the field of computational linguistics. As a result, methods and tools have been developed for tasks such as automated word sense disambiguation, named entity recognition and machine translation (Bird et al., 2009). The usage of these in combination with knowledge bases then enabled content mining, e.g. as in Weichselbraun et al. (2014). In return, the tools and corpora created by the computational linguists and within the area of content mining were equally useful to lexicographers and linguists. As a consequence, new language resources could be created and means to automatically extract linguistic information out of these were developed. New data foundations became available that provided lexical content which was out of reach within the methodologies of lexicography before. Suddenly a large amount of attested new words, senses and usages can be linguistically analysed and investigated. Overall, the adaption

of lexical data, eventually, demonstrates how the impact of digitisation contributes to increased scientific outcomes within single fields of research through the cross-disciplinary usage of language data.

Accordingly, within this thesis **cross-disciplinary usage** is defined as "knowledge acquisition gain in one discipline that is achieved by the reuse of language data that was originally produced within another discipline". From the wide range of disciplines to which this definition applies the main focus will be on the following three research areas being concerned with language data. The first one deals with the compilation and analysis of language data with the purpose of studying natural languages as an epistemic object itself. This scientific area is represented by the branch of traditional linguistics that understands linguistics as an empirical science using and producing language and linguistic data in order to derive and also to verify theories about language. With the technological progress the second research area emerged which focuses on enabling machines to process and generate large amounts of natural language data which exceed the manual capabilities of traditional linguists, i.e. the field of computational linguistics. Content mining constitutes the third application area. It differs, however, from the other two in that it is interested in natural language data as an information source for knowledge extraction.

With regard to language data the scope of this thesis encompasses **morphological language data** in particular as one linguistic data type. As such this data entails the smallest meaning-bearing elements of language and the internal structure of words, i.e. it represents linguistic data on the word and sub-word levels. Morphology is generally not acknowledged as an individual data domain but regarded as a field that is located between lexicon and grammar. Therefore, the granularity and amount of morphological data provided within lexical datasets varies depending on the underlying lexicographic theory. These diverge widely between a minimalist and maximalist view delimiting which morphological components are included (cf. Booij et al., 2000, p. 348). Thus, treating morphological data as an independent language data domain provides the potential to obtain more language data which is hitherto not covered by lexical datasets and, therefore, also contributes to an increase of the data basis for the above mentioned disciplines to create a knowledge acquisition gain. Morphological data poses the possibility to reach a more comprehensive representation of a language in as far as it compensates for the limitations of the lexical data domain. A significant constraint, for instance, involves the aspect of coverage. Traditionally, for practical reasons, dictionaries and lexical data are not meant to be exhaustive (cf. Atkins & Rundell, 2008, p.20). A variety of criteria determine which word or expression is included and to what extent it is described as a lexical entry or dictionary headword[1].

[1]The aspect of exhaustiveness is less strongly pursued since the space limitations of print dictionaries vanished together with growing computational space. However, derivational and compound processes allow for a creation of an infinite number of new

Proper nouns, for instance, designating people, locations or organisations are, by definition, excluded or enter the dictionary only if they are very frequently used. Together with the number of new words that can be created on the basis of proper nouns, e.g. *Darwinism*[2] from *Charles Darwin*, a large amount of language data denoting named entities is not identifiable with lexical data. Also, inflectional data is mostly present for certain selected grammatical forms, e.g. the plural word-form of nouns. Beyond that, inflectional language data is devoted to the grammar and therein mostly represented in exemplary ways, leaving the majority of word-forms undocumented. The reason for this is that the coverage of linguistic information about lexical entries is highly concerned with the definition of lexical meanings. Therefore, mainly very productive derivational affixes like the English prefix *un-* as in *unreal* are more commonly provided in lexical datasets. However, not all lexemes are documented that can be formed with it.

The points just outlined eventually amount to the crucial characteristic of productivity of natural languages. Due to the digitisation a plethora of language data became available that unveils this aspect of the infinite recreation and formation of words in an unprecedented manner. Nevertheless, humans do not need a constantly growing dictionary because they can intuitively assess the meanings of new expressions by instantly applying a lot of interconnected linguistic information - much of it extending to the field of morphology - like part of speech, grammatical categories, transformation rules, the selection of a specific meaning in a given context, the decomposition and analysis of inter-dependencies of the morphemes it is composed of or the identification of phonological adjustments or sub-word language elements of foreign origin. In contrast, machines require this extensive information explicitly in order to process natural language about as well as humans almost effortlessly do.

As a solution to that a mainly computational approach has been established by applying task-specific code, algorithms, systems, tools and computational frameworks whenever linguistic information is needed that exceeds the applicability of lexical data or the capacity of an expert annotation (Heyer et al., 2006). These procedures, however, highly reduce the cross-disciplinary usage of the resulting language data. In many cases a lot of effort is required to understand how the data was exactly created and post processing is needed to transform it into a format that other potential users of this data work with. Moreover, the linguistic quality is questionable if the data is the outcome of several processing steps that rely on mere computational methods rather than linguistic accuracy. As a result,

words, many of which are attested but not included into lexical datasets yet.

[2]This thesis follows the generic style rules for linguistic (Haspelmath, 2014). Italics are used for all object-language forms (such as words and morphs) that are cited within the text or examples and single quotation marks are used for indicating their corresponding linguistic meanings.

linguists tend to dissociate themselves from reusing this language data and prefer to work with specific software or tables which are in turn not usable for computational approaches without any adaptions. Consequently, linguistic and language data is continuously produced within various research fields but most of it stays in data silos. Once created for a specific purpose this data is not further reused by other disciplines even though it might also be of potential research interest. Therefore, cross-disciplinary usage is strongly correlated with the reusability of language data. As a consequence, a data-driven approach evolved that aims at homogenising the data resulting from the computational methods based on highly interoperable data formats. The core of this entails **semantic data modelling** which represents data in terms of a formally defined ontology that achieves a machine-processable meaningful interrelation between different datasets and enables automated inference and reasoning over all datasets sharing the same underlying ontology. With regard to language data in general and morphological data in particular this kind of data representation entails a high potential towards realising a cross-disciplinary usage because it enables the unification of data resources which is necessary to overcome the predominant creation of single-use data.

In fact, since 2011 a new research area called Linguistic Linked Open Data (LLOD; Chiarcos, Hellmann, et al., 2012; Chiarcos, Moran, et al., 2013; Chiarcos, Nordhoff, et al., 2012; McCrae et al., 2016) emerged which aims at complementing the computational approach to language resources by implementing a data-driven approach simultaneously. It is based on Semantic Web technologies and the Linked Data principles as the manifestations of semantic data modelling. Its main effort is grounded in the creation of models, i.e. ontologies, that enable a representation of language data in the Resource Description Framework (RDF)[3] format and the Web Ontology Language (OWL)[4] in order to exploit the main innovation of the Semantic Web which is "a web of things in the world, described by data on the Web" (cf. Bizer et al., 2009, p. 2). According to the Linked Data principles data should be published and connected on the Web following these rules (Berners-Lee, 2006):

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

In the context of scientific data reuse it is further necessary to provide this data under an open license. The LLOD research community took up

[3]https://www.w3.org/RDF/
[4]https://www.w3.org/OWL/

the Linked Data principles and applied them to linguistic data with the aim to enhance the publication of open language data in an interoperable manner. The main advantages of this approach are summarised as follows (quoted according to Chiarcos, McCrae, et al., 2013):

- *Representation and modelling:* RDF is based on labelled directed graphs and thus particularly well-suited for modelling [language resources].
- *Structural interoperability:* Using a common data model eases the integration of different resources. In particular, merging multiple RDF documents yields another valid RDF document, while this is not necessarily the case for other formats.
- *Conceptual interoperability:* The Linked Data principles have the potential to make the interoperability problem less severe in that globally unique identifiers for concepts or categories can be used to define the vocabulary and these URIs can be used by many parties who have the same interpretation of the concept. Furthermore, linking by OWL axioms allows to define the exact relation between two different concepts beyond simple equivalence statements.
- *Expressivity:* Semantic Web languages (OWL in particular) support the definition of axioms that allow to constrain the usage of the vocabulary, thus introducing formal data types and the possibility of checking [language] data for consistency.
- *Federation:* In contrast to traditional methods, where it may be difficult to query across even multiple parts of the same resource, Linked Data allows for federated querying across multiple, distributed databases maintained by different data providers.

As of today this endeavor progressed in so far that more than 200 datasets of linguistic and language data have been newly created in or transformed into RDF. All of them are interconnected with the data of at least one other dataset within the LLOD cloud[5] that constitutes the web of language data. The range of datasets that are available includes language data from various domains, i.e. corpora, lexicons and dictionaries, terminology, thesauri, knowledge bases, linguistic resource metadata, linguistic data categories and typological databases. Given that all these resources are produced and used by researchers from different scientific backgrounds it can be said that Linked Data functions as a significant driver for the cross-disciplinary usage of language data as for instance the creation of Babelfy[6] (Moro et al., 2014) and its resulting applications (Ekinci & İlhan Omurca, 2020; Ekinci & Omurca, 2018; Färber et al., 2018) demonstrate.

[5]https://linguistic-lod.org/llod-cloud
[6]http://babelfy.org/

Despite these advancements and advantages, the domain of morphological language data is highly underrepresented within the area of LLOD. Even though linguists document morphological data in field research and grammars, and a variety of tasks within computational linguistics and content mining require data about the meaning of sub-word units, comprehensive morphological models and datasets for various languages are still missing (Bosque-Gil et al., 2018). In particular this would be an inventory of the smallest meaningful elements of language and their semantic interrelations similar to a dictionary or lexical database for lexical data. Three main reasons for this can be identified which are mutually dependent:

1. Lack of consistent domain documentation: In contrast to the domain of lexical data out of which emerged the field of lexicography, the domain of morphology lacks a dedicated field that deals with a general documentation framework for the compilation and description of morphological data. The extent of morphological data for a language is usually distributed between the lexicon and the grammar of that language (cf. Booij et al., 2004, p. 1870). Both are, however, often created by different linguists which deviate with regard to the scope, granularity and theoretic foundation of the morphological data that is represented. In addition to this, morphological data is mainly contained in an exemplary manner leaving the majority of the data undocumented. Given that morphological language data is, however, regarded as the entirety of the information provided in lexicons as well as grammars – which are in turn also very language-specific – no consistent cross-linguistic domain documentation has evolved.

2. Heterogeneous single purpose data: Without at least a minimal set of scientifically shared and acknowledged representation standards for morphological language data, the existing data landscape is characterised by a large amount of datasets which highly diverge with regard to the quality, granularity as well as the underlying linguistic understanding of the domain of morphology. As a result, interlinear glossed texts, inflection tables, the outputs of morphological analysers as well as lists of morphemic glosses or word formation rules are all equally labelled as "morphological data". Moreover, most of this data has been produced for a specific research purpose and is not used beyond that. Even if a confluence and interconnection of one or more datasets would be envisaged, heterogeneous data formats would impede this endeavor. A lot of the data linguists produce is hidden in unstructured formats such as documents and, hence, not machine-processable. Conversely, computationally produced morphological data, e.g. by morphological analysers, is often not understandable for linguists. Against this background the creation of an appropriate data model that enables a more homogeneous representation of morphological data is rather difficult. Attempts to create such a model

resulted in the consideration of morphological data within the Lexicon Model for Ontologies (OntoLex-*lemon*)[7] (Klimek et al., 2019; McCrae et al., 2017). However, this model only solved the issue of data format interoperability but insufficiently represents morphological data, yet again, only as a part of lexical or grammatical data.

3. Technological limitations: Assuming that a consistent digitisation of morphological language data would be possible, still, issues regarding the technical implementation arise. Due to the necessary explication of a lot of, thus far, only indirectly existing information, morphological datasets would grow significantly in size. This, consequently, entails a need for data storage space as well as increased working memory power for its computational application to which not all data creators or their affiliated institutions have access to. In addition to that, technical infrastructure is required that enables the publication of morphological datasets along with the publication documents that refer to this data. Without that the majority of morphological data will remain inaccessible and vanish on the hard drives of the data creators.

All these reasons cause in parts or their entirety the discouragement of transforming existing morphological language data into LLOD. As a result, less datasets containing morphological information exist than the potential of the available non-Linked Data morphological datasets allows. From the present viewpoint, certainly, these restraints are no longer sustainable. Even though there is no widespread awareness among data creators regarding the technical possibilities provided by the Semantic Web, the technical obstacles outlined above in reason number three are largely solved by the Linked Data principles. Due to the integration of data into the Web itself in the form of Unified Resource Identifiers (URIs), data size is significantly reduced even for large datasets. High memory power is also not required to access the data because it can be browsed like any other information on the internet. Furthermore, the deployment of online services and platforms, such as DataHub[8] or LingHub[9], offers the hosting and publication space to distribute, share and discover the data. Together with the ongoing and quickly improving advancement of personal computers and Web development these Linked Data–specific infrastructures have, therefore, actually overcome the cited obstacles of the technical limitations.

The reasons number one and two in the context of LLOD can be regarded as a chance for the advancement of the linguistic domain of morphological data. In fact, the missing consistent documentation framework for morphological data entails the possibility, for the first time, to work towards

[7] https://www.w3.org/2016/05/ontolex/
[8] https://datahub.io/
[9] http://linghub.org/

a comprehensive domain representation independent of the prevalent constraints. Due to the initial efforts to provide an electronic recreation of the structural setup and implicitly contained semantics of the typography of print dictionaries (cf. Granger & Paquot, 2012, pp.1-2), the interdisciplinary applicability of these electronic counterparts was highly limited. Even though the awareness and usage of lexical data across various areas rapidly increased over the last two decades, problematic issues that are rooted in the transfer of these print dictionary-specific structures impact a multi-functional reuse of lexical databases (Tarp, 2012) and lexical Linked Data-datasets (Bosque-Gil et al., n.d.) which are now part of the broader range of information science and digital humanities. In contrast, the missing prescriptive foundation for morphological datasets enables, reversely, the development of a descriptive data domain representation which directly takes the cross-disciplinary application needs into account. It is due to the inevitable digital setting of language data existing today that the diversity of morphological data created and used in various research fields is uncovered. This provides the visibility that allows to inductively arrive at a discrete representation model that accounts for the scope, granularity and usage of morphological data in its cross-disciplinary occurrence. To eventually obtain comparable representation standards similar to the domain of lexicography the development of an ontology as the foundational data representation framework is suitable. Such an ontology for morphological data will yield interoperable datasets that can be flexibly extended, interconnected with language data of other domains and converted into other formats if required. To this extent it adheres to the reusability needs of cross-disciplinary data usage. In acknowledging the prospect as well as the feasibility just outlined, future Linked Data-based morphological datasets are capable of enabling access to the phenomenology and knowledge which is encoded within the smallest meaningful units of language.

Therefore, the aim of this thesis is to close the gap of missing morphological language data and to investigate its cross-disciplinary usage potential. Under the assumption that many of the aforementioned limitations of lexical data are solvable with comprehensive and interoperable morphological datasets, the underlying overall working hypothesis is that **semantically modelled and represented morphological data will enhance the cross-disciplinary usage of language data in general**. In order to scientifically verify this proposition the following three prerequisites need to be established:

1. Evidence must exist that morphological language data can improve the results of cross-disciplinary tasks which are hitherto performed by relying on other types of language data.

2. An adequate ontology that models the domain of morphological language data is available for the creation of semantically represented

and interoperable morphological data.

3. Morphological datasets based on this ontology have to provide cross-disciplinary usage to a significant degree in that the resulting application of these datasets is attributable to the underlying semantic data structure.

This thesis presents research that can be regarded as the realisation of these requirements and, thus, initiates the induction of cross-disciplinary morphological data usage.

1.2 Overview of Own Contributions

This thesis contains six individual contributions in the form of four conference papers, one workshop paper and one journal article. All publications have been peer-reviewed and successfully published. For five of these six contributions the author holds the main authorship. A detailed declaration of the author's contributions to these publications is given in Chapter 6. Each of the six works deals with a separate thematic area. Their interconnection arises out of the overall working hypothesis as defined in the previous chapter. The following six publications are part of the thesis:

[**P1**] **Bettina Klimek**, Markus Ackermann, Amit Kirschenbaum, and Sebastian Hellmann, 2017. "Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge." In Rehm, G. and Declerck, T. (Eds.): *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*. Springer International Publishing, pp. 130-145.

[**P2**] **Bettina Klimek**, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann, 2020. "MMoOn Core – The Multilingual Morpheme Ontology." In Hitzler, P. and Janowicz, K. (Eds.): *Semantic Web*. IOS Pre-Press, pp. 1-30.

[**P3**] **Bettina Klimek**, 2017. "Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models." In McCrae, J. P. et al. (Eds.): *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*. CEUR Workshop Proceedings 1899, pp. 68-83.

[**P4**] **Bettina Klimek**, John P. McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber, and Christian Chiarcos, 2019. "Challenges for the

Representation of Morphology in Ontology Lexicons." In Kosem, I. et al. (Eds.): *Electronic Lexicography in the 21st Century (eLex 2019): Smart Lexicography.* Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 570-591.

[**P5**] **Bettina Klimek**, Natanael Arndt, Sebastian Krause, and Timotheus Arndt, 2016. "Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory." In Calzolari, N. et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* ELRA, pp. 892-899.

[**P6**] Sonja Bosch, Thomas Eckart, **Bettina Klimek**, Dirk Goldhahn, and Uwe Quasthoff, 2018. "Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment." In Calzolari, N. et al. (Eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* ELRA, pp. 4372-4378.

In the course of the doctorate the following additional publications emerged. Some of them evolved around the research conducted in the six publications that are in the focus of this thesis and will be referred to in the synopsis.

CONFERENCE, PEER-REVIEWED

- McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., **Klimek, B.**, Moran, S. and Osenova, P., 2016. "The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud." In Calzolari, N. et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* ELRA, pp. 2435-2441.

- **Klimek, B.**, McCrae, J. P., Lehmann, C., Chiarcos, C. and Hellmann, S., 2017. "OnLiT: An Ontology for Linguistic Terminology." In Gracia, J. et al. (Eds.): *International Conference on Language, Data and Knowledge 2017.* Springer, Cham, pp. 42-57.

- **Klimek, B.**, Schädlich, R., Kröger, D., Knese, E. and Elßmann, B., 2018. "LiDo RDF: From a Relational Database to a Linked Data Graph of Linguistic Terms and Bibliographic Data." In Calzolari, N. et al. (Eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* ELRA, pp. 2429-2436.

- Eckart, T., **Klimek, B.**, Goldhahn, D. and Bosch, S., 2018. "Using Linked Data Techniques for Creating an IsiXhosa Lexical Resource - a Collaborative Approach." In Skadina, I. and Eskevich, M. (Eds.): *CLARIN Annual Conference 2018.* pp. 26-29.

- Eckart, T., Bosch, S., Goldhahn, D., Quasthoff, U. and **Klimek, B.**, 2019. "Translation-based Dictionary Alignment for Under-resourced Bantu Languages." In Eskevich, M. et al. (Eds.): *2nd Conference on Language, Data and Knowledge (LDK 2019).* Schloss Dagstuhl-Leibniz-Zentrum für Informatik, pp. 17:1–17:11.

PROCEEDINGS

- Eskevich, M., de Melo, G., Fäth, C., McCrae, J.P., Buitelaar, P., Chiarcos, C., **Klimek, B.** and Dojchinovski, M., (eds.) 2019. OASIcs, Volume 70, LDK'19, Complete Volume. In *2nd Conference on Language, Data and Knowledge (LDK 2019).* Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

WORKSHOP, PEER-REVIEWED

- Chiarcos, C., **Klimek, B.**, Fäth, C., Declerck, T. and McCrae, J. P., 2020. "On the Linguistic Linked Open Data Infrastructure." In Rehm, G. et al. (Eds.): In *Proceedings of the 1st International Workshop on Language Technology Platforms.* ELRA, pp. 8-15.

JOURNAL

- **Klimek, B.** and Brümmer, M., 2015. "Enhancing lexicography with semantic language databases." In *Kernerman Dictionary News*, 23. pp. 5-10.

This thesis consists of two main chapters, Chapter 2 and Chapter 3. While Chapter 2 corresponds to the three identified prerequisites that are required for the verification of the working hypothesis, in Chapter 3 it will be elaborated in how far these can be regarded as the initiation of the induction of cross-disciplinary morphological data usage. They are organised as follows.

In the subchapters of Chapter 2 the six contributions are reproduced in their original publication format according to the same order as just outlined. [P1] in Chapter 2.1 represents evidence that motivates the creation of semantically represented morphological data by investigating the effect of the morphological complexity of German on the system performances in a named entity recognition task. The publications [P2], [P3] and [P4] in the Chapters 2.2 to 2.4 are dedicated to the requested modelling for the domain of morphological language data. The MMoOn Core ontology is presented as a new foundation for morphological data representation as Linked Data in [P2]. An alignment of it to the OntoLex-*lemon* vocabulary is further discussed in [P3] and challenges specific to the modelling

of morphological language data are identified in [P4]. Datasets that have been created based on the MMoOn Core ontology are illustrated by the two publications [P5] and [P6] in the Chapters 2.5 and 2.6 respectively. Thereby, the application of the Open Hebrew Morpheme Inventory in [P5] and the Xhosa RDF dataset in [P6] serve as proof for the cross-disciplinary usage of semantically represented morphological language data.

Subsequently, Chapter 3 presents the synopsis of all publications. It will critically elaborate on the validity of the enhancement of the cross-disciplinary usage of language data in general within the realm of the conducted research included in Chapter 2. Therefore, the cross-disciplinary relevance of morphological language data and the semantic modelling approach are explained in Chapter 3.1 which is followed by a summary of the publication outcomes in Chapter 3.2. The resulting implications and pertaining limitations impacting further research are explicated in Chapter 3.3. Finally, the thesis ends with a conclusion in Chapter 4, an outlook on future work in Chapter 5 and the declaration of the author's contribution to the included publications in Chapter 6.

The conducted research of this thesis emerged from active participation in the LLOD research community. Its goal is to contribute to the creation of more openly available morphological language data in the RDF format in order to enhance language data-driven research in general by overcoming data barriers and discipline boundaries.

Chapter 2

Publications

2.1 Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge

This publication addresses the need for the creation of fine-grained morphological language data for morphologically rich languages. A cross-disciplinary applicability for such language data is exemplified for the natural language processing (NLP) task of named entity recognition (NER) by investigating how well systems perform on identifying morphologically complex German named entities based on the GermEval corpus data. A linguistic analysis explicating the complexity of German named entities and lexemes, which are created based on proper nouns, is provided. It motivates the development of a semantic, i.e. a Linked Data-based, modelling approach for morphological language data. Moreover, this work gives insight into the level of granularity that is required for representing morphological language data and can be regarded as valuable information that has to be taken into account for the development of an ontology for morphological data in general.

Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge

Bettina Klimek(✉), Markus Ackermann, Amit Kirschenbaum, and Sebastian Hellmann

AKSW/KILT Research Group, InfAI, University of Leipzig, Leipzig, Germany
{klimek,ackermann,amit,hellmann}@informatik.uni-leipzig.de

Abstract. This paper presents a detailed analysis of Named Entity Recognition (NER) in German, based on the performance of systems that participated in the GermEval 2014 shared task. It focuses on the role of morphology in named entities, an issue too often neglected in the NER task. We introduce a measure to characterize the morphological complexity of German named entities and apply it to the subset of named entities identified by all systems, and to the subset of named entities none of the systems recognized. We discover that morphologically complex named entities are more prevalent in the latter set than in the former, a finding which should be taken into account in future development of methods of that sort. In addition, we provide an analysis of issues found in the GermEval gold standard annotation, which affected also the performance measurements of the different systems.

1 Introduction

Despite initiatives to improve Named Entity Recognition (NER) for German such as in challenges as part of CoNLL 2003[1] and GermEval 2014[2], a noticeable gap still remains between the performance of NER systems for German and English. Pinpointing the cause of this gap seems to be an impossible task as the reasons are manifold and in addition difficult to realize due to their potentially granular (and subtle) nature as well as their inter-relatedness. However, we can name several aspects that might have an influence: (1) lack of linguistic resources suitable for German, (2) less demand (and interest) for improving the quality of NER systems for German, (3) variance of annotation guidelines and annotator consensus, (4) different NER problem definitions, (5) inherent differences between both language systems, (6) quality of provided data and source material, (7) etc. Studying the degree of impact for each of these factors

[1] CoNLL 2003 Challenge Language-Independent Named Entity Recognition, http://www.cnts.ua.ac.be/conll2003/ner/.

[2] GermEval 2014 Named Entity Recognition Shared Task, https://sites.google.com/site/germeval2014ner/, see also (Benikova et al. 2014a).

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 130–145, 2018.
https://doi.org/10.1007/978-3-319-73706-5_11

as a whole revokes any attempt to apply scientific methods for error analysis. However, a systematic investigation of linguistic aspects of proper nouns, i.e., named entities in technical terms[3], in German can reveal valuable insights on the difficulties and the improvement potential of German NER tools. Such an aspect is the morphological complexity of proper nouns. Due to its greater morphological productivity and variation, the German language is more difficult to analyze, offering additional challenges and opportunities for further research. The following list highlights a few examples:

- More frequent and extensive compounding requires correct token decompounding to identify the named entity (e.g., ***Bibel****forscherfrage* - 'bible researchers' question').
- Morphophonologically conditioned inner modifications are orthographically reflected and render mere substring matching ineffective (e.g., *außer****europä****isch*(*Europa*) - 'non-European').
- Increased difficulty in identifying named entities which occur within different word-classes after derivation (e.g., ***luther****ischen*, an adjective, derived from the proper noun *Martin Luther*).

These observations support the hypothesis that morphological alternations of proper nouns constitute another difficulty layer which needs to be addressed by German NER systems in order to reach better results. Therefore, this paper presents the results of a theoretic and manual annotation and evaluation of a subset of the GermEval 2014 Corpus challenge task dataset. This investigation focuses on the complexity degree of the morphological construction of named entities and shall serve as reference point that can help to estimate whether morphological complexity of named entities is an aspect which impacts NER and if it should be considered when creating or improving German NER tools. During the linguistic annotation of the named entity data, issues in the GermEval gold standard (in the following "reference annotation") became apparent and, hence, were also documented in parallel to the morphological annotation. Even though an analysis of the reference annotations was originally not intended, it is presented as well because it effects the measures of tool performance.

The rest of the paper is structured as follows. Section 2 presents an overview of related work in German NER morphology and annotation analysis. The corpus data basis and the scope of the analysis are described in Sect. 3. The main part constitutes Sect. 4, where in Sect. 4.1 the morphological complexity of German named entities is investigated and in Sect. 4.2 the distribution of morphologically complex named entities in the dataset is presented. Section 5 then explains and examines six different annotation issues that have been identified within the GermEval reference annotation. This part also discusses the outcomes. The paper concludes with a short summary and a prospect of future work in Sect. 6.

[3] From a linguistic perspective *named entities* are encoded as *proper nouns*. In this paper both terms are treated synonymously.

2 Related Work

The performance of systems for NER is most often assessed through standard metrics like precision and recall, which measure the overall accuracy of matching predicted tags to gold standard tags. NER systems for German are no exception in this respect. In some cases the influence of difference linguistic features is reported, e.g., part of speech (Reimers et al. 2014) or morphological features (Capsamun et al. 2014; Schüller 2014). The closest to our work, and the only one, to the best of our knowledge, which addresses linguistic error analysis of NER in German is that of Helmers (2013). The study examined different systems for NER, namely, TreeTagger (Schmid 1995), SemiNER (Chrupała and Klakow 2010), and the Stanford NER (Finkel and Manning 2009) trained on German data (Faruqui and Padó 2010). Helmers (2013) applied these systems to the German Web corpus CatTle.de.12 (Schäfer and Bildhauer 2012) and inspected the influence of different properties on NER in a random sample of 100 true positives and 100 false negatives. It reports the odd-ratios for false classification for each of the properties. It was found that, e.g., named entities written exclusively in lower case were up to 12.7 times more likely to be misidentified, which alludes the difficulty of identifying adjectives derived from named entities. Another relevant example was named entities labelled as "ambiguous", i.e., which have a non-named entity homonym as in the case of named entities derived from a common noun phrase. In this case three out of four NER systems were likely to not distinguish named entities from their appellative homonyms with an odd-ratio of up to 13.7. Derivational suffixes harmed the identification in one classifier but inflectional suffixes seemed not to have similar influence. In addition, abbreviations, special characters and terms in foreign languages were features which contributed to false positive results. In comparison with this study, ours addresses explicitly the effect of the rich German morphology on NER tasks.

Derczynski et al. (2015) raise the challenges of identifying named entities in microblog posts. In their error analysis the authors found that the errors were due to several factors: capitalization, which is not observed in tweets; typographic errors, which increase the rate of OOV to 2–2.5 times more compared to newsire text; compressed form of language, which leads to using uncommon or fragmented grammatical structures and non-standard abbreviations; lack of context, which hinders word disambiguation. In addition, characteristics of microblogs genre such as short messages, noisy and multilingul content and heavy social context, turn NER into a difficult task.

Benikova et al. (2015) describe a NER system for German, which uses the NoSta-D NE dataset (Benikova et al. 2014a) for training as in the GermEval challenge. The system employs CRF for this task using various features with the result that word similarity, case information, and character n-gram had the highest impact on the model performance. Though the high morphological productivity of German was stressed in the dataset description as well as in the companion paper for the conference (Benikova et al. 2014a), this method did not address it. What is more, it excluded partial and nested named entities which were, however, used in the GermEval challenge.

As this overview shows, linguistic error analysis is of great importance for the development of language technologies. Error analysis performed for NER tasks has been mostly concentrated on the token level, since this is the focus of most NER methods. However, our analysis differs in that it investigates specifically the role that morphology plays in forming named entities given that German is a language with rich morphology and complex word-formation processes.

3 Data Basis and Approach

3.1 GermEval 2014 NER Challenge Corpus

In order to pursue the given research questions we decided to take the Nosta-D NE dataset (Benikova et al. 2014b) included in the GermEval 2014 NER Challenge as the underlying data source of our investigations. The GermEval challenges were initiated to encourage closing the performance gap for NER in German compared to similar NER annotations for English texts. GermEval introduced a novelty compared to previous challenges, namely, additional (sub-) categories have been introduced indicating if the named entity mentioned in a token is embedded in compounding. Altogether, the named entity tokens could be annotated for the four categories *person*, *location*, *organisation* and *other* together with the information if the token is a compound word containing the named entity (e.g., LOCpart) or a word that is derived from a named entity (e.g., PERderiv). In addition it highlights a second level of 'inner' named entities (e.g., the person "Berklee" embedded in the organisation "Berklee College of Music"). Though the latter was addressed earlier, e.g., in Finkel and Manning (2009), it has been generally almost neglected. For detailed information about the GermEval NER Challenge, its setup, and the implemented systems we refer to Benikova et al. (2014a). Out of the eleven systems submitted to the challenge, only one considered morphological analyses (Schüller 2014) systematically. The best system, however, albeit utilizing some hand-crafted rules to improve common schemes of morphological alterations, did not model morphological variation systematically.

Besides a considerable volume of manual ground truth (31300 annotated sentences), the challenge data favourably was based upon well-documented, predefined guidelines[4]. This allowed us to create our complimentary annotations and to (re-)evaluate a subset of the original challenge ground truth along the same principles as proposed by the guidelines. Table 1 shows example sentences annotated for named entities (which can also be multi-word named entities

[4] The guidelines describing the categorization choice and classification of named entity tokens can be consulted in the following document: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5 (revision 1.6 effective for GermEval is referenced in https://sites.google.com/site/germeval2014ner/data).

consisting of more than one token) and their expected named entity types according to the provided GermEval reference annotation.

Table 1. Example of reference data from the GermEval provided annotated corpus.

Sentence	NE type
1951 bis 1953 wurde der nördliche Teil als Jugendburg des **Kolpingwerkes** gebaut	OTH
Beschreibung Die **Kanadalilie** erreicht eine Wuchshöhe von 60 bis 180 cm und wird bis zu 25 cm breit	LOCpart
Um 1800 wurde im ehemaligen **Hartung'schen** Amtshaus eine Färberei eingerichtet	PERderiv
1911 wurde er Mitglied der **sozialistischen Partei**, aus der er aber ein Jahr später wieder austrat	ORG

3.2 GermEval 2014 System Predictions

In order to obtain insights on the distribution of morphological characteristics of ground truth named entities which were successfully recognized by the systems (true positives) compared to ground truth named entities which were not recognized or categorized correctly[5] (false negatives), we requested the system prediction outputs of GermEval participants from the challenge organizers[6].

Based on the best predictions[7] submitted for each system, we computed (1) the subset of ground truth named entities that all systems recognized (i.e., the true positive intersection, TPi; 1008 named entities) and (2) analogously the subset of ground truth named entities that none of the systems was able to recognize correctly (false negative intersection, FNi; 692 named entities). As performance of participating systems varied widely, we also analyzed (3) the false negatives of Hänig et al. (2014) (FN ExB; 1690 named entities).

3.3 Scope of the Analyses

The three mentioned data subsets were created to pursue two analysis goals: first, to investigate to what extent German named entities occur in morphologically altered forms and how complex these are and second, to report and evaluate issues we encountered in the GermEval reference annotations. The first investigation constitutes the main analysis and targets the question of whether there

[5] We adopted the criteria of the official Metric 1 of Benikova et al. (2014a).

[6] We kindly thank the organizers for their support by providing these and also thank the challenge participants that agreed to have them provided to us and shared with the research community as a whole.

[7] according to F_1-measure.

is a morphological gap in German NER. The second examination evolved out of annotation difficulties during the conduction of the first analysis. Even though not intended, we conducted the analysis of the reference annotation issues and present the results because the outcomes can contribute to the general research area of evaluating NER tools' performances.

The three data subsets build the foundation for both examination scopes. To obtain insights into the morphological prevalence and complexity of German named entities, the annotation was conducted according to the following steps: First, the annotator looked at those named entities in the datasets, which deviated from their lexical canonical form (in short LCF) which is the morphologically unmarked form. From gaining an overview of these named entities, linguistic features have been identified that correspond to the morphological segmentation steps which were applied to these morphologically altered named entities (see Sect. 4.1 for a detailed explanation). These linguistic features enable a measurement of the morphological complexity of a given named entity token provided by the reference annotation (i.e., the source named entity, in short SNE), e.g., "Kolpingwerkes" or "Kanadalilie" in Table 1. This measurement, however, required a direct linguistic comparison of the SNEs to their corresponding LCF form (i.e., their target named entity, in short TNE, e.g., "Kolpingwerk" and "Kanada"). Since the reference annotations provided only SNE tokens but no TNE data, a second annotation step was performed in which, all TNEs of the three subsets were manually added to the morphologically altered SNEs respectively[8]. In the third and last step the SNE has been annotated for its morphological complexity based on the numbers of different morphological alterations that were tracked back.

During the second and the third step of the morphological complexity annotation, problematic cases occurred in which a TNE could not be identified for the SNE given in the reference annotation. The reasons underlying these cases have been subsumed under six different annotation issues (details on these are explained in Sect. 5.1), which can significantly affect the performance measure of the tested GermEval NER systems. Therefore, if a SNE could not be annotated for morphological complexity, the causing issue was annotated for this SNE according to the six established annotation issues.

All three created GermEval data subsets have been annotated manually by a native German speaker and linguist and have been partially revised by a native German Computer Scientist while the code for the import and statistics was developed[9].

[8] The choice of a TNE included also the consideration of the four classification labels PER, LOC, ORG and OTH provided together with the SNE.

[9] The entire annotations of the morphological complexity of the named entities as well as the identified reference annotation error types can be consulted in this table including all three data subsets: https://raw.githubusercontent.com/AKSW/germeval-morph-analysis/master/data/annotation_imports/compl-issues-ann-ranks.tsv.

4 Morphological Complexity of German NE Tokens

4.1 Measuring Morphological Complexity

Morphological variation of named entity tokens has been considered as part of the GermEval annotation guidelines. I.e., next to the four named entity types, a marking for SNEs being compound words or derivates of a TNE has been introduced (e.g., LOCderived or ORGpart). While this extension of the annotation of named entity tokens implies that German morphology impacts NER tasks, it does not indicate which morphological peculiarities actually occur. The linguistic analysis investigating morphologically altered SNEs revealed that SNEs exhibit a varying degree of morphological complexity. This degree is conditioned by the morphological inflection and/or word-formation steps that have been applied to a SNE in order to retrace the estimated TNE in its LCF. The resulting formalization of these alternation steps is as follows:

$$L \in \{\mathcal{C}_k\mathcal{D}_l \mid k, l \in \mathbb{N}\} \times \mathcal{P}(\{c, m, f\}) \text{ where}$$

$\mathcal{C}_k$ denotes that k compounding transformations were applied
$\mathcal{D}_l$ denotes that l derivations were applied
c denotes that resolving the derivation applied to the SNE resulted in a word-class change between SNE and TNE
m denotes that the morphological transformation process applied encompasses an inner modification of the TNE stem compared to its LCF
f denotes that the SNE is inflected.

For convenience, we will omit the tuple notation and simplify the set representation of c and f: $\mathcal{C}_1\mathcal{D}_2f, \mathcal{C}_1\mathcal{D}_1cmf, \mathcal{C}_3\mathcal{D}_0 \in L$. In order to obtain the differing levels[10] of morphological complexity for named entities, we went through the identified morphological transformation steps always comparing the given SNE in the test set with the estimated TNE in its LCF. It is defined that all named entities annotated with a complexity other than $\mathcal{C}_0\mathcal{D}_0$ are morphologically relevant and all named entities with a complexity satisfying $\mathcal{C}+\mathcal{D} \geq 1$ (i.e., involving at least one compounding relation or derivation) are morphologically complex, i.e., these require more than one segmentation step in the reanalysis of the SNE to the TNE in its LCF.

Thus, the SNE token can be increasingly complex, if it contains the TNE within a compound part of a compound or if the TNE is embedded within two derivations within the SNE. An example illustrating the morphological segmentation of the SNE "Skialpinisten" is given in Fig. 1. It shows each segmentation step from the SNE back to the TNE in its LCF in detail and illustrates how deeply German named entities can be entailed in common nouns due to morphological transformations. Overall, the annotation of the three subsets revealed

[10] Although, we use the term level to simplify formulations, no strict ordering between the different possible configurations for the aforementioned formalization of complexity is presupposed.

27 levels of morphological complexity for German named entities. The appendix holds a comprehensive listing in Table 4 of these levels together with examples taken from the corpus[11].

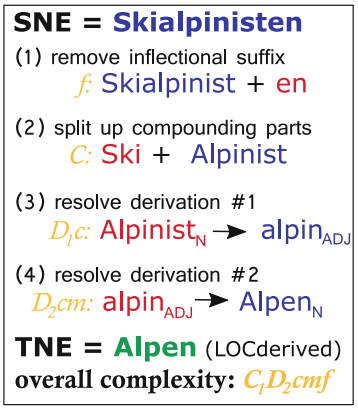


Fig. 1. Example segmentation for annotating the SNE "Skialpinist" with the estimated TNE "Alpen".

4.2 Distribution of Morphologically Complex NE Tokens

Based on our systematization of complexity, we defined more focused complexity criteria such as $\mathcal{C} > 0$ and 'has m' (i.e., inner modification occurred) to complement the criteria morphologically relevant and morphologically complex introduced in Sect. 4.1. Figure 2 shows comparative statistics of the prevalence of named entities matching these criteria for the TPi, FNi and FN ExB[12]. In general, morphologically relevant and morphologically complex named entities are much more prevalent among the false negatives. With respect to the more focused criteria, the strongest increases occur for $\mathcal{C} > 0$, $\mathcal{D} > 0$ and 'is inflected'. In line with the definition of the criterion c, we observe $P(D > 0 \mid c) = 1$. I.e., the occurrence of c in a complexity assignment strictly implies that at least one derivation was applied. The observation of a strong association between inner modification and derivation processes ($P(D > 0 \mid m) = 0.86$) also is in line with intuitive expectations for German morphology.

Figure 3 presents the same comparative statistics between TPi and FNi for the named entities grouped according to their reference classification. In general morphological alteration is more common in named entities annotated with the types PER and LOC. Further, we find lower variance of increase of $C > 0$ across the classes compared to $D > 0$, which is much more common in LOC named

[11] Note, that more levels can be assumed but no occurrences were found in the annotated subsets.

[12] The Scala and Python source code used to prepare the annotations, gather statistics and generate the plots is available at: https://github.com/AKSW/germeval-morph-analysis.

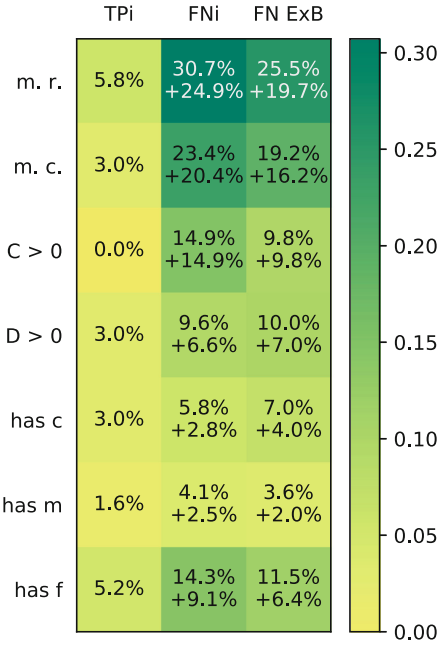


Fig. 2. Prevalence of morphological complexities satisfying specified criteria. Colors encode magnitude of increase of the FN subset compared to the TPi. (m.r. = morph. relevant, m.c. = morph. complex). (Color figure online)

entities (+20.9%) and PER named entities (+12.8%) than in named entities classified ORG and OTH (increase ≤2%). The statistics partitioned by named entity type also reveal that the only types morphologically complex named entities in the TPi subset are LOC named entities with derivations. Analogous statistics between TPi and FN ExB showed similar trends and were omitted for brevity[13].

4.3 Morphological Complexity in Context of NER System Errors

Interestingly, the LOC and PER named entities, that were found to be morphologically complex most often on the one hand are, conversely, the ones covered best by the top GermEval systems according to Benikova et al. (2014a). However, these classes were also deemed more coherent in their analysis, a qualitative impression we share with respect to variety of occurring patterns for morphological alterations. Also, since the morphological complexity of named entities is also one of many factors determining its difficulty to be spotted and typed correctly (besides, e.g., inherent ambiguity of involved lexcial semantics), this might indicate that these two categories might still simply be the ones potentially benefiting most from more elaborate modelling of effects of morphological alteration, as the reported F1 of approx. 84% for LOC and PER still indicates space for improvements.

Further, 19 morphologically complex named entities in FNi could be found, whose TNE was identical with a TNE from the TPi. For example, all systems

[13] The corresponding plot is available at: https://github.com/AKSW/germeval-morph-analysis/blob/master/plots/phrase-partitioned-stats-FalseNegExB.pdf.

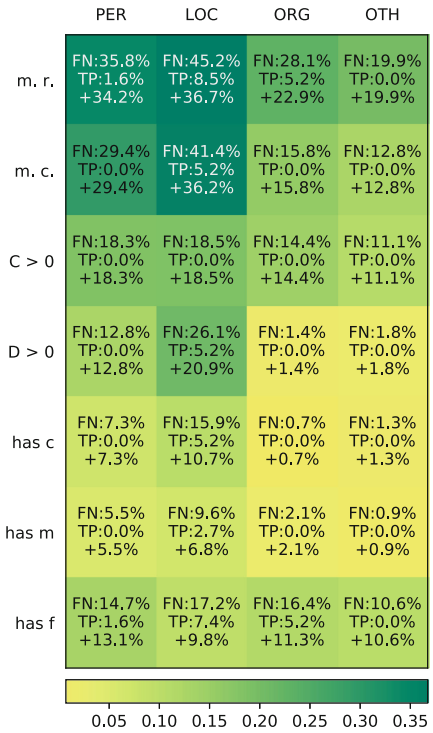


Fig. 3. Prevalence of morphological complexities satisfying specified criteria, grouped be named entity type. Each cell presents ratios in the FNi, the TPi and respective increase. Colors encode magnitude of increase. (m.r. = morph. relevant, m.c. = morph. complex). (Color figure online)

were able to correctly assign LOC-deriv to 'polnischen' (TNE = 'Polen'), however no system was able to recognize 'austropolnischen' (same TNE). Analogously, there is 'Schweizer' in TPi, but 'gesamtschweizerischen' in FNi (common TNE: 'Schweiz'). There were 38 additional morphologically complex named entities in FN ExB with a corresponding TPi named entity sharing the TNE, e.g., 'Japans' (TP) vs. 'Japan-Aufenthaltes' (FN). For all of these pairs, it appears plausible to assume that the difficulty for the corresponding false negative can be attributed to a large extend to the morphological complexity, as simpler variants posed no hindrances to any of the tested systems[14]. For the ExB system, these kind of false negatives constitute 3.4% of all false negatives, which could be viewed raw estimation of potential increase in recall if hypothetically morphological complexity of named entities would be mitigated entirely. It should also be noted that the reported occurrence counts of these pairs for ExB are lower bounds, since not all of its true positives had been annotated at the time of writing.

[14] Still we also acknowledge that several factors of lexical semantics, syntax etc. influence how challenging it is to spot a specific NE occurrence in context and more systematic analysis of these factors would be needed to attribute the error to morphological causes with certainty.

5 Reference Annotation Related Issues

5.1 Reference Annotation Issue Types

During the annotation for morphological complexity issues arose with regard to the GermEval reference annotations which led to various difficulties.

Table 2. Encountered issues pertaining to GermEval reference annotations.

Issue	Example	Prevalence
Not Derived	SNE = *Kirgisische* (LOC-deriv) with TNE = *Kirgistan*	94 (31.5%)
Wrong NE Type	SNE = *barocker* (ORG-deriv) with TNE = *Barock*, "Baroque" is an epoch, it should have been annotated as OTH-deriv	62 (20.8%)
Wrong Spelling	SNE = *Freiburg/31:52* with TNE = *Freiburg*	51 (17.1%)
No NE	SNE = *Junta* - "Junta" is a common noun, there is no TNE	18 (6.0%)
Invalid Reference	SNE = *Was ist theoretische Biologie?* - this is a HTML link label, which is not related to any NE	7 (2.4%)
TNE Unclear	SNE = *Köln/Weimar/Wien* - TNE is unclear, unknown to which of the three named entities is referred to	66 (22.2%)

Overall, six reference annotation issues have been identified and all three subsets have been annotated for these issues (also cf. Table 2):

Issue #1 Not Derived: A significant number of SNEs with the type LOCderived is morphologically not derived from the location TNE but from the inhabitant noun, e.g., "Kirgisisch" is not derived from "Kirgistan" but from "Kirgise".
Issue #2 Wrong NE Type: This issue refers to SNEs which are correctly identified, but are assigned to the wrong named entity category.
Issue #3 Wrong Spelling: SNEs annotated with this issue are either incorrectly spelled or tokenized.
Issue #4 No NE: This issue holds for SNEs, which turn out to be only common nouns in the sentences they occur.
Issue #5 Invalid Reference: SNEs referring to book/film titles, online references or citations which are incomplete, wrong or the online reference is a title for a website given by some person but not the real title or URL.

Issue #6 TNE Unclear: This issue summarizes reasons for preventing a TNE of being identifiable form a given SNE, i.e., it is not possible to morphologically decompose the SNE to retrieve the TNE or there are more than one TNEs included in the SNE.

If Not Derived, No NE, Invalid Referenceor TNE Unclearoccur for a named entity, assignment of a morphological complexity level becomes impossible. Consequently, the corresponding named entities (189) were excluded from the complexity statistics presented in Sects. 4.2 and 4.3. Wrong NE Typeand Wrong Spelling, on the other hand, albeit also implying difficulties for NER systems, do not interfere with identifying the TNE (and thus the complexity level). Hence, such named entities were not excluded.

5.2 Distribution and Effects of Annotation Issues

Table 2 provides, in addition to examples for the aforementioned categories of annotation issues, their total prevalence across TPi and FN ExB (subsuming FNi). Table 3 additionally indicates the distribution of issue occurrences in comparison between the subsets. Overall, occurrence of annotation issues are about three times more likely in the false negative sets compared to TPi, a trend in a similar direction as for the occurrence of morpholoically complex named entities.

Table 3. Frequencies of occurrence of annotation issues by category and subset. Percentages in parentheses are relative frequencies for the corresponding subset.

Issue	TPi	FNi	FN ExB
#1	41 (4.07%)	18 (2.60%)	53 (3.14%)
#2	0 (0.00%)	30 (4.34%)	62 (3.67%)
#3	1 (0.10%)	24 (3.47%)	50 (2.96%)
#4	1 (0.10%)	10 (1.45%)	17 (1.01%)
#5	0 (0.00%)	4 (0.58%)	7 (0.41%)
#6	0 (0.00%)	19 (2.75%)	66 (3.91%)
All	43 (4.27%)	105 (15.17%)	255 (15.09%)

It appears questionable to count named entities with Wrong NE Type, No NE and Invalid Reference that have not been recognized by any NER system as a false negative, as these named entities do not actually constitute named entities as defined by the guidelines (analogously for true positives). Thus, we projected the M1 performance measures on the test split for the ExB system disregarding these named entities[15]. The adjustment results in discounting five

[15] Due to lack of complete screening of all true positives of ExB for annotation issues we linearly interpolated the exemption of one true positive according to TPi to the exemption of five true positives for all true positives of that system.

false positives and 44 false negatives, result in an increase in recall by 0.48% and F1 by 0.34%. Although, this change is not big in absolute magnitude, it can still be viewed relevant considering that the margin between the to best systems at GermanEval was merely 1.28% for F1 as well Benikova et al. (2014a).

6 Conclusion

This study presented an analysis of German NER as reflected by the performance of systems that participated in the GermEval 2014 shared task. We focused on the role of morphological complexity of named entities and introduced a method to measure it. We compared the morphological characteristics of named entities which were identified by none of the systems (FNi) to those identified by all of the systems (TPi) and found out that FNi named entities were considerably more likely to be complex than the TPi ones (23.4% and 3.0% respectively). The same pattern was detected also for the system which achieved the best evaluation in this shared task. These findings emphasize that morphological complexity of German named entities correlates with the identification of named entities in German text. This indicated that the task of German NER could benefit from integrating morphological processing.

We further discovered annotation issues of named entities in the GermEval reference annotation for which we provided additional annotation. We believe that the presented outcomes of this annotation can help to improve the creation of NER tasks in general.

As a future work, we would like to extend our annotation to analyze how these issues affect the evaluation of the three best performing systems more thoroughly. In addition, a formalization to measure the variety of occurring patterns of morphological alteration (used affixes/affix combinations, systematic recurrences of roots...) as a complementary measure for morphological challenges seems desirable. We will further have multiple annotators to morphologically annotate the named entities of the GermEval reference, in order to estimate the confidence of our observation by measuring inter-annotator agreement.

Acknowledgment. These research activities were funded by grants from the H2020 EU projects ALIGNED (GA-644055) and FREME (GA-644771) and the Smart Data Web BMWi project (GA-01MD15010B).

Appendix

Table 4. Distribution of the morphological complexities in the annotated subsets

Compl.	TPi	FNi	FN ExB	Example SNE	Example TNE
$\mathcal{C}_0\mathcal{D}_0$	910 (94.20%)	442 (69.28%)	1149 (74.47%)	Mozart	Mozart
$\mathcal{C}_0\mathcal{D}_0 f$	27 (2.80%)	47 (7.37%)	98 (6.35%)	Mozarts	Mozart
$\mathcal{C}_1\mathcal{D}_0$	0 (0.00%)	62 (9.72%)	101 (6.55%)	Mozart-Konzert	Mozart
$\mathcal{C}_1\mathcal{D}_0 f$	0 (0.00%)	15 (2.35%)	24 (1.56%)	Mozart-Konzerten	Mozart
$\mathcal{C}_1\mathcal{D}_0 m$	0 (0.00%)	3 (0.47%)	5 (0.32%)	Pieterskirche	Pieter
$\mathcal{C}_1\mathcal{D}_0 mf$	0 (0.00%)	3 (0.47%)	4 (0.26%)	Reichstagsabgeordneten	Reichstag
$\mathcal{C}_0\mathcal{D}_1$	0 (0.00%)	9 (1.41%)	20 (1.30%)	Donaldismus	Donald
$\mathcal{C}_0\mathcal{D}_1 f$	0 (0.00%)	1 (0.16%)	4 (0.26%)	Donaldismusses	Donald
$\mathcal{C}_0\mathcal{D}_1 m$	0 (0.00%)	7 (1.10%)	10 (0.65%)	Nestorianismus	Nestorius
$\mathcal{C}_0\mathcal{D}_1 mf$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Spartiaten	Sparta
$\mathcal{C}_0\mathcal{D}_1 c$	5 (0.52%)	16 (2.51%)	61 (3.95%)	Japanisch	Japan
$\mathcal{C}_0\mathcal{D}_1 cf$	9 (0.93%)	8 (1.25%)	14 (0.91%)	Japanischen	Japan
$\mathcal{C}_0\mathcal{D}_1 cm$	1 (0.10%)	1 (0.16%)	6 (0.39%)	Europäisch	Europa
$\mathcal{C}_0\mathcal{D}_1 cmf$	10 (1.04%)	8 (1.25%)	19 (1.23%)	Europäischen	Europa
$\mathcal{C}_2\mathcal{D}_0$	0 (0.00%)	3 (0.47%)	5 (0.32%)	Bibelforscherfrage	Bibel
$\mathcal{C}_2\mathcal{D}_0 mf$	0 (0.00%)	1 (0.16%)	1 (0.06%)	Erderkundungssatelliten	Erde
$\mathcal{C}_1\mathcal{D}_1$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Benediktinerstift	Benedikt
$\mathcal{C}_1\mathcal{D}_1 f$	0 (0.00%)	2 (0.31%)	2 (0.13%)	Transatlantikflüge	Atlantik
$\mathcal{C}_1\mathcal{D}_1 m$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Römerstrasse	Rom
$\mathcal{C}_0\mathcal{D}_2$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Geismarerin	Geismar
$\mathcal{C}_0\mathcal{D}_2 f$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Hüttenbergerinnen	Hüttenberg
$\mathcal{C}_0\mathcal{D}_2 m$	0 (0.00%)	0 (0.00%)	1 (0.06%)	Rheinländerin	Rheinland
$\mathcal{C}_0\mathcal{D}_2 cf$	0 (0.00%)	1 (0.16%)	1 (0.06%)	Austropolnischen	Polen
$\mathcal{C}_0\mathcal{D}_2 cmf$	4 (0.41%)	0 (0.00%)	3 (0.19%)	Transatlantischen	Atlantik
$\mathcal{C}_3\mathcal{D}_0$	0 (0.00%)	1 (0.16%)	1 (0.06%)	25-US-Dollar-Marke	US
$\mathcal{C}_1\mathcal{D}_2 cf$	0 (0.00%)	2 (0.31%)	2 (0.13%)	Gesamtschweizerischen	Schweiz
$\mathcal{C}_1\mathcal{D}_2 cmf$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Skialpinisten	Alpen
Total	966	638	1543		

References

Benikova, D., Biemann, C., Kisselew, M., Padó, S.: GermEval 2014 named entity recognition shared task: companion paper. In: Workshop Proceedings of the 12th Edition of the KONVENS Conference, Hildesheim, Germany, pp. 104–112 (2014a)

Benikova, D., Biemann, C., Reznicek, M.: NoSta-D named entity annotation for German: guidelines and dataset. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, pp. 2524–2531. European Language Resources Association (ELRA) (2014b)

Benikova, D., Muhie, S., Prabhakaran, Y., Biemann, S.C.: GermaNER: free open German named entity recognition tool. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, Duisburg-Essen, Germany, pp. 31–38. German Society for Computational Linguistics and Language Technology (2015)

Capsamun, R., Palchik, D., Gontar, I., Sedinkina, M., Zhekova, D.: DRIM: named entity recognition for German using support vector machines. In: Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany, pp. 129–133 (2014)

Chrupała, G., Klakow, D.: A named entity labeler for German: exploiting Wikipedia and distributional clusters. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta. European Language Resources Association (ELRA) (2010). ISBN 2-9517408-6-7

Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Inf. Process. Manag. **51**(2), 32–49 (2015)

Faruqui, M., Padó, S.: Training and evaluating a German named entity recognizer with semantic generalization. In: Proceedings of KONVENS 2010, Saarbrücken, Germany (2010)

Finkel, J.R., Manning, C.D.: Nested named entity recognition. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 141–150. Association for Computational Linguistics (2009)

Hänig, C., Bordag, S., Thomas, S.: Modular classifier ensemble architecture for named entity recognition on low resource systems. In: Workshop Proceedings of the 12th Edition of the KONVENS Conference, Hildesheim, Germany, pp. 113–116 (2014)

Helmers, L.A.: Eigennamenerkennung in Web-Korpora des Deutschen. Eine Herausforderung für die (Computer)linguistik. Bachelor thesis, Humboldt-Universität zu Berlin (2013)

Reimers, N., Eckle-Kohler, J., Schnober, C., Gurevych, I.: GermEval-2014: nested named entity recognition with neural networks. In: Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition, Hildesheim, Germany, pp. 117–120 (2014)

Schäfer, R., Bildhauer, F.: Building large corpora from the web using a new efficient tool chain. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 486–493. European Language Resources Association (ELRA) (2012)

Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop, pp. 47–50 (1995)
Schüller, P.: MoSTNER: morphology-aware split-tag German NER with Factorie. In Workshop Proceedings of the 12th Edition of the KONVENS Conference, Hildesheim, Germany, pp. 121–124 (2014)

2.2 MMoOn Core – The Multilingual Morpheme Ontology

In this publication the MMoOn Core ontology as the first proposal for a discrete representation model for the domain of morphological language data is presented. An underlying domain analysis as well as the design principles and its integration into related works are explained. In addition to the description of the main classes and properties created within MMoOn Core, the architectural set up that allows to create MMoOn-based datasets, so called MMoOn morpheme inventories, is introduced. It will be clarified how this ontology constitutes a possibility to serve the interdisciplinary need for language resources. On the one hand, MMoOn-based datasets enable linguists to gain more analytical insights into their own language data. On the other hand, computational linguistic researchers are enabled to include more specific morphological language data into their system settings to reach better results. Furthermore, several use cases for the research fields of linguistics and NLP are illustrated. These show how the ontology contributes to the representation of morphological language data by (i) accounting for the fine-grained and specific empirical data linguists compile and document and (ii) by exemplifying its application in computational linguistics.

Semantic Web 4 (2020) 1–30
IOS Press

MMoOn Core - The Multilingual Morpheme Ontology

Bettina Klimek [a], Markus Ackermann [a], Martin Brümmer [b], Sebastian Hellmann [a].
[a] *KILT Research Group, Universität Leipzig, Institut für Angewandte Informatik (InfAI e.V.), Germany.*
Email: (klimek|ackermann|hellmann)@informatik.uni-leipzig.de
[b] *Independent Researcher, Leipzig, Germany.*
Email: der.bruemmer@gmail.com

Editor: Philippe Ciminiano, Universität Bielefeld, Germany
Solicited reviews: John McCrae, National University of Ireland Galway, Insight Centre for Data Analytics, Ireland; One anonymous reviewer

Abstract. In the last years a rapid emergence of lexical resources has evolved in the Semantic Web. Whereas most of the linguistic information is already machine-readable, we found that morphological information is mostly absent or only contained in semi-structured strings. An integration of morphemic data has not yet been undertaken due to the lack of existing domain-specific ontologies and explicit morphemic data. In this paper, we present the Multilingual Morpheme Ontology called MMoOn Core which can be regarded as the first comprehensive ontology for the linguistic domain of morphological language data. It will be described how crucial concepts like morphs, morphemes, word forms and meanings are represented and interrelated and how language-specific morpheme inventories can be created as a new possibility of morphological datasets. The aim of the MMoOn Core ontology is to serve as a shared semantic model for linguists and NLP researchers alike to enable the creation, conversion, exchange, reuse and enrichment of morphological language data across different data-dependent language sciences. Therefore, various use cases are illustrated to draw attention to the cross-disciplinary potential which can be realized with the MMoOn Core ontology in the context of the existing Linguistic Linked Data research landscape.

Keywords: MMoOn, Linguistic Linked Data, morphology, morpheme ontology, inflection, derivation, interlinear morphemic glossing, OntoLex-*lemon*

1. Introduction

Morphological language data (MLD) plays a crucial role across various interdisciplinary research fields. Traditionally, linguists have fundamentally studied morphology on both language-independent and language-specific levels for centuries in order to investigate the underlying mechanisms that a) allow new words to emerge that are not yet recorded in dictionaries (i.e. word formation), b) are required to alter words so that they take the appropriate form within a certain syntactic environment (i.e. inflection) and c) explain to what extent languages structurally differ in encoding lexical or grammatical meanings within words (i.e. comparative linguistics). This work is the basis for the far younger research field of natural language processing (NLP) which strives to apply linguistic knowledge on morphology (in conjunction with other linguistic areas) on large amounts of text in order to automatically analyze, process or create natural language content. While the methods and aims of linguistics and NLP differ, both sciences can highly benefit each other. Within an ideal cycle of interdisciplinary exchange NLP would take the insights on morphology provided by linguists, apply them to large amounts of text and feed back their results to the linguists who could refine their studies on morphology, which in turn would lead to a better research basis that can be taken up by NLP research again.

Both research fields heavily rely on MLD. The realization of the described scientific exchange and ad-

vancement is, however, prevented because of the existing data silos on both sides which use many different and non-interoperable data formats, thus, impeding an easy data transfer. Due to the emergence of Semantic Web technologies this state can change. Being based on the principles of Linked Data, they have proven to evoke true data-driven interdisciplinarity for research domains shared by different sciences. This research manifests itself in the area of Linguistic Linked Open Data (LLOD) which was initiated in 2010 with the foundation of the Working Group on Open Data in Linguistics (OWLG) [9, 40]. Since then a significant rise of language data on the Semantic Web emerged. Academic, industrial and technological interest into Linguistic Linked Data appeared and materialized in three areas: (1) W3C community groups such as Linked Data for Language Technology (LD4LT)[1] or BPMLOD[2], and (2) European research projects such as LIDER[3], Falcon[4] or FREME[5] as well as (3) scientific workshops and special issues such as the workshop series on Linked Data in Linguistics[6], the Multilingual Semantic Web workshop series[7] or the special issue of the Semantic Web Journal on Multilingual Linked Open Data [27][8].

A cross-disciplinary usage of LLOD has already been proven to be achievable in the case of the OntoLex-*lemon* model[9] [41] which successfully unified linguistic and NLP research data for lexical language data (LLD). However, a similar approach for MLD is not yet established. While a plethora of linguistic resources[10] for the LLD domain exists and is highly reused, there is still a great gap for equivalent morphological datasets and ontologies [6, 27]. Therefore, the aim of this paper is to present the **M**ultilingual **Mo**rpheme **On**tology, in short MMoOn Core. The goal of the MMoOn Core ontology is to represent the domain of morphology in a granular way and to assign semantics at the appropriate subword layers in order to derive compositional semantics on the morph, morpheme and word levels. In particular it enables the representation of the morphemes including their written representations and meanings as well as their relations to the words in which they can occur. It is designed to meet the documentary needs of linguists and the applicatory needs of NLP researchers alike. MMoOn Core serves as an extensible schema conceptualizing the domain of morphology and is not bound to any specific natural language but also enables the creation of language-specific MMoOn morpheme inventories. Because of the language-independent conceptualization as well as the evolutionary process of the model, MMoOn Core is suitable for describing any inflectional language. Multilingualism is accounted for automatically since the created MMoOn morpheme inventories are inherently interconnected through the MMoOn Core ontology. With a rising number of morpheme inventories multilingual interlinking will constantly increase over time, hence, the name **Multilingual** Morpheme Ontology. Ultimately, MMoOn Core has been created to serve as a shared semantic model for representing MLD and to enable the exchange, reuse and enrichment of MLD across different data-dependent language sciences.

Extracting and explicating the morphological semantics of words, however, requires not only a domain expert with detailed linguistic knowledge about morphology but also close to native-speaker level knowledge about the language. Even though ontologies such as the OntoLex-*lemon* model [41], LexInfo [11], OLiA [8], or GOLD [18] partially define a minimal RDF vocabulary to describe morphemes and morphological data as such, a dedicated morpheme ontology capturing and formalizing semantics is still missing.

This becomes obvious through the fact that morphological information is predominantly still attached to the lexeme (the unit that carries lexical meaning) or the whole word form (cf. Example 1[11]) and not to the morphological segment (cf. Example 2). The current research gap has two dimensions: First, none of the above-mentioned ontologies provides sufficiently granular terminology to properly describe and tag word segments and second, interoperable morphological data is consequently not available.

[1] https://www.w3.org/community/ld4lt/
[2] https://www.w3.org/community/bpmlod/
[3] http://www.lider-project.eu/
[4] http://falcon-project.eu/
[5] http://www.freme-project.eu/
[6] http://ldl2018.linguistic-lod.org/
[7] http://msw4.insight-centre.org/
[8] http://www.semantic-web-journal.net/blog/call-multilingual-linked-open-data-mlod-2012\-data-post-proceedings
[9] https://www.w3.org/2016/05/ontolex/
[10] Cf. the emergence and development of the Linguistic Linked Open Data Cloud: http://linguistic-lod.org/llod-cloud.

[11] This paper follows the generic style rules for linguistic [25]. This means that italics are used for all object-language forms (words and morphs) that are cited within the text or examples and single quotation marks are used for indicating linguistic meanings (morphemes).

(1) Word form: *players*
Annotation: NNP
Meaning: 'noun, plural, common'[12]

(2) Word form: *players*
Morphs: *player-s*
Morphemes: 'player'-PL[13]

In contrast to digital and Linked Data dictionaries or lexicons, morphemic language resources are mostly available in layout-centric formats, such as HTML website contents, PDF documents, tables or even only in printed media. What is more, the domain of morphology is to a large extent treated by linguists who do not only differ in their understanding of this linguistic area but also compile morphological data with a focus on consumption by humans and not on machine processability. The creation of the MMoOn Core model consequently strives to tackle these challenges and will add the following contributions:

- Provide a fine-grained and extensive semantic model for representing MLD suitable for linguistic and NLP tasks.
- Publication of MMoOn Core as a language-independent conceptualization of the MLD domain as a freely available, reusable and extensible linguistic resource.
- Linking of MMoOn Core to already existing linguistic data models.
- First compilation of derivational meanings.
- Representation of morphemic glosses as Linked Data.
- Usage of MMoOn Core as a unifying building block to compile language-specific morpheme inventories which:
  - integrate heterogeneous data sources with semantic consistency,
  - provide resource descriptions for word forms and morphemic language data,
  - interrelate language elements across the morph, word form and lexeme level,
  - include direct extensions of the vocabulary with language-specific meanings,
  - are automatically multilingually interconnected through an underlying shared semantic,
  - result in a compilation of natural language data in a machine-readable manner by adhering to Linked Data principles and interlinking.

The remainder of the paper is structured as follows: Section 2 states the motivation and background and is followed by an outline of related work in Section 3, also pointing to gaps in existing resources. After introducing a brief domain analysis in Section 4, the main part of the paper – the Multilingual Morpheme Ontology – will be presented in detail in Section 5. This part includes its architectural setup, design principles as well as its basic elements. A more detailed comparison of MMoOn Core to OntoLex-*lemon* is provided in Section 6 by taking a closer look at the currently developing morphology module. Furthermore, use cases for the application of MMoOn Core for linguistic and NLP research will be outlined in Section 7. Finally, the paper closes with concluding remarks and a prospect of the future work in Sections 8 and 9.

2. Motivation and background

The need for the development of a data model that is able to describe the morphemic inventories of natural languages was expressed by two major research communities. The first one centers around the community groups OWLG[14], LD4LT[15] and BPMLOD[16] and consists of researchers coming from the areas of computational linguistics, NLP, machine translation and language technologies. They express a high demand on interoperable and fine-grained (multilingual) linguistic data that models subword information and which can be integrated in and applied to the existing content and language analyzing systems. The abovementioned groups also expressed a strong preference for free and open data to increase reusability and reproducibility.

The second group of researchers involves linguists whose main subject area is the investigation of natural language per se. Especially linguists who document endangered and under-resourced languages as

[12] Taken from the Lancaster tagset: http://www.scs.leeds.ac.uk/amalgam/tagsets/lob.html

[13] This kind of morphological representation is well established practice in linguistics and widely known as interlinear morphemic glossing [13, 37].

[14] https://linguistics.okfn.org

[15] https://www.w3.org/community/ld4lt

[16] https://www.w3.org/community/bpmlod

well as general comparative linguists both produce and rely on adequate linguistic data. A rising awareness of methodological standards in the compilation of language data has emerged in linguistic research "for the sake of [the] speech communities [of languages threatened by extinction] and their interest in their cultural tradition and for the sake of the very database of the discipline itself" [36]. In linguistics the usage of interlinear or morpheme-by-morpheme glosses as a means for the representation of the segments and meanings of text are an established common practice. Due to their widespread application, efforts of standardization have been introduced [13, 37]. As a result, a great amount of interlinear-glossed text resources exist in linguistic databases or as text examples in linguistic publications. Unfortunately, this wealth of data is not easily accessible or reusable due to the (1) technical heterogeneity, (2) license restrictions or unavailability of licenses, and (3) nonformal description of linguistic documentation. Here, the field of linguistic documentation is in need of a model that allows for the (automatic) creation, retrieval, processing and publishing of its morphological data in compliance with the granularity of the linguistic representation levels.

In order to fulfill the demands of both research communities just outlined, the MMoOn Core ontology has been created. It presents a new vocabulary which is easily integrable into already existing lexical resources and expressive enough to capture the various correspondences between subword elements and their associated meanings. Hence, all specific MMoOn language inventories will contribute to the development of natural language analyzing methods and tools. At the same time, MMoOn allows linguists to adequately represent their high-quality language data using a vocabulary with well-defined semantics and in a data format that ensures interoperability with a large range of formats and systems. Thus, we believe that, both the NLP research area and linguistics as an empiric discipline will benefit from the reuse of the MMoOn Core vocabulary.

The developmental approach underlying the creation process of the MMoOn Core ontology is grounded in a thorough domain analysis (cf. Section 4) and guided by a defined set of requirements as well as design choices (which are explained in detail in Section 5.3). To this extent, it has been developed from scratch as a standalone ontology without originating from any existing vocabulary or model. On the contrary, the aim of the MMoOn Core ontology is to unite morphological data represented in differing formats or underlying varying linguistic theories and descriptions. Since MMoOn Core further pursues the aim to function as a language-independent domain ontology for MLD, the generalizable elements, relations and characteristics which have been identified for the linguistic research field of morphology [4, 24] have been derived and transformed within the semantic modeling of the ontology. These include linguistic concepts such as *affix*, *inflection*, *derivation*, *segmentation*, *meaning* or *interlinear glossing* as described in the foundational linguistic works about morphology and are not only assumed to be applicable to a wide range of languages but also to be familiar concepts to linguists. Under consideration that linguists create the most fine-grained MLD, MMoOn Core is motivated by the provision of as many descriptive domain elements as possible to keep the entry barrier into working with RDF for linguists as low as possible. To conclude, the MMoOn Core ontology can be regarded as the first extensive representation model for MLD to create inventories of the smallest meaningful elements of language similar to dictionaries or lexical databases within the lexical data domain.

3. Gaps in existing resources and related work

An inventory of morphemes requires an appropriate data model on the one side and morphemic data on the other side. In what follows an overview will be given that investigates the applicability of existing linguistic ontologies as well as existing Linked Data morphological resources but also datasets and sources that are based on other formats.

3.1. Vocabularies modeling MLD

Within the last few years, ontologies emerged that contain vocabularies partially describing morphological aspects of language. These include the *lemon* model [39] and the decomp and ontolex submodules of the OntoLex-*lemon* model [41], LexInfo [11], OLiA [8] and the GOLD [18] ontology. Even though, none of these vocabularies were explicitly designed to capture the domain of MLD, they include conceptual information on the meaning side of morphemes and/or information of morphemic elements. For that reason the MMoOn Core ontology has been interlinked to some of these vocabularies (cf. section 5 and section 6) in order to comply to the Semantic Web best practices

for reusing existing data models. In this context LexInfo, OLiA and GOLD are mainly reusable as terminological datasets providing the theoretical description of the linguistic concepts involved in lexicography and morphology.

With regard to the representation of subword units *lemon* and OntoLex-*lemon* provide elements that belong to the domain of MLD. *Lemon* was the first model to offer a morphology module[17] that allows the representation of different forms of lexical entries including `lemon:Part` which describes affixes. This module evolved to be a standalone ontology called LIAM (Lemon Inflectional Agglutinative Morphology)[18]. However, this vocabulary focuses on a regular expression based description of morphological processes and pattern transformation [41]. The crucial information – namely the morphemic segment – is contained as string in the data type property `liam:rule` and, therefore, not machine processable and not further interrelatable to other segments. In addition to that, the applicability of *lemon* and the LIAM ontology with regard to language-specific modeling of morphological data has been questioned in previous work [7].

The latest advancement in modeling MLD is presented in the W3C report of the OntoLex-*lemon* model specification[19]. Especially the ontolex and decomp modules are highly reused for representing lexical data but also compositional morphology. Still, the morphological elements such as `decomp:Component` and `ontolex:Affix` are too coarse grained and mainly intended to represent compounding morphology. Further, specific elements like roots and stems or more specific affixes like the transfix or empty morph are missing together with the necessary relations that represent the segmentation steps and relations between the morphemic elements. Additionally, word forms are only encoded as strings via the `ontolex:otherForm` datatype property which prohibits a further specification of the derivational and inflectional segments a word form may consist of. Nonetheless, the OntoLex-*lemon* model serves as the ontological standard for modeling linguistic language data to a large extent of the LLOD community and is highly reused. For that reason – and because of the significant overlap of the two domains of lexical and morphological language data – it was out of question to interconnect MMoOn Core with OntoLex-*lemon* in order to enable an interconnection but also the supplementation of both domain models [33] (cf. section 6).

The recently published Ligt vocabulary has to be mentioned as a possibility for representing morphological data as well [10]. It is specialized to enable the transformation of interlinear glossed text into RDF data. In particular, it can be used to transform resources based on Toolbox, FLEx and Xigt (eXtensible Interlinear Glossed Text) to Ligt-RDF. The main contribution of Ligt is the unification of several heterogeneous interlinear glossed text resources based on different formats within a homogeneous RDF data graph. With respect to its usability for representing MLD, however, the Ligt vocabulary differs fundamentally from MMoOn Core and OntoLex-*lemon* in that the morphemic elements it describes identify single occurrences of morphs within an interlinear text similar to tokens within a corpus. As a result, the only element relevant for the domain of MLD in Ligt is the class `ligt:Morph` which is specified with a string and for its position within the morph tier, paragraph and document it occurs. No semantics is established interrelating `ligt:Morph` resources or specifying them, e.g. as suffixes, derivational or inflectional morphs or for their meanings. In fact, a gloss tier that would interrelate the morphs with the abstract identities of their morphemic meanings, i.e. the glosses, is not provided in Ligt. Since the objective of Ligt is to represent unique occurrences of morphs instead of unique morphemic concepts that can be applied to an unlimited number of occurrences in primary language data, reasoning over the `ligt:Morph` resources to obtain more insights is not possible. Whereas MMoOn Core is intended to provide a vocabulary for obtaining domain knowledge about the morphological inventory of a language, in the realm of the MLD domain Ligt-based datasets rather function as the attestations for the morphs of a language. In that respect, the Ligt creators deliberately decided to consider the provision of comprehensive MLD semantics out of scope for this vocabulary in favor of gaining unified representations of various interlinear glossed text formats. This choice is especially advantageous because it not only facilitates the application of the vocabulary in practice but also allows for an easier interlinking – if required – with already existing semantically richer domain vocabularies for MLD, including MMoOn Core. Even though no published dataset based on Ligt exists to date, the significance and need of such datasets is already obvious given that interlinear text resources are quite often the only existing documented language resources for

[17] http://lemon-model.net/lemon-cookbook/node35.html
[18] http://lemon-model.net/liam
[19] https://www.w3.org/2016/05/ontolex/

less- or under-resourced languages (cf. Section 3.3). In this respect Ligt datasets could be potential sources to derive an attested MMoOn morpheme inventory for a language from interlinear text resources, similarly to a dictionary that is derived from corpus data.

3.2. Overview of Linked Data resources

So far, two datasets have been created and published based on the MMoOn Core model and architecture (cf. Section 5.2), i.e. the Hebrew Morpheme Inventory [32] and the Xhosa RDF dataset [5] together with a dictionary alignment to Kalanga and Ndebele lexical datasets [17].

To the best of our knowledge, all other existing Linked Data resources including MLD are based on the *lemon*/LIAM model or the OntoLex-*lemon* model. As a consequence, these datasets contain morphological data only to a limited extent, e.g. the decomposition of compounds or unrelated affix resources (e.g. [16]). As a specific example for a dataset containing inflectional language data, the Dbnary "morpho" Wiktionary extractions for German, French, English and Serbo-Croation need to be mentioned[20]. These datasets contain the Wiktionary headwords and inflected word forms in *lemon*-RDF and are annotated for their inflectional meanings with OLiA [46]. However, in a strict view of the domain of MLD (cf. Section 4) this representation of morphological data covers only the morphological meanings as word form annotations instead of segmented morphs that correspond to a specific meaning. Notwithstanding the fact that the Wiktionary data does not contain segmentations of word forms, an adequate representation of these segments and their interrelation to each other and within the word forms is not possible with the existing vocabularies, with the exception of the MMoOn Core ontology.

3.3. Overview of non-Linked Data resources

Due to the fact that the Linked Data paradigm is in comparison to linguistic research and documentation very young, it is not surprising that the majority of MLD exists in non-Linked Data formats. In fact, the largest part of linguistic data is preserved in documents. However, this overview of MLD will not touch upon such data in unstructured formats but focuses on structured data only. Among the datasets which can be found a high variance with regard to aspects like accessibility, data quality, reusability, complexity of morphological data, covered languages and data format can be observed:

a) MLD in linguistic field work data: This kind of data entails fine-grained, complex and segmented MLD documented in interlinear glossed texts that are edited with specific tools like FieldWorks[21] or FLEx. Usually the data is compiled by one linguist for an undocumented, small or endangered language. Hence, the resulting datasets are of high quality but often not very large and commonly meant for linguistic research. The formats of the field linguists' tools are very specific and the output dataset is not seldom published at all. Instead, only a part of it is used for giving language examples in resulting text publications, i.e. in PDF documents. However, efforts like TypeCraft[22] [2] and Dictionaria[23] emerged that aim at providing an open and data driven publication platform for publishing full FieldWorks datasets. What is more, they also provide the data in common formats like XML, CSV, JSON and XLS[24].

b) MLD as a part of large language databases: For large and well documented languages usually more linguistic data is available to date. Whole research groups and institutes are devoted to collecting and editing resources such as word lists, dictionaries and corpora and also strive to organize and manage all the linguistic data available in large databases. These datasets also cover MLD like word forms, inflection tables and affix lists. These language resources are the outcome of a collaborative work between linguists and computer linguists that merge and structure manually compiled data as well as automatically transformed or created language data. Examples include the Oxford Online Database of Romance Verb Morphology[25], the work of the Surrey Morphology Group [26] and the French project

[20] http://kaiko.getalp.org/about-dbnary/download/

[21] https://software.sil.org/fieldworks/

[22] https://typecraft.org/

[23] http://dictionaria.clld.org/

[24] Even though the datasets published by Dictionaria are also provided in RDF, this information is omitted here because no standard vocabulary for linguistic Linked Data has been used and only a part of the original data is transformed into RDF, i.e. only the headwords encoded in literals. Instead, very basic vocabularies such as SKOS and DCTERMS have been used. As a consequence, the morphological data that is entailed in the original source dataset is either missing completely or not differentiable from the lexical data within the delivered RDF datasets.

[25] http://romverbmorph.clp.ox.ac.uk

[26] http://www.smg.surrey.ac.uk/

ALEXINA[27][45] which develops morphological NLP lexicons. For German language data in particular, the German Institute of Language (IDS[28]) poses a considerable source for basic words and word forms[29] and also provides the dictionary of affixes[30].

In this context, the Lexical Markup Framework (LMF) [20, 21] has to be mentioned as well. It enables the representation of machine-readable dictionaries (MRD) and NLP lexicons and has been applied to create numerous datasets, e.g. ALEXINA, including morphological data based on the morphological extension of the LMF core model. It provides two strategies for representing word forms. The first one applies to an extensional listing of all forms of a lexical entry which are specified for linguistic categories and values. This approach, however, does not explicitly contain morphemes. The second strategy allows for an intensional modeling of so called morphological patterns and inflectional paradigms. These are formalized in detail and specific to lexical entries, however, with no explicit listing of the forms in the lexicon. While the usage of the morphological extension of LMF is very powerful in terms of machine-processing, it is less suitable as a human-understandable basis for a linguistic analysis of the morphology of a language. The lexicon-centric view on morphology additionally reduces morphology to the lexical entry level and impedes the identification of the smallest meaning bearing units of a language on the word form level. Moreover, LMF-based databases are often realized in structured formats such as XML and very customized. As a result, a considerable effort to understand the data is required and a direct data reuse and interoperability is, therefore, reduced.

c) MLD as morphological segmentation tool output: One of the most challenging tasks in computational linguistics is the creation of segmentation tools. Irrespective of the accuracy and quality of the segmentations, such data outputs also create MLD which can be used in several NLP tasks and linguistic research alike. The IDS developed the Morphisto segmentation tool[31] which is freely available. It analyzes a word form with regard to its grammatical features, the lexical word it belongs to as well as it identifies prefixes and suffixes. Nonetheless, the corresponding morphemic parts of the word, even though involved in the analysis process, are not given in the segmented output. Furthermore, morphological data and tools are provided by the Morpho Challenge workshops[32] which aim at discovering morphemes from text input by statistical machine learning algorithms. One considerable development in this area is the Morfessor tool[33]. In contrast to Morphisto, Morfessor is a generic language-independent segmentation tool that outputs a morphological lexicon on the basis of probabilistic measurements. While the initial effort did not go beyond the identification of morphemes as string sequences [14], it has been extended to consider meaning parameters as well [15]. Albeit, these comprise rather formal aspects again, such as frequency and length, with the authors admitting that "so far the modeling of meaning has only been touched upon" [15]. It has to be stressed that, even though, such tools present a promising method for obtaining MLD for any language, the actual application of these tools requires a lot of time, i.e. time to understand the customized (and often proprietary) output data as well as time for the postprocessing needed for the quality assessment or even data clean up.

The presented overview of Linked and non-Linked Data resources for MLD illustrates two research fields which develop independently from another, even though, both would increase their scientific outcomes by joining their methods and resources as it has been shown for the domain of lexical language data already. In line with the need for lexical data there is also a demand for morphological data that applies both to the language specific morphological domain requirements and to cross-lingual interoperable data modeling. Given the current state of the art, Linked Data vocabularies are not suitable enough to represent the various existing morphological data that will stay isolated and hard to reuse without the unifying RDF data format.

[27] Atelier pour les LEXiques INformatiques et leur Acquisition, http://gforge.inria.fr/projects/alexina

[28] http://www1.ids-mannheim.de/start

[29] http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html

[30] http://hypermedia.ids-mannheim.de/call/public/gramwb.ansicht?v_app=g

[31] http://www1.ids-mannheim.de/lexik/home/lexikprojekte/lexiktextgrid/morphisto.html

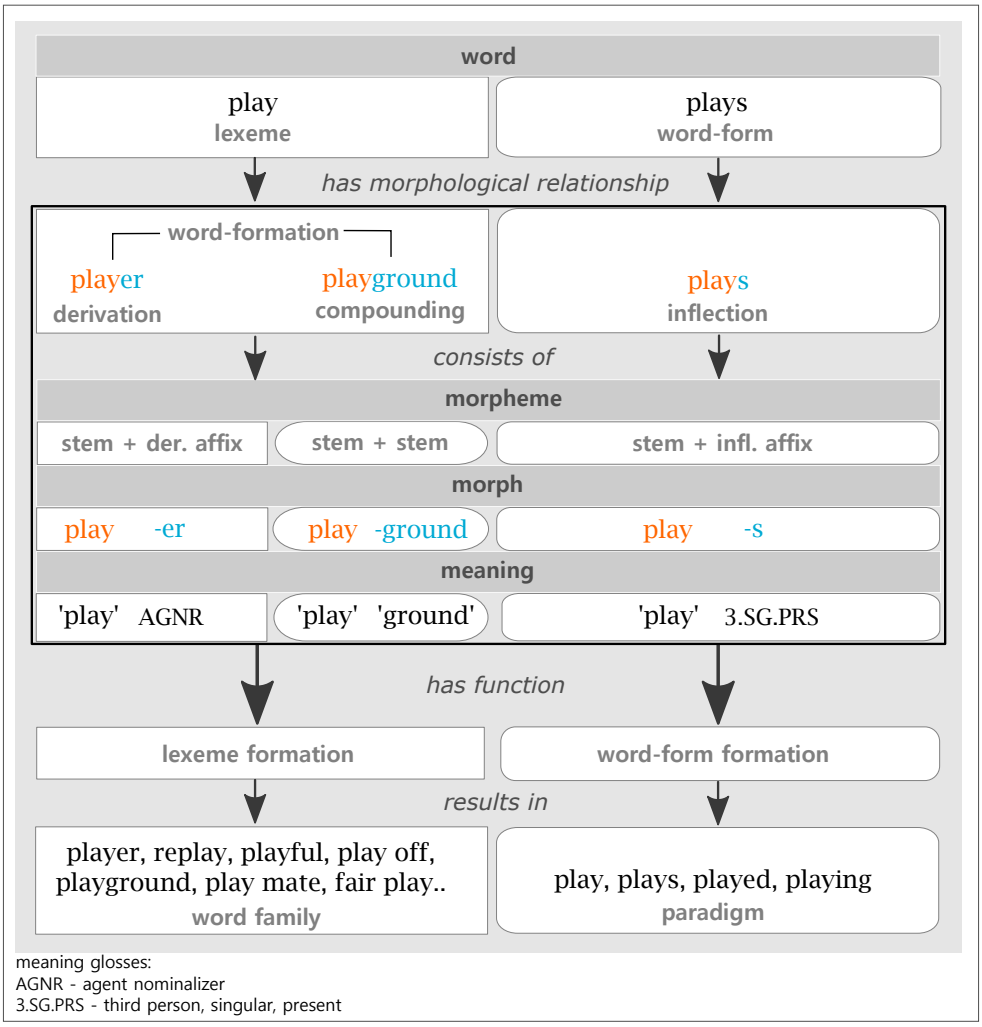


Figure 1. Overview of the linguistic domain of morphology with the English example lexeme *play* (verb).

4. Domain analysis

The development of MMoOn Core is based on the following domain analysis for MLD. It has been conducted in order to clarify and decide which linguistic elements and relations need to be represented. The linguistic domain of morphology deals with the internal structure of words including the elements and meanings of which they consist, i.e. the morphs and morphemes of a language. In the context of MMoOn Core we define the term *morpheme* as the smallest component of a word that contributes some sort of meaning, or a grammatical function to the word to which it belongs, whereas the term *morph* is defined as the perceivable side, i.e. the written or spoken realization, of a single morpheme. Just as other linguistic domains, e.g. syntax or phonology, the study of morphology can either refer to that part of language in general or to the morphological system of a specific language. For the purpose of outlining the domain this section is concerned with the first sense of *morphology*, although, the second meaning plays a crucial role when it comes to the description and investigation of the MLD of a specific language.

Figure 1 gives a basic overview of the conceptualization of the domain. It depicts a condensed summary based on linguistic works that outline the area and study of morphology in a general way [4, 24] and which can be assumed to portray the common agree-

[32] http://research.ics.aalto.fi/events/morphochallenge/
[33] http://www.cis.hut.fi/projects/morpho/

ment among linguists as to what elements and relations are part of morphology. The word level is divided into **lexemes** and **word forms**. The former are abstract words which contain a core meaning and are usually listed as entries in dictionaries. The latter are concrete realizations of a lexeme which combine the lexical core meaning with additional grammatical meanings that are relevant for their embedding in a syntactic environment. Lexemes and word forms can enter two morphological relationships, i.e. **word formation** and **inflection**, respectively. Word formation can be further divided into **derivation** and **compounding**. These terms address the morphological components of which they can consist. The major part of morphology is then devoted to "the study of the systematic covariation in the form and meaning of words that can be identified by segmentation" [24]. For the English example of the verb *play* it is shown in Figure 1 that these segments can be divided into free and bound realizations, i.e. **stems** and **affixes**. Stems are morphs that can usually stand alone whereas affixes are always attached to a stem. The two lexemes *player* and *playground* and the word form *plays* all contain the lexical stem *play*. The difference between these three types of words lies in their morphological building patterns. Derived lexemes consist of a stem and a **derivational affix**, which is in this example the suffix *-er* that encodes the meaning of 'agent noun' and also entails a word-class change from verb to noun. The morph *-er* is very productive in English and can be used to form a variety of agent nouns from verbs, e.g. *winner* (noun) from *win* (verb) or *writer* (noun) from *write* (verb). Compound lexemes, in contrast, consist of two stems, i.e. *play* and *ground* in the given example. Both processes of word formation have the function to form new lexemes, by extending the meaning of a lexeme with additional meaningful elements. As a result, **word families** of lexemes emerge which contain all lexemes that share the same lexical core meaning. Accordingly, all lexemes of the word family *play* in Figure 1 are derivatives or compounds encoding some extended but related lexical meaning of the verb *play*.

In contrast to word formation, inflection does not result in new lexemes. Rather, it involves the morphological modification of a lexeme in order to use the word form of it in a certain syntactic environment. Consequently, word forms consist of a lexical stem and an **inflectional affix**. In the example *plays* is a word form of the lexeme *play* and consists of the stem *play* and the suffix *-s* which encodes 'third person', 'singular' and 'present tense'. Thus, the process of inflection has the function to build word forms of a lexeme. This results in **paradigms** that contain all word forms that can be build from one lexeme. Usually, an inflectional paradigm is a cross-classification according to the grammatical features involved. These are often linguistic categories such as person, number and tense in inflectional languages. Since English marks only the word forms encoding the third person, singular and present tense with the suffix *-s*, the paradigm is not very extensive and encompasses only four word forms. Similarly to the derivational affixes, the inflectional affixes occur in other word forms with the same (grammatical) meaning.

Overall, the domain of morphology is mainly concerned with the identification of the smallest meaning bearing units of language and the investigation of their concrete realization, meaning, function, relation to each other and the systematization of the underlying building (ir)regularities.

5. MMoOn Core - The Multilingual Morpheme Ontology

Everything developed by us around MMoOn Core can be accessed under the following websites: http://mmoon.org/ and https://github.com/MMoOn-Project. The ontology is published under http://mmoon.org/core.rdf and open for any kind of reuse under a CC BY 4.0 license. Altogether, the MMoOn Core model comes with 430 classes, 37 object properties, five datatype properties and 301 instances which have been all created manually. An overview of the model is given in Figure 2 that illustrates the eight main classes and their division into further subclasses. As will be shown in the following subsections, the seemingly large setup of MMoOn Core is well structured and can be used from a reduced extent up to its full possibilities, which will enable a sufficient description of MLD according to the conducted domain analysis.

5.1. MMoOn Core basic elements

In the following an overview of the eight main classes and central properties provided in MMoOn Core will be given. Due to the size of the ontology vocabulary it is recommended to additionally consult the ontology file to receive more detailed insights into the definitions and interrelations established between the ontology elements.

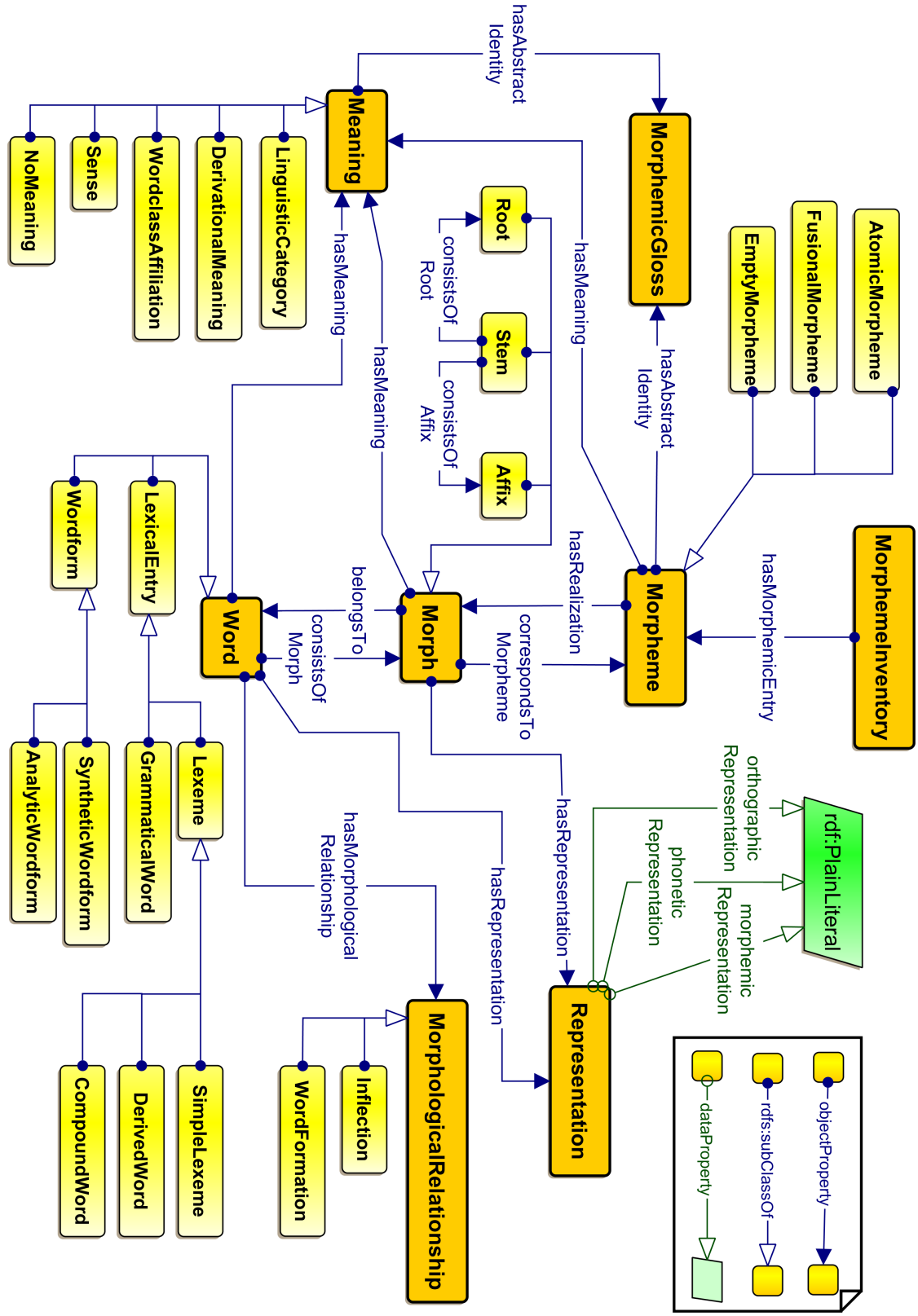


Figure 2.: Overview of the MMoOn Core main classes and properties.

5.1.1. Main classes

MorphemeInventory: Each compilation of morphemic data with MMoOn Core will result in a morpheme inventory that is specified for the language of the data by using the object property `mmoon:forLanguage`. Every MMoOn language inventory should be named according to its given lexvo ISO language code and is an instance of `mmoon:MorphemeInventory`. Since MMoOn shall describe morphemes, each morpheme inventory consists of `mmoon:Morpheme` and/or `mmoon:Morph` resources.

Word: The word is the basic constituent at the phrase level and unit of morphological analysis. MMoOn Core further subdivides this class into `mmoon:LexicalEntry` and `mmoon:Wordform`, which both consist of further subclasses (cf. Figure 2). The `mmoon:Word` class serves as a very broadly defined superclass subsuming everything that consists of a contiguous sequence of letters or phonemes. In this sense both `mmoon:LexicalEntry` and `mmoon:Wordform` are subclasses of `mmoon:Word` and differ in that the former class instances represent abstract words and the latter class instances represent concrete words. Instances of `mmoon:LexicalEntry` are, therefore, words as they appear as entries in a dictionary. The two subclasses `mmoon:Lexeme` and `mmoon:GrammaticalWord` distinguish between lexical entries that have a lexical or a grammatical meaning. The instances of the class `mmoon:Wordform` are inflectional variants of `mmoon:LexicalEntry` instances and represent words as they are used in text or speech [24]. The classification of words in MMoOn Core is more fine-grained than in vocabularies modeling lexical language data. It mainly serves to distinguish words according to their morphological formation. In particular this entails that morphs occurring in `mmoon:LexicalEntry` instances are morphs that are involved in word formation processes and morphs occurring in `mmoon:Wordform` instances are part of word form formation processes.

In order to allow for an easy extension of an existing lexical dataset with morphological data, `mmoon:LexicalEntry` is interconnected with the `ontolex:LexicalEntry` class via the `rdfs:subClassOf` property and with `gold:LexicalItem` via `skos:broadMatch`.

MorphologicalRelationship: This class serves as a means to specify the relationship between word forms of a lexical entry (inflection) or the relationship between lexical entries of a word family (derivation and compounding). Accordingly, the two subclasses `mmoon:Inflection` and `mmoon:WordFormation` are established. Several subclasses for both of them are also provided, e.g. the class `mmoon:Declension` that can be used to document nominal inflectional paradigms as they are provided in inflection tables. All word forms that are included in such a table can be then associated with its respective declension class, for instance a Latin noun belonging to the first declension paradigm. Similarly, the two classes `mmoon:Derivation` and `mmoon:Compounding`, being subclasses of `mmoon:WordFormation`, provide more specific subclasses that are ready to use. The derived word *smallish*, for instance, is a lexeme that can be specified for the derivational relation `mmoon:DeadjectivalAdjective`. This allows for a morphological classification of the words of a language which is usually described in the grammatical sections of language descriptions discussing inflectional paradigms and word families. In this regard, however, the MMoOn Core `mmoon:MorphologicalRelationship` subclasses are primarily designed to cover an extensional representation of inflection and derivation classes by listing `mmoon:Lexeme` and `mmoon:Wordform` instances which are interconnected with the `mmoon:hasWordform` or `mmoon:isDerivedFrom` object properties and point to the same `mmoon:MorphologicalRelationship` instance. An intensional usage of the `mmoon:MorphologicalRelationship` class is also possible, however, not in an explicit machine-processable manner (as provided in LMF, for instance). Morphological patterns that subsume inflected or derived forms sharing the same transformation processes for inflection or word formation can be only described with `rdfs:comment` or a similar annotation property. The reason for this is the inability to explicitly specify a `mmoon:MorphologicalRelationship` class or instance for grammatical or derivational categories contained in the `mmoon:Meaning` class. Additionally, a specific object property that would allow to interconnect `mmoon:Lexeme` or `mmoon:Wordform` instances with each other as prototypical references to the shared morphological patterns would have to be created. In this respect, the generation of word forms and lexemes based on explicitly defined morphological patterns from within an ontology is regarded out of scope of MMoOn Core which – being an ontology – is re-

garded as a means to the describe and not generate MLD[34].

Moreover, with the two classes `mmoon:NoInflection` and `mmoon:NoWordFormation` words that exhibit an inability to undergo certain morphological processes can be explicitly represented.

`Morph`: The morph resources are concrete realizations of a single morpheme which usually result from segmentation. In the MMoOn Core vocabulary they are the manifestations of the form side of a linguistic sign and as such constitute perceivable elements in the form of graphemes or phonemes. Therefore, a `mmoon:Morph` has a corresponding `mmoon:Morpheme` (see below) and together both form one linguistic sign based on a one-to-one correspondence between form and meaning. Several subclasses enable the specification of the morph type, e.g. `mmoon:Affix`, `mmoon:Stem` and `mmoon:Root`. Again, the MMoOn Core vocabulary provides here a more fine-grained classification. Especially the affix subclasses `mmoon:Simulfix`, `mmoon:Transfix`, `mmoon:EmptyMorph` and `mmoon:ZeroMorph` constitute a valuable addition next to the commonly provided prefix, suffix, infix and circumfix classes that exist already in other vocabularies, e.g. GOLD, OLiA or OntoLex-*lemon*, but also in MMoOn Core as well. What is unique to MMoOn in addition to these classes, is the possibility to interrelate morph instances with the `mmoon:isAllopmorphTo` and `mmoon:isHomonymTo` object properties.

`Morpheme`: The morpheme class contains the smallest meaning-bearing elements of a language. These comprise all semantically distinct concepts which are encoded by the morph the morpheme realizes, i.e. the morpheme resources are manifestations of the inseparable meaningful side of corresponding morphs in a language. These meanings can be lexical meanings, grammatical meanings or senses. Determined by the occurring kind of morph-to-morpheme correspondence, morpheme resources can be further specified for being 1) a `mmoon:AtomicMorpheme`, i.e. the realization by the morph resource entails exactly one meaning, or 2) a `mmoon:FusionalMorpheme`, i.e. more than one meaning is encoded within the morph realizing such a morpheme but these are not separately identifiable by further segmentation, or 3) a `mmoon:EmptyMorpheme`, which is by definition a morpheme that has no meaning but is realized by an empty morph. This class has been established to explicitly capture the non-existing meaning correspondence of `mmoon:EmptyMorph` instances and the statement `mmoon:EmptyMorpheme mmoon:hasRealization mmoon:EmptyMorph` is already provided with the vocabulary for convenience.

`Meaning`: The `mmoon:Meaning` class is the largest class in MMoOn Core. It comprises meanings a word, morph or morpheme can be associated with, e.g. `mmoon:LinguisticCategory`, `mmoon:DerivationalMeaning` or `mmoon:WordclassAffiliation`. Since the domain of MLD is concerned with meanings, MMoOn Core aims at providing already a wide range of meanings that are attested among many of the world's languages. With the advanced usage of the vocabulary it is planned to extend it with meanings that are currently not available in MMoOn Core at the moment but will be necessary for dataset creators of specific languages. The linguistic categories are collected from three different sources, i.e. the OLiA ontology, the GOLD ontology and the LiDo Glossary of Linguistic Terms database[35]. They contain usually obligatory expressed linguistic features such as person, number, tense and case, but also clusivity, relative person or social deixis. In contrast, MMoOn Core is the first vocabulary that also provides and collects derivational meanings which are useful to represent word formation processes. These include, for instance, diminution, inhabitant, aktionsart or applicative. The modeling of word classes as a type of meaning might seem unusual but follows the narrow purpose to provide the possibility to express conversion which is also called zero-derivation. Conversion is regarded as the formation of a lexeme from a lexeme with another part of speech which contains no further derivational meaning except that which is entailed in the word class change, e.g. the noun *call* derived from the verb *(to) call*. Further, for describing the meanings of lexemes, stems and roots the `mmoon:Sense` class can be used. Providing sense resources here, however, exceeds the domain scope of MMoOn Core. Thus, senses must be defined based on existing data or can point to an appropriate external sense resources, e.g. synsets from WordNet RDF, by using

[34] Efforts to achieve this goal are currently under development within the OntoLex-*lemon* morphology module [35].

[35] http://linguistik.uni-regensburg.de:8080/lido/Lido

the mmoon:senseLink object property. Finally, the class mmoon:NoMeaning is established to explicitly state that an empty morph has no meaning.

MorphemicGloss: The morphemic gloss is the abstract identity of a morpheme and serves as a metalinguistic representation of meanings. MMoOn Core already contains 300 instances of morphemic glosses, most of which are taken from the Leipzig Glossing Rules [13] or from Lehmann's glossing list [37]. Furthermore, for each mmoon:Meaning class and every instance that will have a type assertion to one of these classes, glosses are established that are interrelated to the meanings, e.g. mmoon:Singular mmoon:hasAbstractIdentity mmoon:MorphemicGloss_SG. The glosses can be also used to represent mmoon:Morpheme resources, e.g. the English morpheme for 'third person', 'singular' and 'present tense' is represented as eng_inv:FusionalMorpheme_3P_SG_PRS mmoon:hasAbstractIdentity mmoon:MorphemicGloss_3P, mmoon:MorphemicGloss_SG, mmoon:MorphemicGloss_PRS. The provision of morphemic glosses and their association to meanings in MMoOn Core fulfills the following three objectives. First, the existence of gloss instances facilitates the data compilation and saves the time for creating glosses. Second, consistency of glosses among different MMoOn morpheme inventory datasets is guaranteed because of a shared set of preassigned glosses. Nonetheless, if necessary or desired, new glosses can be created as well but should be linked via owl:sameAs to the existing MMoOn Core gloss. Finally, the glosses enable a cross-linguistic comparison of how specific meanings are morphologically encoded across different languages.

Representation: In this class the linguistic representations of mmoon:Morph and mmoon:Word resources are collected as abstract representation instances, e.g. eng_inv:Suffix_er mmoon:hasRepresentation eng_inv:Rep_er. These instances can be further specified for their string realization with the four different datatype properties mmoon:orthographic-, phonetic- and morphemicRepresentation as well as mmoon:transliteration. Morphemic representation literals include the marking of the morph boundary according to the defined typographic conventions of mmoon:morphemicRepresentation that demarcate them from plain orthographic representations, e.g. the instance eng_inv:Rep_er points to the morphemic representation literal "-er"@en. For the reason of consistency the morphemic representation for the mmoon:ZeroMorph instances, which have by definition no phonological and orthographic representation, has been already established, i.e. mmoon:Representation_ZM mmoon:morphemicRepresentation "-Z"^^xsd:string. Together with the mmoon:Meaning resources the mmoon:Representation data enables the identification and explication (cf. Section 5.1.2) of allomorphs (two morphs that link to the same meaning but to different representations) and homonymous morphs (two morphs that link to the same representation but to different meanings) within a dataset.

As this overview of the eight main classes shows, the class hierarchies in MMoOn Core are very elaborate. Irrespective of the level of granularity of the source data both the very specific subclasses and the more general superclasses enable the representation, identification and classification of the linguistic elements that are involved in the domain of MLD.

5.1.2. Properties

A key feature of modeling the domain of MLD constitutes a sufficient set of relations that is able to capture the segmentation of words. Altogether, the MMoOn Core vocabulary provides 37 object properties which can be used to state more or less specific relations for modeling the morphemic elements of the data that should be represented. Figure 3 illustrates a part of the example data that has been introduced in Figure 1 by using the most specific properties, i.e. the subproperties which are lowest within the hierarchy of an object property.

In practice, datasets containing morphological data highly differ in terms of coverage and granularity. As a result, the variety of the created object properties emerged because of the intention to increase the applicability of the MMoOn Core vocabulary to as many differing kinds of morphological datasets as possible. This aspect is not trivial, since morphological data does not exist to the same extent as lexical language data and ranges from simple tables containing lexemes, stems and affixes over texts with interlinear morphemic glosses to morphological segmentation tool outputs. In what follows, it will be first outlined how morphological data is ideally expressed with the MMoOn Core vocabulary and second, further possi-

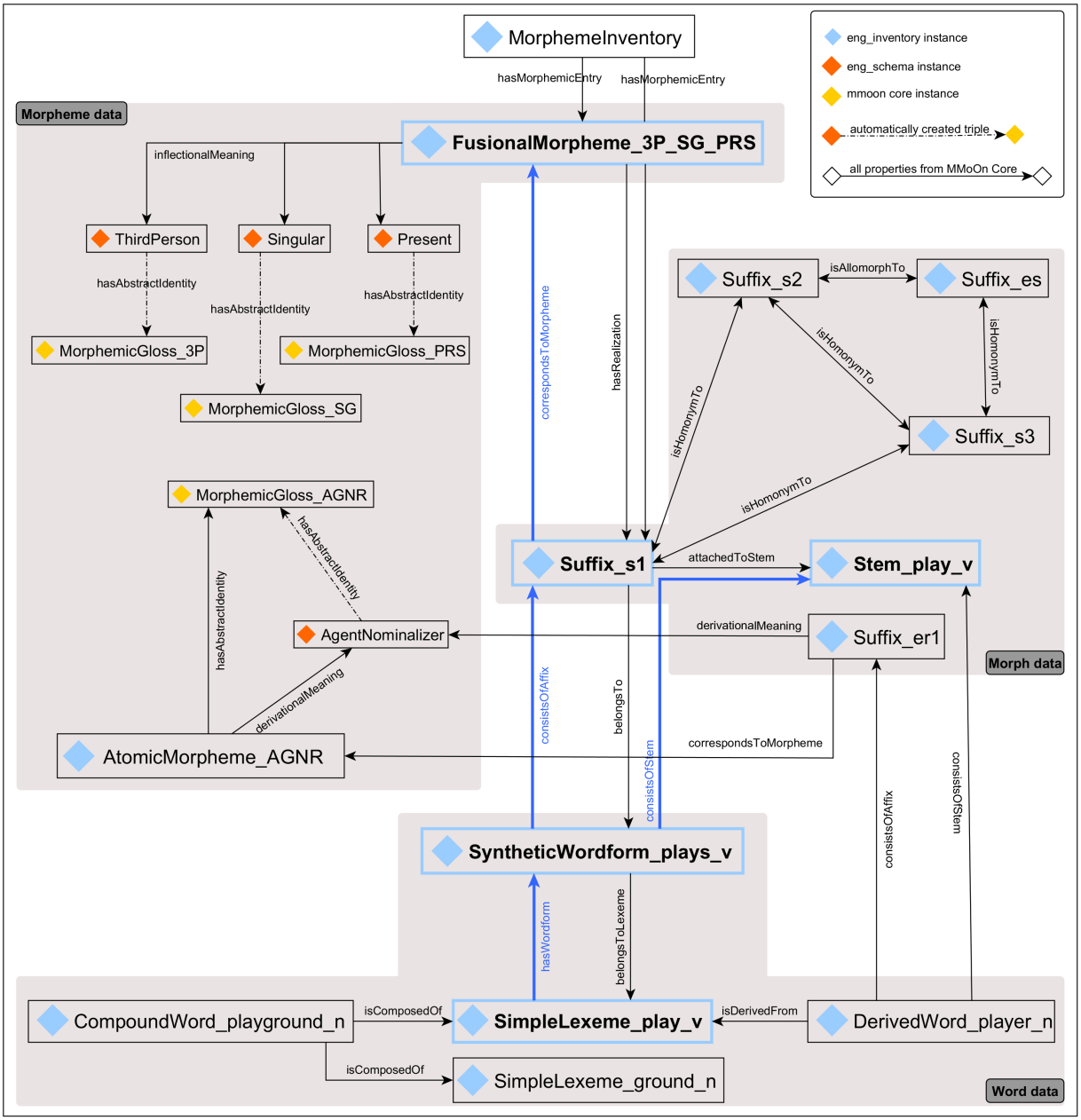


Figure 3. Modeling of relations between morphological data with the example segmentation of the word form *plays*.

bilities for deviating data representations will be motivated.

An ideal MMoOn-based dataset contains instances of the three main classes `mmoon:Word`, `mmoon:Morph` and `mmoon:Morpheme`. They are interrelated according to the part of the graph that is highlighted in blue. This case is exemplified for the word form *plays* in Figure 3. It is classified as a `mmoon:SyntheticWordform` which can be segmented into the two `mmoon:Morph` instances `Stem_play_v` and `Suffix_s1`. These in turn are interconnected with their corresponding `mmoon:Morpheme` instances, in this example `Suffix_s1` with `FusionalMorpheme_3P_SG_PRS`. This modeling is chosen because it enables an explicit distinction between the form and meaning side of

subword elements. It resolves a prevalent ambiguity that exists in the discourse about the morphology domain when, for example, speaking of "the third person, singular, present *-s* morpheme". Therefore, within the MMoOn vocabulary the *-s* is referred to as a `mmoon:Morph` instance and the "third person, singular, present" as a `mmoon:Morpheme` instance. Since the form and meaning sides of linguistic signs are inseparable, both resources are interrelated with the `mmoon:correspondsToMorpheme` object property and its inverse property `mmoon:hasRealization`.

The grey areas in Figure 3 illustrate how the instances of the three main classes in this ideal modeling can be further described to represent word, morph and morpheme data.

On the word level the interrelation between different types of words can be stated. Word form resources are always interconnected to lexemes by using the property `mmoon:belongsToLexeme` which is inverse of the property `mmoon:hasWordform` as exemplified for the instance `SyntheticWordform_plays_v`. Further, an assignment to an inflectional paradigm can be stated. The property `mmoon:inflectionalRelation` is used to express which verbal inflection class applies, similar to inflection tables in dictionaries. In the given example the following statement can be realized:

```
SyntheticWordform_plays_v
mmoon:inflectionalRelation
ex:regularConjugation.
```

The segmentation of word forms into morphs consists only of stem or root and inflectional morph segments. Derivation and compounding relations are expressed between `mmoon:LexicalEntry` resources. This can be done by using the object properties `mmoon:isDerivedFrom` and `mmoon:isComposedOf` as is illustrated for the derived word *player* and the compound word *playground* in Figure 3. Similar to the declaration of an inflectional relation for verbal word forms, a derivational and compounding relation can be also stated for derived and compound words, e.g.:

```
DerivedWord_player_n
mmoon:derivationalRelation
ex:agentNoun.
CompoundWord_playground_n
mmoon:compoundingRelation
ex:nominalCompound.
```

The segmentation into derivational affixes takes place on the lexeme level. Therefore, in Figure 3 the derivational morph `Suffix_er1` is interconnected with the derived word `DerivedWord_player_n` and would not be part of the segmentation of the word form instances belonging to this lexeme. This outlined ideal usage of the MMoOn Core vocabulary on the word level takes up the split-morphology hypothesis [43]. This modeling choice renders an explicit declaration of a morph expressing either an inflectional or derivational meaning unnecessary, since the derivational segmentation operates pre-syntactically to form new lexemes and the inflectional segmentation operates post-syntactically providing the grammatical features to yield a word form.

On the morph level the `mmoon:Morph` instances as the perceivable side of the morphemes are represented as strings via the three datatype properties `mmoon:phoneticRepresentation`, `mmoon:othographicRepresentation` and `mmoon:morphemicRepresentation` for rendering phoneme, grapheme and morphemic representations. The latter consists of a morphemic boundary marking and the conventional orthographic representation of it, e.g.: `Rep_Suffix_s1 mmoon:morphemicRepresentation "-s"@de`. It is further possible to interrelate affixes with stems or roots by using the superproperties `mmoon:attachedTo` and `mmoon:consistsOfMorph` or their more specific subproperties, e.g.:

```
Suffix_s1
mmoon:attachedToStem
Stem_play_v.
Stem_player_n
mmoon:consistsOfAffix
Suffix_er1 ;
mmoon:consistsOfRoot
Root_play.
```

The introduced one-to-one correspondence between morphs and morphemes enables the identification of allomorphs and homonymous morphs in the data. All `mmoon:Morph` instances that correspond to the same `mmoon:Morpheme` instance but not the same representation can be, therefore, interrelated with the object property `mmoon:isAllomorphTo`. Conversely, all `mmoon:Morph` instances that point to the same representation but to different corresponding `mmoon:Morpheme` instances are interrelated with the object property `mmoon:isHomonymTo`. Both properties are symmetric so that this interconnection need to be stated only for one morph. In Figure 3 both cases are exemplified by the instances `Suffix_s2`, `Suffix_s3` and `Suffix_es` given that the first and second morph correspond to the 'nominal plural'

morpheme and the last to the 'genitive' morpheme. This is not restricted to inflectional morphs but can be also used to express allomorphy between derivational morphs, e.g. for the English adjectival morph corresponding to the 'comparative' morpheme (i.e. `Suffix_er2`):
`Suffix_er1`
`mmoon:isAllomorphTo`
`Suffix_er2`.
Even though this modeling choice requires a numbering of `mmoon:Morph` resources it is taken up because it allows to identify and establish allomorph and homonymous relations within morphemic datasets which often contain information about meanings and representations but lack an explicit declaration of their interrelations.

On the morpheme level the meanings that are encoded by the morphs are assigned to the `mmoon:Morpheme` instances. In accordance to this, `Suffix_s1` corresponds to the fusional morpheme `FusionalMorpheme_3P_SG_PRS` which is further specified with the object property `mmoon:inflectionalMeaning` for consisting of the non-segmentable inflectional meanings `ThirdPerson`, `Singular` and `Present`. This property is a subproperty of `mmoon:hasMeaning` next to other properties that can be used to declare derivational, grammatical, contextual or inherent inflectional meanings and senses. The URI of `mmoon:Morpheme` resources reuses the morphemic glosses that are already interconnected with the meanings within the MMoOn Core ontology. This is done since morphemes are concepts and as such need some kind of representation in order to be referenceable. The abstract identities provided by the morphemic glosses are widely known and, therefore, suitable to serve this purpose. Moreover, since the `mmoon:Morpheme` resources represent concepts only, statements about their perceivable forms, for example their ordering, segmentation or position within a word, are made by means of the corresponding morphs by which they are realized. To this extent, the modeling of the linguistic concepts of 'morph' and 'morpheme' in MMoOn Core formalizes the distinction between signifier and signified which constitute the – usually inseparable – sides of the linguistic sign. By explicitly separating them, information about both – as just illustrated – can be described in detail by avoiding ambiguities at the same time.

However, comprehensive datasets containing resources that are involved in the blue graph just explained are rather an exception. Especially the `mmoon:Morpheme` instances as defined in MMoOn Core only exist in interlinear glossed text sources. Therefore, the object properties are modeled in a way that allows to represent any fraction of MLD with MMoOn Core. As single requirement, a MMoOn based dataset needs to have at least one morphemic entry, i.e. a `mmoon:Morpheme` or a `mmoon:Morph` resource. Apart from that, one can start representing data from any level. The three inverse object properties `mmoon:hasRealization`, `mmoon:belongsTo` and `mmoon:belongsToLexeme` enable the representation of data in the opposite direction of the blue graph in Figure 3 from the morpheme or morph to the word form and word data. In addition to that, it is necessary that MLD can be modeled independently from the complexity of the data. Especially the possibility to assign meanings not only to the morpheme resources but also to morph and word resources had to be considered carefully. For datasets containing only morphs together with the information of the meanings they encode, `mmoon:Meaning` instances can be also directly explicated. This is illustrated in Figure 3 with the instance `Suffix_er1` which can also be directly associated with the derivational meaning instance `AgentNominalizer` in case the `AtomicMorpheme_AGNR` instance does not exist to declare the morph-to-morpheme correspondence. What can be also seen is that the morphemic gloss instance `mmoon:MorphemicGloss_AGNR` already exists in the MMoOn Core vocabulary and is automatically assigned to the meaning instance `AgentNominalizer` (cf. Section 5.1.1). Since the URIs of the `mmoon:Morpheme` instances are based on the labels of the `mmoon:MorphemicGloss` instances, `mmoon:Morpheme` instance data might be later derived from the established meaning-to-gloss associations that are given for `mmoon:Morph` instances lacking corresponding `mmoon:Morpheme` data. Likewise, meanings can be directly assigned to `mmoon:Word` resources (however, not shown in Figure 3). This might be useful for datasets similar to DBnary that contain only word forms of a lexeme that are annotated with the corresponding grammatical meanings on the word level. Albeit, for this case a fully valid MMoOn dataset can not evolve, because no morph or morpheme resources are contained. It is, however, possible to use the MMoOn vocabulary then as an extension of another vocabulary for lexical data such as OntoLex-*lemon*.

The decision to define not only `mmoon:Morpheme` but also `mmoon:Morph` and `mmoon:Word` as do-

mains of the mmoon:hasMeaning object property compensates for the lack of morpheme data as defined in the MMoOn Core vocabulary. Under the assumption that dataset creators start with the most suitable usage of the ontology according to their source data and make use of a later generation of mmoon:Morpheme resources from the initial MMoOn-RDF data, it can be expected that the dataset is likely to become semantically over-expressive. It might be the case, for example, that a later addition of the instance AtomicMorpheme_AGNR creates two more triples; one that interlinks it with the morph Suffix_-er1 via mmoon:hasRealization and another that interconnects this mmoon:Morpheme instance to AgentNominalizer via the mmoon:hasMeaning property. This leads to a semantic overload but does neither reduce the interoperability nor the quality of a dataset. Overall, the heterogeneity of existing non-RDF morphological data representations had to be taken into account. Therefore, this modeling option is regarded as a reasonable compromise to enable a less constrained data modeling which can in turn serve as a basis to arrive at the intended usage of MMoOn Core due to the possibility to create mmoon:Morpheme resources from mmoon:Meaning data. The alternative would have been to restrict the usage of mmoon:hasMeaning to mmoon:Morpheme instances and to accept a largely reduced applicability of the vocabulary and, consequently, less morphemic datasets in RDF.

The presented overview of the MMoOn Core object properties illustrated the possibilities of their usage for representing MLD of different complexity and coverage. On this basis MLD (if newly created) can be modeled according to the ideal graph just exemplified or (if covering only a part of the domain data) extended later on to include more fine-grained MLD. It shall be noted that datasets containing morpheme, morph, word form and lexeme resources that are interconnected in the most granular way will allow to derive the greatest insights into the morphological elements and structures of a specific language that is represented with the MMoOn Core vocabulary.

5.2. Architectural setup of MMoOn morpheme inventories

Given the complexity of the MMoOn Core ontology the question arises how language-specific MMoOn morpheme inventories are meant to be built. Therefore, an integrational architectural setup (cf. Figure 4) has been developed which interconnects the language data of each morpheme inventory with MMoOn Core and, thus, ensures the multilinguality of all MMoOn datasets. The architectural setup comprises three data layers that serve to cover the following two aspects of linguistic data, i.e. 1) the difference between primary and secondary language data and 2) their description by assuming either language-independent or language-specific linguistic categories. The first aspect is based on the general assumption that most linguistic datasets comprise primary as well as secondary language data [36]. The former data type is defined here as language data which originates from a certain text compilation or could be applied to any text or token in order to identify the word forms, morphs and morphemes of the morpheme inventory. The latter is then defined as the kind of data which enables the description of the primary language data. E.g. the German plural suffix deu_inventory:Suffix_er1 is a primary language data instance which is specified with the secondary language data instance deu_-schema:Plural for its grammatical meaning. The second aspect is then concerned with the assignment of both instances to language-independent or language-specific categories. In this respect, linguistic categories like suffix or plural tend to be modeled as language-independent concepts, even though, in practice they are used in the context of describing the data of a specific language and consequently then carry a more specific meaning.

In what follows, the three data layers of MMoOn morpheme inventories will be described and how they allow to model primary and secondary language data simultaneously in the context of a language-independent data model that subsumes and interrelates language-specific data.

The first layer builds the MMoOn Core ontology as the underlying formal and conceptual model shared by all morpheme inventories. Since it models the domain of morphology as a subfield of the study of language it functions at the **language-independent schema level** describing the domain of morphology in a general way. It aims at providing the starting point for creating language-specific models of the morphology of a certain language based on unifying and comparable generic concepts. In that respect, it can be seen in Figure 4 that the eight main classes are divided into four classes for the representation of secondary language data which can be directly applied to describe the primary language data that is represented by means of the other four main classes. This modeling satisfies

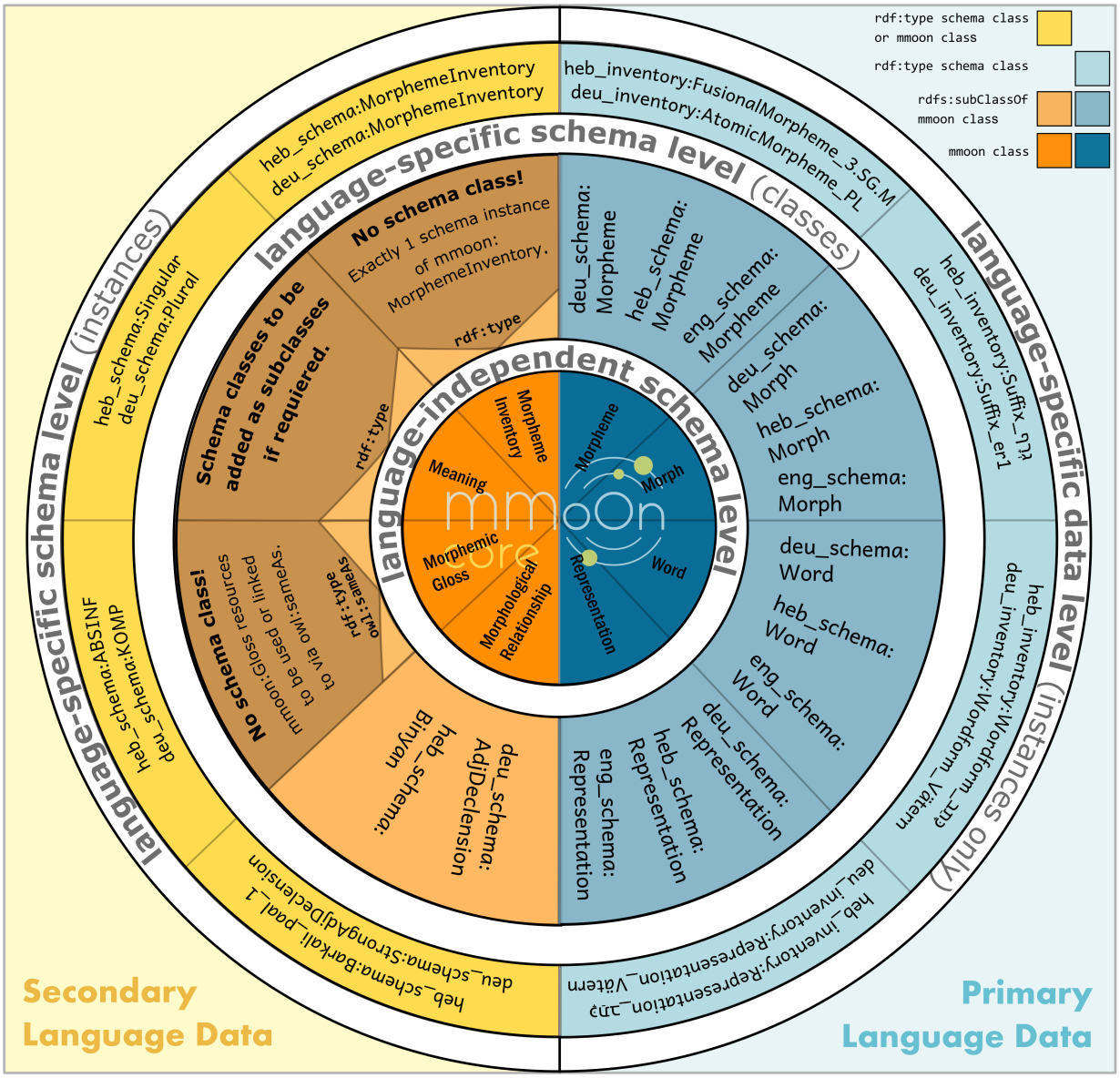


Figure 4. Architectural setup of MMoOn morpheme inventories exemplified with morphological German and Hebrew data.

the practical implication that primary language data is rarely collected on its own but most often accompanied with respective secondary language data that needs to be specified as well. Especially the provision of the numerous fine-grained grammatical and derivational meanings facilitates, thus, the creation of a morphological dataset because it reduces the time which is usually required to search for the necessary linguistic meanings in other external vocabularies.

The second layer in the architectural setup builds the **language-specific schema level** (i.e. the entire middle and outer left circle) being exemplified for a German and Hebrew morpheme inventory in Figure 3. This level is meant to provide the formalized schematic vocabulary which enables a description of the general linguistic concepts provided in MMoOn Core in compliance to their actual language-specific realization. Consequently, the domain of morphology on this level is modeled as the descriptive linguistic part of a certain language. In practice this layer is realized by a language-specific ontology that imports the MMoOn Core ontology and contains language-

specific extensions via added subclasses and instances. These include subclasses of all four MMoOn Core main classes (and their subclasses) representing primary language data as well as subclass extensions of the `mmmon:MorphologicalRelationship` class. Morphemic glosses, however, are not meant to be created but preferably reused from the MMoOn Core vocabulary to ensure consistency across multiple MMoOn-based datasets. On this level MMoOn Core class `mmoon:MorphemeInventory` is populated with only one instance specifying the language of the morpheme inventory according to the dataset it contains. Moreover, the language-specific variants of the MMoOn Core `mmoon:Meaning` class are realized as instances. Assuming that MMoOn Core does by far not cover all grammatical and derivational meanings that exist across the languages of the world, missing meanings can be added by creating a new (sub)class. As a result, the deu_schema ontology and the heb_schema ontology are derived as extensions of MMoOn Core by creating appropriate subclasses and instances, e.g. `deu_schema:Word rdfs:subClassOf mmoon:Word` or `deu_schema:Plural rdf:type mmoon:Plural`. Similarly, necessary but missing relations can be added by creating new object or datatype properties. However, it is assumed that the properties already provided by the MMoOn Core ontology will be sufficient for representing most of the existing morphological data and can be, therefore, directly used. This language-specific ontology as an extension of the MMoOn Core model serves the purpose to enable the definition of the linguistic elements according to their language-internal peculiarities by being interconnected with a higher cross-linguistic meta layer at the same time. In order to facilitate the creation of MMoOn morpheme inventories a schema template file that contains the MMoOn Core import as well es the described class extensions is available for immediate reuse[36]. Further, an advantage of the language-specific ontologies is that they can be directly reused by other researchers who have morphemic language data on the same language and would like to contribute their dataset as a MMoOn morpheme inventory as well. An example for this is the Bantu Language Model, a schema ontology for the whole language family of the Bantu languages[37], which served to create the Xhosa, Kalanga and Ndebele morpheme inventories.

The largest part of each dataset constitutes the primary language data. Within the architectural setup it is realized by instances on the **language-specific data level** (i.e. the outer blue circle in Figure 4). Given that the primary language data is formally described by the secondary (but language-specific) language data, the former usually takes the subject position while the latter takes the object position within a RDF statement. Further, language-specific data instances can also take the object position, whenever primary language data is interrelated, e.g. when two suffixes are explicated to be allomorphs to each other.

In sum, the aim of this architectural setup is to create a unified multilingual data graph of all MMoOn morpheme inventories to come. The presented layers correspond in practice to three RDF files, i.e. mmoon.ttl, schema.ttl and inventory.ttl[38]. Even though the creation process of a MLD dataset as outlined with MMoOn seems more complex or even tedious, we like to encourage data set creators to adhere to the creation of MLD according to the design of the architectural setup of MMoOn Core-based morpheme inventories because it directly impacts the following four indirect outcomes:
1) Facilitated multilingual Linked Data usage: Due to the unifying function of the MMoOn Core model language-specific instance data of different languages can be cross-linguistically traversed through a single data graph.
2) Exploitation of linguistic data in NLP tasks for linguistics and vice versa: The rather flat structured language data NLP systems rely on could be supported and extended by also taking fine-grained linguistic data into account to arrive at more stable data-driven approaches. Conversely, empirical linguistic research benefits from vast amounts of language data that can be collected in a structured way with NLP methods, which in turn, can serve as a starting point to create more accurate and interrelated linguistic datasets[39].

[36] The template file can be downloaded here: https://github.com/MMoOn-Project/MMoOn/blob/master/schema_template.ttl and only needs to be specified for the language of the morpheme inventory.

[37] https://github.com/MMoOn-Project/OpenBantu/blob/master/bnt/schema/bantulm.ttl

[38] Usually the schema and inventory files are specified for the language of the morpheme inventory, e.g. deu_schema, heb_schema and deu_inventory, heb_inventory in Figure 3.

[39] For more details on how the MMoOn dataset creation setup is involved here, cf. Section 7.2

3) Enable onomasiological and semasiological data retrieval: Most linguistic datasets only allow for unidirectional data retrieval. A MMoOn morpheme inventory, however, is more flexible in this respect. Because it provides the means to represent the association of a linguistic meaning with its language-specific expression within the same model, the meanings a certain morph or word form encodes as well as the kind of morphemic expressions that are used to encode a certain meaning can be retrieved simultaneously.
4) Development of a meta-collection of linguistic concepts: Every MMoOn Core based language-specific schema ontology automatically adds to the extension of the MMoOn Core `mmoon:Meaning` class and its subclasses. E.g. the generic meaning of the language-independent `mmoon:Singular` class is extended by all language-specific `Singular` instances. At the same time, additional and newly created linguistic concepts that appear in the schema ontologies indicate missing language-independent MMoOn Core concepts which will be regularly complemented. In this respect the MMoOn Core ontology is under constant development. As a result, the MMoOn Core ontology will evolve to a kind of meta-collection for linguistic concepts that also comprises and interconnects their language-specific realizations. To the best of the authors' knowledge another ontology offering such an explicit distinction for representing language-independent and language-specific linguistic concepts does not exist.

5.3. Domain requirements and design principles

The creation of a domain ontology is guided by several influencing aspects ranging from the granularity of the domain representation, the intended usage of the resulting datasets and possible user groups to the choice of the vocabulary as well as the technical possibilities and limitations of the data format. Thus, modeling the MMoOn Core ontology entailed several design decisions. In order to comprehend the motivations that accompanied the development of MMoOn Core, the design principles and determining domain requirements will be outlined in what follows.

5.3.1. Design principles for the domain of MLD

Domain delimitation: The elements and relations of the domain of MLD in MMoOn Core are based on the domain analysis as outlined in Section 4. However, some of the included linguistic elements such as lexemes, word forms and morphs overlap with other linguistic domains, e.g. lexicography, phonology and syntax. Study areas like morphophonology and morphosyntax indicate that basic linguistic concepts are considered to be part of several linguistic domains depending on their defined characteristics and functions. As a consequence, the domain of morphology can be either described in a very strict way, ignoring possible domain interrelations or in a broader way which would result in an overlap with other domains. The MMoOn Core model takes up the strict approach and, thus, provides anything that is necessary to describe words and the meaningful segmentable subword elements of which they consist. Accordingly, the mentioned overlapping elements are not further specified for postulated functions and usages in other linguistic domains. In that respect, the model strives to be as detailed as possible (on a language-comparative level) and as broad as necessary at the same time. Therefore, MMoOn Core constitutes a quite narrow and fine-grained vocabulary for the domain of MLD but also provides prominent classes, such as `mmoon:LexicalEntry`, that appear across various linguistic domains and can be used as interlinking or alignment points. Furthermore, explicit cross-domain information can be also added by directly linking resources of a MMoOn morpheme inventory to an already existing dataset providing the necessary phonological or syntactic domain information for the same language. The reuse and interlinking to vocabularies describing other linguistic domains is recommended whenever possible.

Framework neutrality: Even though no model comes without any predisposition, MMoOn Core aims at completeness and a comprehensive application rather than fitting the descriptive needs of a certain linguistic framework, model or theory of morphology. It is a first proposal of modeling MLD comprising the relevant categories and relations in order to extend and integrate morphemic data into already existing linguistic datasets which are mainly framework neutral models as well. However, if required, the MMoOn Core vocabulary is easily adjustable so that the data that shall be represented is integrable according to strict theoretical descriptive needs.

Modeling of linguistic concepts and categories: One of the main challenges when it comes to the description of language data is the choice and modeling of the

concepts for linguistic categories. A highly controversial debate exists among the linguistic research community about the treatment of concepts such as 'case', 'gender' or 'noun' as being interlingual comparative or language-specific descriptive categories (cf. for example [23] and [38]). Given that MMoOn Core serves as an upper ontology to create language-specific morpheme inventories both kinds of concepts needed to be considered. Due to the RDF format this particular issue could be solved by adhering to the Semantic Web's standard which already entails the representation of commonality and variability through the hierarchy of classes [1]. In line with this, MMoOn Core classes are regarded as prototypical interlingual concepts and consequently function as the least common denominator for a linguistic category. Every instance of the classes is then a language-specific concept of the upper interlingual MMoOn class concept as described in the setup of the language-specific schema file. According to this, MLD of different MMoOn morpheme inventories can be described with all language-specific features while staying comparable because of the shared MMoOn class membership. As a result, all MMoOn Core based datasets will contribute to a multilingual data graph of interconnected MLD of specific languages.

Coverage: The MMoOn Core model covers concepts and relations that are necessary for synchronic language description, i.e. the representations and meanings of the words, morphs and morphemes are given according to a certain point in time (present or past). Thus, etymological and historical information is not considered in the class or property modeling. As Section 5.1 outlined, MMoOn Core encompasses a fine-grained vocabulary that enables the identification and description of linguistic elements that are necessary for representing MLD. Also, a considerable set of object properties allows for a detailed specification of the relations that hold among the words and the morphemes and morphs of which they consist. As mentioned before, the morphological rules underlying the data are not considered explicitly and need to be inferred indirectly from the data or have to be described by using another vocabulary along with MMoOn Core. The main approach pursued provides granular descriptive means for the morph and morpheme elements and their interrelations to word elements by outsourcing granular phonological, lexicographic or syntactic concepts at the same time. This is not seen as a disadvantage because including them would entail the preference of some theoretical framework which is meant to be avoided.

Target user groups: The use of the MMoOn Core model is directed towards linguists, computational linguists, NLP researchers, lexicographers and anyone who has an interest in compiling and managing MLD. It is anticipated that MMoOn language inventories will be set up by data compilers of the various user groups mentioned. That way synergies can evolve between the smaller but high-quality and mainly manually compiled datasets that are expected from the linguists and the large but not as fine-grained data produced by users with an interest in the machine-processable aspect of linguistic data. The emergence of these cross-disciplinary synergies are assumed to advance the whole LLOD community in general.

5.3.2. Data modeling requirements

Linked Data principles: The choice to model MMoOn Core in the RDF format is motivated by the underlying Linked Data principles [3] which promote the creation of structurally and semantically interoperable datasets. This aspect adheres to the aim of providing a data-unifying domain modeling that is based on technical integrability. Furthermore, due to the creation of unique resources as URIs, the ontology is easily accessible on the Web. Consequently, all emerging MMoOn-based datasets will, therefore, contribute to a growing interconnected data graph and, thus, not join the ranks of the already existing morpheme data silos on the Web.

Reuse: In general it is understood as a good practice to reuse existing vocabularies when creating a new ontology. Since the largest part of the MMoOn Core vocabulary aims at representing meanings, we decided to create a new taxonomy within the `mmoon:Meaning` class and to describe every subclass as a MMoOn Core-specific resource, even though other vocabularies contain similar or the same linguistic meanings and categories as well. By doing so, the assignments of meanings to morphemic elements or words when creating a MMoOn dataset should be facilitated and, moreover, a consistent assignment of morphemic glosses to vocabulary-internal elements could be achieved. Nonetheless, the considerate overlap with other vocabularies for representing language data is ac-

counted for by interrelating mostly `mmoon:Meaning` but also `mmoon:Morph` classes to the highly used GOLD [18] and OLiA [8] ontologies. Classes that are regarded as either equivalent, similar or usable as a defining description for a MMoOn Core class are interrelated via the `owl:equivalentClass`, `rdfs:seeAlso` or `rdfs:isDefinedBy` properties. Furthermore, an alignment with MMoOn Core and the OntoLex-*lemon* model has been established by stating that `mmoon:LexicalEntry` is a subclass of `ontolex:LexicalEntry`. This enables a more specific description of `mmoon:LexicalEntry` resources by using the OntoLex-*lemon* vocabulary for lexicographic information and prevents an overload of the MMoOn Core model by including already existing lexical data.

Extensibility: Finally, a data compilation is rarely ever complete and a single domain model can never capture all practical and theoretical aspects of MLD in general and even less the aspects of MLD of single languages. Given these circumstances, the MMoOn Core model serves as a starting point for morphological data description that might be sufficient for a considerable number of datasets, but must be also prepared to allow for necessary extensions and/or adjustments. This requirement is also assured by the Linked Data format meeting these needs by taking up the assumption of an open world [1]. Consequently, the RDF format allows for a liberate reuse of all classes and properties as well as for an unrestricted extension of the model with new classes and properties. It is, however, assumed that the central comprehensive elements are provided by MMoOn Core and shared by the majority of the emerging MMoOn-based datasets.

URI design: As outlined in Section 5.2 every MMoOn morpheme inventory consists of three files with the MMoOn Core ontology being shared by all datasets. In order to facilitate the identification of and navigation through a dataset, the following URI scheme is implemented for all MMoOn datasets created by the authors: `http://mmoon.org/`**`lang`**`/schema/`**`pi`**`/` for the language-specific schema ontologies and `http://mmoon.org/`**`lang`**`/inventory/`**`pi`**`/` for the language data, where **lang** is replaced by the ISO 639-3 language code and **pi** by an identifier for the project name, e.g. `http://mmoon.org/deu/schema/og/`. For all other dataset creators it is recommended to adhere to the following URI pattern for establishing greater consistency among all MMoOn-based datasets to come: `http://`**`hostname`**`/`**`lang`**`/schema/`**`pi`**`/` and `http://`**`hostname`**`/`**`lang`**`/inventory/`**`pi`**`/`, respectively.

In sum, it appears that the data modeling requirements posed by the morphology domain are very well accomplished by the underlying Linked Data format. The MMoOn Core model as a proposal to start with a homogeneous morphemic data compilation fulfils the needs of a specified linguistic data description model and integrates the resulting data into the Semantic Web environment, thus, benefiting from all of its advantages.

6. MMoOn and OntoLex-*lemon*

In contrast to existing ontologies for describing language data, linguistic datasets rarely contain linguistic information that neatly corresponds to one single linguistic domain. The OntoLex-*lemon* model [41] being a W3C community group specification tackled this issue by covering the domain of lexicology by enabling the representation of related linguistic domains via dedicated submodules. With this modular extensible approach the representation of a wide range of the existing linguistic data can be already realized. Consequently, an all-encompassing vocabulary covering any potential or existing kind of linguistic data point is neither feasible nor desirable. Rather, the development and usage of more fine-grained and specific vocabularies that are interconnected with a commonly shared ontological basis, i.e. OntoLex-*lemon*, will provide the necessary means to enable an appropriate modeling of existing or future linguistic data as Linked Data.

This holds true especially for the domain of MLD, which tends to include lexical as well as morphological data. Depending on the use case and dataset, OntoLex-*lemon*, i.e. the ontolex and decomp submodules in particular, may be used for describing MLD. This has been, for instance, done for representing the components of compound words [16]. Nonetheless, as already mentioned in Section 3.1 for linguistic data corresponding to the domain analysis of MLD (cf. Section 4), the ontolex and decomp modules are mostly limited to compositional morphology and, hence, leave the larger part of the MLD domain to be non-expressible with the provided vocabulary.

A comparative overview based on detailed examples that shows how data on the lexeme, word form, morph

and morpheme levels can be described by using either OntoLex-*lemon* or MMoOn Core can be consulted in Klimek 2017 [33]. Here, a list shall suffice that summarizes the main results, i.e. aspects that reach representability through the MMoOn Core vocabulary and which are not covered in OntoLex-*lemon* respectively:
1) Inflectional affixes: Since inflectional information is usually no central part of lexical data, means to represent inflectional affixes are not part of OntoLex-*lemon*. In fact, even consistently collected number information for nouns by providing the respective morph together with the lexical entry, is not describable with it. Instances that are allowed within the `ontolex:Affix` class are restricted to affixes that form new lexical entries, i.e. derivational affixes. However, a huge part of MLD is comprised by inflectional affixes that are necessary to represent the formation of word forms. The MMoOn Core vocabulary, in contrast, does not distinguish between derivational and inflectional affixes in its assertion being of the type `mmoon:Affix`. Instead, the inflectional or derivational meaning underlying a specific affix is contained in the corresponding morpheme instance as well as its occurrence within a lexical entry or word form, respectively.
2) Stems and roots: Those two elements are crucial for describing MLD, not only for decomposing word forms but also lexical entries. While OntoLex-*lemon* provides the possibility to identify the underlying stems in compound words only (which are not termed as stems but widely included within the class `decomp:Component`), it is not possible to represent the stems or roots of word forms. MMoOn Core provides classes for both elements. Even though the granularity of a segmentation differs from dataset to dataset and depends on the applied linguistic analysis, in many languages root resources are the building blocks of lexical data, e.g. in Arabic languages, and, hence, should be covered as well. As a result, MMoOn allows for the representation of whole inflectional paradigms, including the decomposition into underlying roots, stems and inflectional affixes of the word forms belonging to a specific paradigm.
3) Morphemic interrelations: Part of the description of morphemic elements is also the representation of their relation to other morphs. Therefore, stating the allomorphs and homonyms of a morph is important for their identification, function and the combinatoric rules that apply to them. While the MMoOn Core vocabulary contains two object properties to specify allomorphy and homonymy between morphs, these relations are not part of the lexical domain and, hence, not expressible with OntoLex-*lemon*.
4) Morphemes and meanings: Also not part of the lexical domain is the representation of morphemes. Meanings, i.e. lexical senses in OntoLex-*lemon*, differ largely from the grammatical and derivational meanings that are necessary for describing MLD. The 300 meaning classes provided in MMoOn Core are far from being extensive with regard to the large variety across languages. However, they are a first step towards collecting and documenting meanings that are encoded by morphs and constitute a useful starting point for representing morpheme resources.

As a result of the introduced suggestion to create an interconnection between OntoLex-*lemon* and MMoOn Core in Klimek 2017 [33] both domain ontologies have been aligned, as already mentioned in Section 5.3.2, with the established subclass relation between `mmoon:LexicalEntry` and `ontolex:LexicalEntry`. The two ontologies are intended to be separately usable to describe morphological as well as lexical data in an independent and specific manner by simultaneously maintaining the semantic interconnectivity between all data elements. Consequently, the MMoOn Core model shall not be understood as an OntoLex-*lemon* extension but serves as a standalone vocabulary that can be used in conjunction with OntoLex-*lemon*. Still, the MMoOn Core ontology and its proposed alignment raised awareness within the W3C Ontology-Lexica Community group[40].

As a result, the creators of MMoOn Core have been invited to lead the development of a new OntoLex-*lemon* morphology module which is currently under development[41]. As the interim results for this emerging OntoLex-*lemon* module report [35], the morphology module aims to represent MLD in the context of lexical language data and is not intended to be a vocabulary for the domain of MLD per se. MMoOn Core has built the main orientation basis in the module creation process, however, with the goal to reduce complexity. Especially the morph and its specification of affix types is taken up from MMoOn Core and also the possibility to express inflectional and derivational morphs is now considered. A novelty in the morphology module will be the creation of a means to automat-

[40] https://www.w3.org/community/ontolex/
[41] https://www.w3.org/community/ontolex/wiki/Morphology

ically generate word forms for a lexical entry which is not an integral part of MMoOn Core. In general this module differs from MMoOn Core in that it is more suitable for advanced users of Semantic Web technologies and the Linked Data framework. This is due to the embedding of new vocabulary elements into the existing OntoLex-*lemon* modules and the outsourcing of meanings and glosses by referring to recommended external vocabularies as well as the considerable data preprocessing that is required for the automatic generation of word form data. After all, the data creators, their level of training with Linked Data and their intended usage of the MLD in RDF will influence the choice for MMoOn Core or OntoLex-*lemon* (including the future morphology module) or both models in conjunction. On the whole, it is advisable to start the initial transformation to RDF with the vocabulary that is more expressive with regard to the underlying linguistic domain of the source data, i.e. OntoLex-*lemon* for lexical or MMoOn Core for morphological data.

7. Use cases

In what follows, possible usages of the MMoOn Core ontology will be outlined. This serves to exemplarily indicate the research potential it entails for the two application areas of linguistics and NLP it has been designed for. It shall be noted that all mentioned usages are equally realizable with the commonly applied methods of language representation and analysis in these fields. However, special awareness should be given to this Linked Data-based approach of MLD representation by using MMoOn Core (alone or in conjunction with other ontologies) because it yields the benefit of interdisciplinary reuse, extension and application as an opportunity to overcome the current limitations of scientific progress caused by data silos and heterogeneous formats.

7.1. Use cases for linguistic research

7.1.1. Enhancement of morphological data in dictionaries

Dictionaries and lexical datasets contain a considerable amount of MLD. This includes derivational morphs and the lexical entries they can be productively combined with but also elements and building patterns of inflectional paradigms that vary in the degree of their descriptive granularity across dictionaries of different languages. In dictionaries of Semitic languages, for instance, headwords are collected around roots which are followed by the full list of word forms but also lexemes which can be derived from them. For the description of such fine-grained morphological data, the creation of MMoOn morpheme inventories enables the representation of this data in an appropriate manner which can serve as an addition to vocabularies that are usually used for representing lexical data. The Hebrew Morpheme Inventory can be seen as a proof for this application of the MMoOn Core ontology [32].

7.1.2. Language acquisition

With the availability of more and more language data the applied linguistic research area of (second) language acquisition is provided with new possibilities for creating language learning materials and tools. Within this setting morphological data plays a significant role for the acquisition of inflection and formation patterns of words. The future morphological datasets, therefore, have the potential to broaden and complement already existing data-driven learning tools and techniques for corpus linguistics [22] with valuable morphological data. Provided by MMoOn morpheme inventories, inflection tables, word families and the grammatical as well as lexical morphs with their usage restrictions can be obtained. In this respect single MMoOn-based datasets can be already regarded as source data for language learning and teaching materials. The created Xhosa RDF dataset [5] is an example for a MMoOn-based dataset with an intended usage for language revitalization efforts for Bantu languages by using the MMoOn Core ontology as the uniting model for collecting interoperable data of multiple Bantu languages [17] to develop various learning materials.

7.1.3. Language documentation

The area of language documentation has the intention to "to provide a comprehensive record of the linguistic practices characteristic of a given speech community"[28]. Since the publication of this paper in 1998, this area has sparked a community which aims to create linguistic resources for endangered and minority languages. As mentioned in Section 3.3, due to the work of the language documentation community, a great amount of interlinear-glossed text resources exist in linguistic databases or as text examples in linguistic publications. However, these linguistic resources do not use the same representation format. Hence, sharing it within and especially outside of this community is difficult. If a language was documented using the MMoOn Core ontology, it would be possible to create other output formats such as tables, dictionaries,

etc. That way the resulting language resource could not only be shared with the language documentation community but, moreover, this data would become usable by the NLP and Semantic Web communities to create tools supporting minority languages.

7.1.4. Representation of morphemic glosses in linguistic literature

Morphemic glosses are part of many linguistic publications and usually used in given examples. A standardized set for interlinear morphemic glosses does not exist and each publication is accompanied with a customized list of glosses. Nonetheless, an adoption of the proposed standardized application within the Leipzig Glossing Rules [13] as well as the reuse of the therein provided set of glosses can be observed. However, the majority of glosses being used is still heterogeneous in that different glosses are used for the same morphemic concepts across the literature. The morphemic glosses provided in MMoOn Core can be regarded as a reference set of glosses since MMoOn Core already reused the existing glosses provided within the Leipzig Glossing Rules which are already widely accepted and applied by linguists. Given that the links between all `mmoon:MorphemicGloss` instances and the linguistic concepts they represent, i.e. the instances of all `mmoon:Meaning` subclasses, are already created, an unambiguous reference can be established. Consequently, including the morphemic gloss URIs within the digital versions of publications of linguistic works can not only contribute to a more consistent usage of glosses but also to a better findability of language examples that are, hitherto, hidden in unstructured text documents.

7.1.5. Comparative linguistics

The internally provided links between the `mmoon:Meaning` classes and `mmoon:MorphemicGloss` instances that come with MMoOn Core entail another possibility, i.e. they are especially suitable for comparative linguistic analyses. This is because a multilingual semantic interconnection is automatically established since all schema ontology files of the MMoOn morpheme inventories are interconnected within a single graph via the imported MMoOn Core ontology. As a result, this allows for a flexible conversion or newly created representation of multiple language datasets taking language-specific characteristics into account while maintaining semantic interoperability simultaneously. Due to this architectural setup of MMoOn Core, reasoning is enhanced and specific queries enable exact investigations of comparative synchronic cross-lingual phenomena and, moreover, tracing historical linguistic changes across multiple datasets at once. In particular the use of the morphemic glosses is facilitating semasiological as well as onomasiological querying because every created language-specific meaning in a morpheme inventory is automatically interlinked to the respective language-independent gloss.

7.2. Use cases for NLP research

7.2.1. Conversion of Wiktionary datasets

The already mentioned MLD provided by Wiktionary (cf. Section 3.2) is one of the largest openly available datasets. In the context of Linked Data-based NLP research it is desirable to create an RDF version of this data. The existing Dbnary morpho dataset is, however, not appropriate for NLP tasks because it covers only four languages, uses an outdated *lemon* vocabulary and contains only a morphological annotation of the grammatical meanings of the word forms given in the Wiktionary inflection tables. Instead, it seems promising to convert existing data provided by UniMorph [30, 31] and paradigm extractions[42] [19] which have already normalized and segmented Wiktionary data into structured formats. The UniMorph 2.0 [31] dataset contains data of 47 languages from Wiktionary that has been normalized with regard to the differing inflection tables and that is semantically annotated with a set of grammatical features which correspond essentially to the `mmoon:MorphemicGloss` instances in MMoOn Core. The data provided by paradigm extract [19] covers only nine languages but is of special interest because the inflectional paradigms extracted from Wiktionary also contain the segmented morphs of a word form. Combined, these two datasets constitute a substantial foundation to convert the word forms and morphs contained within Wiktionary inflection tables into RDF. The architectural setup for creating MMoOn morpheme inventories is suitable to represent the UniMorph and paradigm extract data. Hence, the existing data could not only be made available as Linked Data but also merged within a single data graph in which they would be automatically semantically enriched (by the interlinking of the glosses to meanings and the meanings to morphs) and multilingually interconnected due to the uniting function of the underlying MMoOn Core model.

[42] https://github.com/marfors/paradigmextract

7.2.2. Morphological text annotation

Morphological annotation tools could be created with a data-driven approach based on MMoOn datasets similar to the task of part-of-speech tagging. The initially required RDF representation of corpora can be provided by using the Natural Language Processing Interchange Format (NIF) [26, 44]. The resulting NIF corpus can be then extended with several layers of annotations depending on the granularity of the interconnected MMoOn dataset. This could range from the identification of lexemes, stems, morphosyntactic meanings and also part-of-speech data on the word form level of the tokens up to the segmentation into their morphs together with the underlying inflectional and derivational meanings on the morph level of the tokens. In any case, the `mmoon:MorphemicGloss` resources can be regarded as a ready-to-use tagset for meanings which facilitates the creation of annotations. Such a MMoOn-based morphological text annotation approach could also provide suggestions for unknown tokens due to the possible lookup of their contained morphs (which are likelier to exist in the dataset). The more fine-grained the underlying MMoOn dataset for such an annotation tool is the more detailed linguistic information can be automatically extracted from large amounts of texts. This can in turn impact the results of other NLP tasks and might even lead to the automatic creation of interlinear glossed text.

7.2.3. Named entity recognition

Recent work in the field of named entity recognition (NER) in German has revealed that the complexity of morphology is rarely considered in existing NER tools, even though considering it could lead to improved results [34]. This holds true especially for the identification of NEs (or linguistically termed: proper nouns) which have undergone several morphological transformations and appear within complex lexemes. E.g. in order to retrieve the NE *Alpen* (engl. 'the Alps') within the inflected german noun *Skilalpinistinnen* (engl. 'female ski alpinists') all compositional, derivational and inflectional transformations that have been applied to *Alpen* have to be deconstructed. But also nontransformed proper nouns that are only obligatory affected by inflectional marking can already pose a challenge for NER tools. Within a German MMoOn morpheme inventory the involved morphs *-en*, *-in(1)*, *-ist*, *Ski*, *alp* and *-in(2)* would be available and could help to identify the NE within the common noun. A very elaborate MMoOn dataset could also contain the complete token with its full segmentation, which allows for a direct retrieval of the underlying NE from within the data graph. Since the MMoOn Core ontology enables a comprehensive explication of morphological data, the lack of appropriate morphological data can be overcome. Consequently, future morpheme inventories could be a promising consideration in the development of NER tools and systems.

7.2.4. Machine translation

Machine translation belongs to one of the most complex and challenging tasks in NLP. Dictionaries and lexical data play a crucial role as one of the sources that are utilized for identifying the sense of a word in a text in one language and the respective expressions used for this sense in another language. However, depending on the morphological type of the languages that are to be translated this task is getting increasingly difficult the more the word-to-morpheme ratio deviates from one-to-one correspondences. Machine translation systems that would be complemented by MMoOn-based datasets could rely on the more fine-grained morphological data. This might be especially improving when translating from analytical languages, e.g. Vietnamese, to polysynthetic languages (marking the extremes of the typological continuum) or vice versa. A lexical approach only will not be able to capture for instance sentences like *angya-ghlla-ng-yug-tuq*, 'I have a fierce headache' (Siberian Yupik) [12] because it consists of a single word. Within the MMoOn representation, however, the individual morphs are explicated and could be translated into an isolating or agglutinative language through the senses and grammatical meanings they consist of. Since all MMoOn datasets share the MMoOn Core ontology within the unified graph of a multilingual dataset the atomic morphemes of isolating languages and the fusional morphemes of polysynthetic languages can be identified and translated in an onomasiological way (in contrast to the semasiological approach of lexical data).

7.2.5. Sentiment analysis

Comprehensive MLD also has the potential to contribute to the NLP research field of sentiment analysis. Subjective information about topics within texts is not only encoded lexically but also by morphological means. E.g. the detection of negation, being one of the main issues for sentiment analysis [47], could benefit from a morphological data source such as a MMoOn morpheme inventory because negation can be very productively expressed by using prefixes like *un-* for English together with adjectives. Furthermore, bound morphemes for comparative, superlative or in-

tensification can be easily retrieved from such a dataset and also identified even if the lexemes they are attached to are unknown. In general, MLD represented with MMoOn can explicitly describe obligatory grammatical and highly productive lexical morphemes that express various concepts relevant for sentiment analysis. Consequently, an integration of MLD in the form of MMoOn morpheme inventories poses a promising application case for extending existing resources, algorithms, models and frameworks in the field of sentiment analysis.

8. Concluding remarks

The development of the MMoOn Core ontology started in 2015. Since then, the ontology has been evaluated for its applicability resulting in the Hebrew Morpheme Inventory [32] as proof of concept. Simultaneously, the architectural setup has been developed, morphemic glosses and meanings have been extended and refined. The interim status of the ontology has been presented at various scientific events to gain feedback from the target user groups which has been considered and integrated into the final publication state of MMoOn Core as well. Despite this longstanding process from conceptualizing to actually publishing this accompanying article for the MMoOn Core ontology, no comparable advancement in creating a domain ontology for representing MLD is recorded [6].

As far as the vocabulary use of MMoOn Core is concerned, it achieves a four out of the five star ranking of Linked Data vocabulary use [29]. According to this, MMoOn Core contains dereferencable human-readable information about the used vocabulary (1 star), available information as machine-readable explicit axiomatization of the vocabulary (2 stars), a linking to other vocabularies, i.e. OntoLex-*lemon* (3 stars) and provides metadata about the vocabulary (4 stars). At the current state the fifth star, i.e. vocabularies that link to MMoOn Core, is not achieved. With the awareness that exists already for this domain ontology, however, it is very likely that other vocabularies, e.g. OntoLex-*lemon* or Ligt will create links to MMoOn Core in the future.

In summary, the presentation of the MMoOn Core ontology in this paper has explained how this model will enable the conversion of existing as well as the creation of new morphological datasets and, thus, reaches its aim of contributing to a rising number of homogenized, interoperable linguistic datasets. This result is mainly based on two characteristics of the ontology. First, the rather unusual granularity of the provided meaning classes and their interlinkings with their respective glosses reduce the time for mapping source data of different formats with the ontology and enhances the consistency across datasets. Being embedded within the whole MMoOn Core ontology, these concepts explicate the large part of the linguistic domain of morphology and, therefore, enable the creation, transformation and semantic enrichment of the of MLD that was hitherto inaccessible for machine-processing, e.g. inflection tables, interlinear glossed text, morphological data accompanying lexical databases and dictionaries. The second crucial characteristic of MMoOn Core is its capacity to strengthen the interdisciplinary reuse of MLD originating from the linguistic, NLP and Semantic Web communities. Due to the architectural setup that is based on MMoOn Core, both, language-independent as well as language-specific representations of MLD can be realized. Therefore, depending on the use case and the intended application of the MLD that shall be described as Linked Data either the MMoOn Core ontology can be used to create a very generic and language-independent morpheme inventory or a language-specific schema file that enables specific extensions. Due to the fact that all emerging MMoOn-based datasets are inherently interconnected through the MMoOn Core ontology, datasets that had been of potential interest for a specific user group but have been eventually rejected for an actual reuse (because they were considered too general or too specific in their description) can be now directly adjusted to the required granularity of the representation needs. In this respect it is through the architectural setup of MMoOn Core that the creation of MLD is enabled not only for different user groups and usages but also that all resulting morpheme inventories are semantically unified, thus, leading to an enhanced interoperability and reusability. To conclude, it could be shown that the MMoOn Core ontology contributes to a facilitated and flexible cross-disciplinarity MLD data generation and exchange.

9. Future work

Even though the MMoOn Core ontology as it is published now can be regarded as a ready to use domain

ontology, it is intended to evolve in the future. Collecting and representing all concepts that can be morphologically expressed across the word's languages can not be achieved by a few scientists. Therefore, the meanings provided in MMoOn Core can be regarded as a starting point of the ontology which shall be constantly adapted and extended according to emerging MMoOn morpheme inventories and their schema files. Especially the list of derivational meanings is envisioned to be enlarged and integrated into MMoOn Core from the language-specific datasets.

Another prospective step entails to outreach to other LLOD communities in order to strengthen collaborative research. This is desirable in order to reach the most consistent usage of existing linguistic domain models and data since the considerable overlap of linguistic data compilations of different research areas can not be avoided. Given that MMoOn Core presents a further addition to existing ontologies for the representation of linguistic domains it is advisable to reach a shared agreement on aligning phonological, morphological and lexical data by interconnecting PHOIBLE [42], MMoOn Core and OntoLex-*lemon* respectively.

Similarly, the connection of MMoOn Core and the Ligt ontology will be promoted. In doing so, a higher number of semantically richer morphological datasets from interlinear glossed text sources, especially for less-resourced languages, can be expected in the future.

Finally, work on Linked Data-based solutions for an integration and the transformation of non-RDF resources such as the Typecraft or UniMorph datasets into LLOD based on MMoOn is planned.

Acknowledgements

We acknowledge support from Leipzig University for Open Access Publishing.

References

[1] Allemang D. and Hendler, J.: *Semantic web for the working ontologist: Effective modeling in RDFS and OWL.* Elsevier, 2011. doi:10.1016/c2010-0-68657-3.

[2] Beermann, D. and Mihaylov, P.: *TypeCraft collaborative databasing and resource sharing for linguists.* In: Language Resources and Evaluation **48**(2), Elsevier, 2014, pp. 203–225. doi:10.1007/s10579-013-9257-9.

[3] Berners-Lee, T.: *Linked data-design issues.* URL http://www.w3.org/DesignIssues/LinkedData.html (2011), 2006.

[4] Booij, G.: *The grammar of words: An introduction to linguistic morphology.* Oxford University Press, 2012.

[5] Bosch, S., Eckart, T., Klimek, B., Goldhahn, D. and Quasthoff, U.: *Preparation and usage of Xhosa lexicographical data for a multilingual, federated environment.* In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T. et al. (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 4372–4378.

[6] Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. and Gómez-Pérez, A.: *Models to represent linguistic linked data.* In: Mitkov, R., Tait, J. and Boguraev, B. K. (eds) Natural Language Engineering **24**(6), 2018, pp. 811–859. doi:10.1017/s1351324918000347.

[7] Chavula, C. and Keet, C. M.: *Is lemon sufficient for building multilingual ontologies for Bantu languages?* In: Keet, C. M. and Tamma, V. (eds) 11th OWL: Experiences and Directions Workshop (OWLED), 2014, pp. 61–72.

[8] Chiarcos, C.: *An ontology of linguistic annotations.* In: Mönnich, U. and Kühnberger, K. (eds) GLDV-Journal for Computational Linguistics and Language Technology **23**(1), 2008, pp. 1–16.

[9] Chiarcos, C. and Hellmann, S.: *Working group for open data in linguistics: Status quo and perspectives.* In: Hellmann, S., Frischmuth, P., Auer, S. and Dietrich, D. (eds) Proceedings of the 6th Open Knowledge Conference (OKCon 2011), ceur-ws.org, 2011.

[10] Chiarcos, C. and Ionov, M.: *Ligt: An LLOD-native vocabulary for representing interlinear glossed text as RDF.* In: Eskevich, M. et al.(eds) 2nd Conference on Language, Data and Knowledge (LDK 2019), Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2019.

[11] Cimiano, P., Buitelaar, P., McCrae, J. and Stintek, M.: *LexInfo: A declarative model for the lexicon-ontology interface.* In: Journal of Web Semantics **9**(1), Elsevier, 2011, pp. 29–51. doi:10.1016/j.websem.2010.11.001 .

[12] Comrie, B.: *Language universals and linguistic typology: Syntax and morphology.* University of Chicago press, 1989.

[13] Comrie, B., Haspelmath, M. and Bickel, B.: *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.* In: Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig. Available online at https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf (July 2016), 2008.

[14] Creutz, M. and Lagus, K.: *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0.* Helsinki University of Technology, Helsinki, 2005.

[15] Creutz, M. and Lagus, K.: *Inducing the morphological lexicon of a natural language from unannotated text.* In: Honkela, T., Könönen, V., Pöllä, M. and Simula, O. (eds) Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05) **1**, 2005, pp. 106–113.

[16] Declerck, T.: *Representation of polarity information of elements of German compound words.* In: McCrae, J. P., Chiarcos, C., Montiel Ponsoda, E., Declerck, T., Osenova, P. and Hellmann, S. (eds) Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL 2016), 2016, pp. 46-49.

[17] Eckart, T., Bosch, S., Goldhahn, D., Quasthoff, U. and Klimek, B.: *Translation-based dictionary alignment for under-resourced Bantu languages.* In: Eskevich et al. (eds) 2nd Conference on Language, Data and Knowledge (LDK 2019), Schloss

Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019, pp. 17:1–17:11. doi:10.4230/OASIcs.LDK.2019.17.

[18] Farrar, S. and Langendoen, D. T.: *An OWL-DL implementation of gold.* In: Witt A. and Metzing D. (eds) Linguistic Modeling of Information and Markup Languages. Text, Speech and Language Technology **41**, Springer, Dordrecht, 2010, pp. 45–66. doi:10.1007/978-90-481-3331-4_3.

[19] Forsberg, M. and Hulden, M.: *Learning transducer models for morphological analysis from example inflections.* In: Jurish, B., Maletti, A., Würzner, K. and Springmann, U. (eds) Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata, 2016, pp. 42–50. doi:10.18653/v1/w16-2405.

[20] Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. and Soria, C.: *Lexical markup framework (LMF).* In: Calzolari, N., Choukri, K., Gangemi, A. et al. (eds) Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC 2006), 2006, pp. 233–236.

[21] Francopoulo, G. and Paroubek, P. (eds): *LMF lexical markup framework.* John Wiley & Sons, 2013. doi:10.1002/9781118712696.

[22] Gaëtanelle, G. and Granger, S.: *How can data-driven learning be used in language teaching?* In: O'Keeffe, A. and McCarthy, M. (eds) The Routledge Handbook of Corpus Linguistics. Routledge, London, 2010, pp. 387–398. doi:10.4324/97802038569 49.ch26.

[23] Haspelmath, M.: *Comparative concepts and descriptive categories in crosslinguistic studies.* In: Carlson, G. N. (ed) Language **86**(3), Linguistic Society of America, 2010, pp. 663–387. doi:10.1353/lan.2010.0021.

[24] Haspelmath, M. and Sims, A.: *Understanding morphology.* Routledge, 2013. doi:10.4324/9780203776506.

[25] Haspelmath, M.: *The Leipzig style rules for linguistics.* Max Planck Institute for Evolutionary Anthropology, Leipzig, URL http://www.uni-regensburg.de/sprache-literatur-kultur/sprache-literatur-kultur/allgemeine-vergleichende-sprachwissenschaft/medien/pdfs/haspelmath_2014_style_rules_linguistics.pdf, 2014.

[26] Hellmann, S.: *Integrating natural language processing (NLP) and language resources using linked data.* Ph.D. Dissertation, Leipzig University, Leipzig, 2014.

[27] Hellmann, S., Moran, S., Brümmer, M. and McCrae, J. (eds): *Special Issue on Multilingual Linked Open Data (MLOD).* Semantic Web **6**(4), IOS Press, 2015.

[28] Himmelmann, N. P.: *Documentary and descriptive linguistics.* In: Linguistics **36**(1), Walter de Gruyter, Berlin, 1998, pp. 161–196. doi:10.1515/ling.1998.36.1.161.

[29] Janowicz, K., Hitzler, P., Adams, B., Kolas, D. and Vardeman II, C.: *Five stars of linked data vocabulary use.* In: Semantic Web **5**(3), IOS Press, 2014, pp. 173–176. doi:10.3233/sw-140135.

[30] Kirov, C., Sylak-Glassman, J., Que, R. and Yarowsky, D.: *Very-large scale parsing and normalization of Wiktionary morphological paradigms.* In: Calzolari, N., Choukri, K., Declerck, T. and Goggi, S. et al. (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 3121–3126.

[31] Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M. et al.: *UniMorph 2.0: universal morphology.* In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T. et al. (eds) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 1868–1873.

[32] Klimek, B., Arndt, N., Krause, S. and Arndt, T.: *Creating linked data morphological language resources with MMoOn - The Hebrew morpheme inventory.* In: Calzolari, N., Choukri, K., Declerck, T. and Goggi, S. et al. (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 892–899.

[33] Klimek, B.: *Proposing an OntoLex - MMoOn alignment: Towards an interconnection of two linguistic domain models.* In: McCrae, J. P., Bond, F., Buitelaar , P. et al. (eds) Proceedings of the LDK Workshops: OntoLex, TIAD and Challenges for Wordnets, 2017, pp. 68–83.

[34] Klimek, B., Ackermann, M., Kirschenbaum, A. and Hellmann, S.: *Investigating the morphological complexity of German named entities: The case of the GermEval NER challenge.* In: Rehm, G. and Declerck, T. (eds) Proceedings of the 27th biennial Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2017: Language Technologies for the Challenges of the Digital Age, Springer, Cham, 2017, pp. 130–145. doi:10.1007/978-3-319-73706-5_11.

[35] Klimek, B., McCrae, J., Bosque-Gil, J., Ionov, M., Tauber, J. K. and Chiarcos, C.: *Challenges for the representation of morphology in ontology lexicons.* In: Kosem, I., Zingano Kuhn, T., Correia, M. et al. (eds) Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 conference., 2019, pp. 570–591.

[36] Lehmann, C.: *Data in linguistics.* In: The Linguistic Review **21**(3-4), De Gruyter Mouton, 2004, pp. 175–210. doi:10.1515/tlir.2004.21.3-4.175.

[37] Lehmann, C.: *Interlinear morphemic glossing.* In: Booij, G., Lehmann, C., Mugdan, J. and Skopeteas, S. (eds) Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung 2, Walter de Gruyter, 2004, pp. 1834–1857. doi:10.1515/9783110172782.2.

[38] Lehmann, Christian: *Linguistic concepts and categories in language description and comparison.* In: Chini, M. and Cuzzolin, P. (eds) Typology, acquisition, grammaticalization studies, Franco Angeli, Milano, 2018, pp. 27–50.

[39] McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J. et al.: *Interchanging lexical resources on the semantic web.* In: Language Resources and Evaluation **46**(4), Springer Science and Business Media (LLC), 2012, pp. 701–719. doi:10.1007/s10579-012-9182-3.

[40] McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S. and Osenova, P.: *The open linguistics working group: Developing the Linguistic Linked Open Data cloud.* In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S. et al. (eds) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 2435–2441.

[41] McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P.: *The OntoLex-Lemon model: Development and applications* In: Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S. and Baisa, V. (eds) Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference : Lexicography from Scratch, 2017, pp. 587–597.

[42] Moran, S., McCloy, D. and Wright, R.: *PHOIBLE online.* URL http://phoible.org/, Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014.

[43] Perlmutter, D., Hammond, M. and Noonan, M.: *The split morphology hypothesis: Evidence from Yiddish.* In: Hammond, M. and Noonan, M. (eds) Theoretical morphology: Approaches in modern linguistics, San Diego: Academic Press, 1988, pp. 79–100.

[44] Röder, M., Usbeck, R., Hellmann, S. and Gerber, D.: *N^3-A Collection of datasets for named entity recognition and disambiguation in the NLP interchange format.* In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H. et al. (eds) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), 2014, pp. 3529–3533.

[45] Sagot, B.: *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French* In: Calzolari, N., Choukri, K., Maegaard, B. et al. (eds) Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), 2010, pp. 2744–2751.

[46] Sérasset, G.: *Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF* In: Hellmann, S., Moran, S., Brümmer, M. and McCrae, J. (eds) Special Issue on Multilingual Linked Open Data (MLOD), Semantic Web **6**(4), IOS Press, 2015, pp. 355–361. doi:10.3233/sw-140147.

[47] Wiegand, M., Balahur, A., Roth, B., Klakow, D. and Montoyo, A.: *A survey on the role of negation in sentiment analysis.* In: Morante, R. and Sporleder, C. (eds) Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, 2010, pp. 60–68.

2.3 Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models

Given that lexical language data often also contains a considerable amount of morphological language data it is regarded as common practice to create an interconnection between the MMoOn Core ontology and other existing vocabularies that model lexical data. Since the OntoLex-*lemon* model is the most used model of published lexical language data as Linked Data it is desirable to establish an alignment between the two models. This publication encourages data owners of lexical data based on the OntoLex-*lemon* vocabulary to use the MMoOn Core ontology for representing the morphological data that remained indescribable with OntoLex-*lemon*. Therefore, [P3] provides a detailed comparison of the representability of language elements in MMoOn Core and OntoLex-*lemon* on the lexeme, word-form, morph and morpheme levels. On this basis the conceptual overlap of both ontological models is investigated and specific alignments are proposed.

Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models

Bettina Klimek[1]

AKSW/KILT, University of Leipzig,
klimek@informatik.uni-leipzig.de,
WWW home page: http://aksw.org/Groups/KILT.html

Abstract. This paper motivates and proposes to align the OntoLex and MMoOn Core models. It deals in particular with the ontolex and decomp modules and their potential as ontological foundation to represent the domain of morphological language data (MLD). It will be argued that ontolex and decomp provide only a basic modelling of the domain, which is not sufficient for representing fine-grained MLD, but suitable for interconnecting OntoLex with the Multilingual Morpheme Core Ontology (MMoOn Core). Both models each offer a modelling of a linguistic domain - OntoLex for lexical language data and MMoOn Core for morphological language data - that exhibits a notable amount of conceptual overlap. Thus, this paper investigates the potential of exploiting the overlap of both models for initiating an ontology-based interconnection of lexical and morphological datasets.

Keywords: MMoOn, OntoLex, morphology, ontology alignment

1 Introduction

The development of OntoLex as a standardized model for the ontological representation of lexical language data has gained high acknowledgement within the Linguistic Linked Open Data (LLOD) community. A reason for that lies in the far reaching modelling of lexical language data (LLD) that goes beyond the domain of lexicography. By providing five modules, the OntoLex model can be used according to the needs of a dataset creator to also represent morphological, syntactical, semantic and translational information about a lexical entry as well. I.e. the OntoLex model encompasses the representation of other linguistic domains as well.

This paper deals in particular with the ontolex and decomp modules and their potential as ontological foundation to represent the domain of morphological language data (MLD). It will be argued that ontolex and decomp provide only a basic modelling of the domain, which is not sufficient for representing fine-grained MLD, but suitable for interconnecting OntoLex with the Multilingual Morpheme Core Ontology (MMoOn Core)[1]. Both models each offer a modelling

[1] http://mmoon.org

of a linguistic domain – OntoLex for LLD and MMoOn Core for MLD – that exhibits a notable amount of conceptual overlap. The aim of this paper is, thus, to investigate the potential of exploiting this overlap of both models for initiating an alignment of both ontologies. Dataset creators of either OntoLex or MMoOn datasets would benefit from such a unification in that it enables the seamless extension of lexical OntoLex data with morphological MMoOn data or of morphological MMoOn data with lexical OntoLex data.

The remainder of the paper is structured as follows. Section 2 gives a brief overview of the domain of MLD and its overlap to the domain of LLD. In Section 3 the MMoOn Core model is summarized and presented as a suitable model for the domain of MLD. The main part of the paper constitutes Section 4 which investigates the representation of MLD in both models in a comparative way and which points to the overlapping aspects. It further serves to not only prove that the MMoOn Core model is qualified to be interconnected with OntoLex but also to show that both models would benefit from an alignment with regard to the representation of language data in both linguistic domains. Thereafter, in Section 5, specific interconnection points between both models are proposed together with practical issues that need to be considered for implementing an alignment of both ontologies. The paper closes with a summary in Section 6.

2 Scope and Delimitation of the Domain of Morphology

In traditional linguistics research fields such as phonology/phonetics, morphology, lexicology, syntax, semantics and pragmatics are distinguished. However, the study of one field reveals considerate inter-dependencies to other fields. E.g. the field of morphophonology investigates the interface of phonology and morphology. Similarly, there is an overlap of morphology and lexicology which in the view of linguistic data representation makes it hard to state where the domain of morphology ends and the domain of lexicography begins. Since "lexical items are the fundamental building blocks of morphological structure" [4] it is not satisfactory to represent lexical data only in lexicons and morphological data only in morphemicons. Even though such lexicons and morphemicons constitute valuable data resources, it is desirable to interconnect both. E.g. a lexicon entry might be the English adjective *unreal* and a morphemicon entry might be the negation prefix *un-*. In both separate dataset types the information that the adjective consists of this very prefix and the information which other lexical entries also contain this prefix is missing.

In the scope of Linked Data such information can be modelled in an ontology, which provides the necessary relations that interconnect lexical items and morphological items. But then the question arises: what kind of data should be represented in the linguistic domain of morphology? Figure 1 illustrates a data-driven view of the domain with the English example lexeme *(to) play*. The box in the middle indicates the narrow scope of the domain, i.e. which elements and their relations need to be modelled in order to represent MLD. The central entries of MLD are thus morphs, morphemes and meanings. But also information

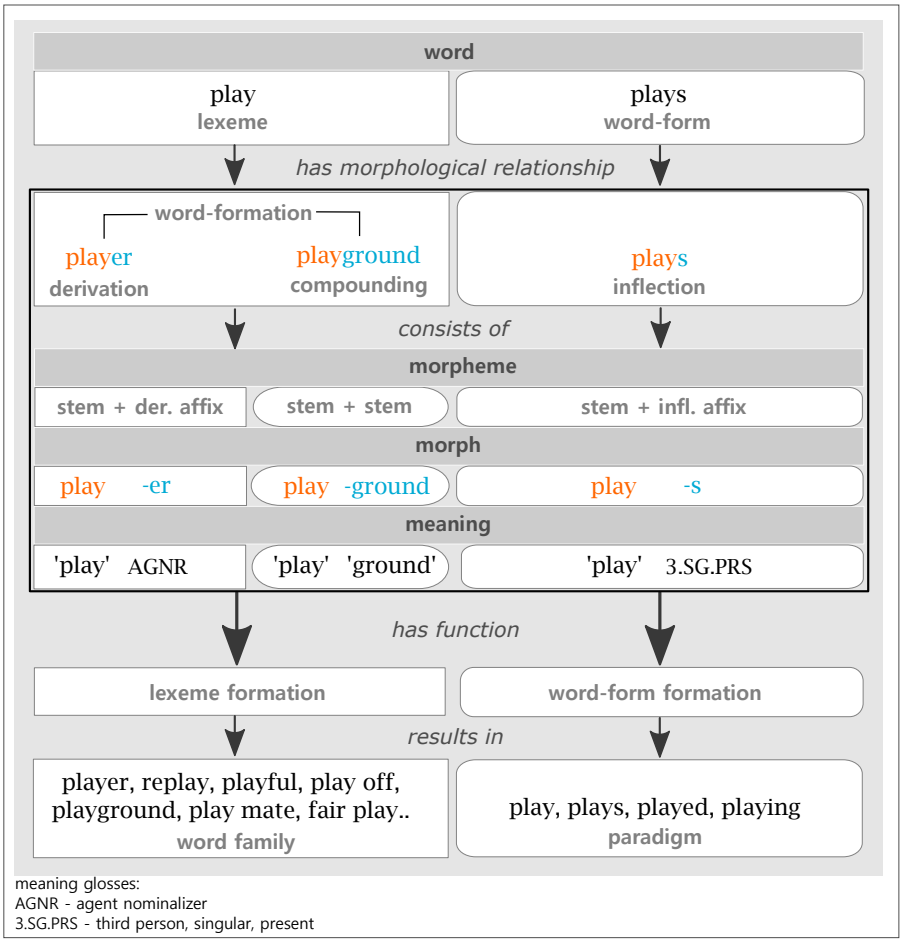


Fig. 1. Overview of the linguistic domain of morphology with the English example lexeme "play" (verb).

on the function as a derivational, compounding or inflectional morph/morpheme within a given word needs to be provided. The rest of the Figure shows how the the narrow MLD is interrelated to corresponding LLD. As a result, Figure 1 as a whole shows the wide scope of the MLD domain, which then also includes lexemes and word-forms. Consequently, only in the wide scope of MLD interrelating information between lexical and morphological items can be obtained, i.e. the identification of word-families and word-forms. This, however, means that there is no clear cut delimitation between the two domains of LLD and MLD and especially word-families and word-forms could be regarded as elements of both domains, since their representation requires lexical as well as morphological knowledge and data. Nonetheless, in the research field of LLOD this situation is

not problematic. To the contrary, the Linked Data format is open for extension, so that existing lexical or morphological datasets in RDF can be interconnected across various vocabularies, e.g. by aligning the two domain models of OntoLex and MMoOn.

3 Why MMoOn?

To my knowledge, the *lemon* model [10, 9] and the resulting OntoLex model were the first models to provide an ontological representation of the domain MLD. Given that the central domain of these vocabularies is LLD, it is not surprising that the domain of MLD as shown in Figure 1 is only partially covered and questions arose on the applicability of both models for representing more fine-grained MLD [2].

In order to fill the gaps in these vocabularies (which will be discussed in Section 4) and to obtain a more extensive model that covers the full domain of MLD, the MMoOn Core ontology has been developed. The MMoOn model has proven to be applicable for the representation of inflectional languages, even for those exhibiting non-concatenative morphology, such as Hebrew [6]. Due to the shortage of space in this paper an image giving an overview of the MMoOn Core ontology can be found here: `http://mmoon.org/mmoon-core-model`, the full vocabulary here: `http://mmoon.org/core.owl` and the documentation of anything related to the ontology and emerging datasets here: `https://github.com/MMoOn-Project`. The MMoOn Core ontology is designed as a language-independent and theory-neutral model to create language-specific morpheme inventories. It consists of eight main classes: `MorphemeInventory`, `MorphologicalRelationship`, `MorphemicGloss` and `Meaning` which enable the representation of secondary language data and `Word`, `Morph`, `Morpheme` and `Representation` which are used to describe primary language data[2]. With regard to the modelling of secondary data, the OntoLex developers declare that the model "does not prescribe any vocabulary for doing so [i.e. recording linguistic properties] , but leaves it at the discretion of the user of the model to select an appropriate vocabulary [...]"[3]. As this complies to the common best practice for Linked Data to reuse existing vocabularies, such descriptive secondary language data will remain undiscussed within the modelling of MLD in OntoLex in this paper. It shall be noted, that MMoOn Core comes with nearly 300 meanings to which morphemic glosses are already assigned. Even though there is an overlap to vocabularies such as LexInfo [3], meaning resources are included in MMoOn because it includes also derivational meanings and facilitates the creation of a

[2]The former includes descriptive data which enables the assignment of linguistic features (or properties), e.g. grammatical categories or part of speech, and the latter contains all elements and their relations within a given language that are part of the domain, e.g. morphs, morphemes, word-forms. (For more detail see [7].)

[3]Every reference to the OntoLex model or any of its modules is made with regard to the model specification here: `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Linguistic_Description`.

MMoOn dataset (especially for linguists, who then do not have to deal with various vocabularies). The class hierarchies in MMoOn Core are fine-grained and interrelated with various object properties. This allows for explicitly stating which parts of the words are morphologically formed as well as to which words morphs and morphemes belong. As a result, a MMoOn morpheme inventory is more than a mere morphemicon: it is a semantically structured data graph that can be traced in both directions from words to morphemes in a semasiological and an onomasiological way. In particular the modelling encompasses the elements and their relations of the domain of MLD as shown in Figure 1. A dataset created with MMoOn is called a MMoOn morpheme inventory. Every morpheme inventory consists of three files: 1) The Core model, which functions as a cross-linguistic template for the domain of MLD, 2) a schema file, which is language-specific and describes the secondary language data and 3) an inventory file that contains only primary language data, i.e. only instance data. This schema file – or language-specific morpheme ontology – is derived from and imports the Core ontology. Hence, it contains all elements that are already provided in MMoOn Core and can be easily further adjusted and extended according to the morphological phenomena that shall be represented in a given language. Thus, the MMoOn Core model is suitable for the semantic modelling of MLD of any inflectional language and, therefore, an appropriate candidate for an alignment with the ontolex and decomp module.

4 Representing Morphological Data

In the following sections it will be shown how MLD is representable with MMoOn on the one side and with the ontolex and decomp modules on the other side. This direct comparison takes up Figure 1 as running example and aims at stressing why an interconnection of MMoOn and the two modules can be regarded as a valuable contribution to the ontological modelling of LLD and MLD in general.

4.1 Morphology on the Lexeme Level

A fundamental distinction in the domain of morphology is inflection and word-formation. The former involves word-form formation and the latter lexeme formation. Inflectional information on the lexeme level contains information on the building pattern of the word-forms of a lexeme.

As Example 1 shows, the ontolex object property `morphologicalPattern` can be used to express the inflectional class of a lexeme. The "?" in the subject slot indicates that the OntoLex model specification states that "the implementation of these patterns is not specified [...] but should be provided by some suitable vocabulary such as [the Lemon Inflectional and Agglutinative Morphology Module for OntoLex] LIAM[4]". What is more, the object property provided, does not differentiate inflectional and derivational relations of lexemes. The MMoOn Core

[4] `http://lemon-model.net/liam`

ontolex/decomp	MMoOn
Example 1: Representing the inflectional morphological relationship of a lexeme.	

`ontolex:lex_play a ontolex:Word ;`
`ontolex:morphologicalPattern ? .`
	`eng_inv:SimpleLexeme_play_v`
	`a eng_schema:SimpleLexeme ;`
	`mmoon:inflectionalRelation`
	`eng_schema:RegularConjugation .`

Example 2: Representing the derivational morphological relationship of a lexeme.	

`ontolex:lex_play a ontolex:Word ;`
`ontolex:morphologicalPattern ? .`
	`eng_inv:DerivedWord_player_n`
	`a eng_schema:DerivedWord ;`
	`mmoon:derivationalRelation`
	`eng_schema:AgentNoun .`

Example 3: Representing that one lexeme is derived from another lexeme.	

`ontolex:lex_player a ontolex:Word ;`
`?decomp:subterm ontolex:lex_play .`
	`eng_inv:DerivedWord_player_n`
	`a eng_schema:DerivedWord ;`
	`mmoon:isDerivedFrom`
	`eng_schema:SimpleLexeme_play_v .`

Example 4: Representing that a lexeme is a compound word that is composed of two other lexemes.	

`ontolex:lex_playground`
`a ontolex:Word ;`
`decomp:subterm ontolex:lex_play ,`
`ontolex:lex_ground .`
	`eng_inv:CompoundWord_playground_n`
	`a eng_schema:CompoundWord ;`
	`mmoon:isComposedOf`
	`eng_schema:SimpleLexeme_play_v ,`
	`eng_schema:SimpleLexeme_ground_n .`

model, however, already contains a basic modelling of classes for inflection and word-formation within the `MorphologicalRelationship` main class, which are automatically reused and provided in every language specific MMoOn schema ontology, e.g. `eng_schema` in the provided examples.

The case of Example 2, representing the derivational morphological relationship of a lexeme, is similar to Example 1. While the MMoOn vocabulary provides object properties that indicate an inflectional or derivational relation and also the kind of this relation in the subject slot of the triple, the ontolex object property remains ambiguous. Given that this property has no range declaration, it is, however, possible to use the MMoOn vocabulary to fill the subject slot. Further, it is important to note, that the LIAM vocabulary does not provide a general ontological modelling of morphological relationships such as MMoOn. Rather, it models the transformation rules that apply to a pattern underlying a specific morphological relation, which could then be applied for instance to `eng_schema:RegularInflection` or `eng_schema:AgentNoun`[5].

[5] `AgentNoun` is part of the class hierarchy: MorphologicalRelationship> WordFormation>Derivation>DerivedNoun>DeverbalNoun>AgentNoun

The Examples 3 and 4 show which other lexemes are involved in a word-formation process. The MMoOn vocabulary provides the two object properties `isDerivedFrom` and `isComposedOf` to state from which lexeme a derived word is derived and of which two lexemes a compound word is composed. The decomp object property `subterm` can be equivalently used for compound words in Example 4. The "?" in Example 3, however, indicates that this predicate is not appropriate for stating that the noun *player* is derived from the verb *play*, because `subterm` is defined as a property that "relates a compound lexical entry to one of the lexical entries it is composed of"[6].

As the examples show, the ontolex and decomp vocabulary is not accurate enough to represent the morphological relationship, either inflectional or derivational, of lexemes. In the cases of stating which lexemes are involved in the word-formation process, the model clearly favours compound words, while lacking an object property that interconnects a lexeme as the basis of a derived word. For such cases the MMoOn vocabulary would be a valuable addition to represent more fine-grained lexical data because it provides more specific object properties and also a more precise classification of lexical entries, i.e. it distinguishes simple lexemes, which are neither composed nor derived from other lexemes, derived words and compound words as subclasses of the MMoOn `LexicalEntry` class. What is more, an alignment of the ontolex `LexicalEntry` class with these classes would be cruicial in order to interconnect an OntoLex lexical dataset with a MMoOn morpheme inventory.

4.2 Morphology on the Word-form Level

In the domain of MLD word-forms play a central role, because these are the entities which contain the inflectional affixes that mark the grammatical variant of a lexical entry. Consequently, all word-forms of a lexeme need to be represented as separate resources in a dataset. As can be seen in Example 5, both the ontolex module and the MMoOn ontology provide properties and classes to do so[7]. While ontolex has one class, `Form`, in MMoOn the class `Wordform` is further specified for the two subclasses `SyntheticWordform` and `AnalyticWordform`[8]. In order to enable the extraction of inflectional paradigms of lexemes, word-form instances in MMoOn can be assigned to more specific morphological relationships. I.e. the synthetic word-forms *play* and *plays* belong to a regular present tense conjugation paradigm (which is not shown in the example but works similar to lexemes shown in Example 1). The analytic word-form *has played*, however, belongs to

[6] cf.:`https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[7] Note that the word-forms in Example 6 are not complete.

[8] The two concepts of 'synthetic' and 'analytic word-form' correspond to the definitions of Christian Lehmann: "A word form is synthetic [...] iff all its semantic and grammatical components are represented in one word form." and "A word form is analytic iff it consists of more than one word form such that the lexical meaning provides the root of one of them, while the grammatical meaning components are coded in the other word forms [...]. Cf. the entries "analytic structure" and "synthesis" at `http://linguistik.uni-regensburg.de:8080/lido/Lido`.

a regular past tense conjugation paradigm and consists of a word-form of *have* and the past participle of *play*. With the object property `consistsOfWord` every word or word-form which is contained in an analytic word-form can be explicitly stated and further represented as well. This is not possible in ontolex, since the appropriate object properties that are available do not take the class `Form` as range but only `LexicalEntry`[9].

Example 5: Representing the word-forms of a lexeme.

ontolex/decomp

```
ontolex:lex_play a ontolex:LexicalEntry ;
    ontolex:canonicalForm ontolex:form_play ;
    ontolex:otherForm ontolex:form_plays ,
                      ontolex:form_played ,
                      ontolex:form_playing ,
                      ontolex:form_has_played.
```

MMoOn

```
eng_inv:SimpleLexeme_play_v a eng_schema:SimpleLexeme ;
     mmoon: hasWordform  eng_inv:SyntheticWordform_play_1_v  ,
                         eng_inv:SyntheticWordform_plays_v ,
                         eng_inv:SyntheticWordform_played_v ,
                         eng_inv:SyntheticWordform_playing_V ,
                         eng_inv:AnalyticWordform_has_played_v .
```

Example 6: Representing the inflectional features that are encoded in a word-form.

ontolex/decomp

```
ontolex:form_plays a ontolex:Form ;
    lexinfo:number lexinfo:singular ;
    lexinfo:person lexinfo:thirdperson ;
    lexinfo:tense lexinfo:present .
```

MMoOn

```
eng_inv: SyntheticWordform_plays_v a eng_schema:SyntheticWordform ;
     mmoon:inherentInflectionalMeaning eng_schema:ThirdPerson ,
                                       eng_schema:Singular ,
                                       eng_schema:Present .
```

Example 7: Representing the morphs of which a word-form consists.

MMoOn

```
eng_inv: SyntheticWordform_plays_v a eng_schema:SyntheticWordform ;
     mmoon:consistsOfStem eng_inv:Stem_play_v ;
     mmoon:consistsOfAffix eng_inv:Suffix_s1 .
```

Example 5 further shows that MMoOn provides the property `hasWordform`, which is inverse of `belongsToLexeme`, for interrelating word-forms and lexemes. In ontolex, given that it is primarily concerned with lexical data, two properties

[9] Otherwise, analytic word-forms could be similarly representable in ontolex/decomp as constituents of multiword expressions.

are provided, i.e. `canonicalForm` and `otherForm`. This specification is clearly useful for compiling dictionaries. For stating all word-forms of a lexeme it might, however, not always be appropriate. At a first glance it seems as if the classes `ontolex:Form` and `mmoon:Wordform` could be used equivalently. That would be true, if all `Form` instances which are connected to a lexical entry via the two mentioned object properties could be regarded as word-forms of a lexical entry. In languages like German for instance, the canonical lexical entry of verbs is the infinitive, which is not an inflected word-form of the lexical verb entry. Querying all `Form` instances as 'word-forms' of a `LexicalEntry` in ontolex might thus return incorrect results. It needs to be mentioned here, that it is not clear from the OntoLex model specification if the representation of word-forms by using the `Form` class is considered or even intended. From the examples given in the specification one can conclude that different forms ("non-lemma") of lexical entries should be describable, but for a specific representation of the word-forms of a lexical entry the vocabulary seems not explicit enough with regard to the provided object properties and the rather general `Form` class[10].

Next to representing word-forms as separate resources, stating information about the grammatical features for which a word-form inflects is also part of the MLD domain. Example 6 shows that ontolex proposes here the use of the LexInfo vocabulary. Since one of the purposes of the MMoOn model is to enable a language-specific description of linguistic categories, a wide range of grammatical meanings is provided in the MMoOn Core vocabulary which are reused in every language-specific MMoOn schema ontology, e.g. `eng_schema:Singular rdf:type mmoon:Singular`. In addition, various differentiating object properties, such as `inherentInflectionalMeaning` in Example 6 or `contextualInflectionalMeaning` which are based on [1], are also established.

This kind of "annotating" word-forms or lexemes for their grammatical features is quite common, but of more significance in the domain of MLD is the identification of those meaningful parts within a word-form that encode the grammatical features and which are identifiable by segmentation, i.e. the morph entities. Consequently, it is necessary to state of which morphs a word-form (or word in general) consists. At this point the ontolex/decomp modules delimit the ontological representation to lexical data. Although, the ontolex class `Affix` is part of the vocabulary, the usage of this class remains quite limited. Because word-forms are not considered as ontolex `LexicalEntry`, but only as `ontolex:Form` instances, none of the ontolex/decomp object properties can be used for making more statements about the components of word-forms. Example 7 illustrates how morphs are explicated as segments of word-forms in MMoOn. A word-form always consists of a stem, which is the semantic core shared with the corresponding lexeme, and some inflectional affix(es). With the dedicated property `consistsOfMorph`[11], which is inverse of `belongsTo`, morph resources can

[10] Also in the model specification a property `ontolex:form` is used multiple times, even though not specified in the vocabulary.

[11] The two MMoOn Core object properties used in Example 7 are subproperties of `mmoon:consistsOfMorph`.

be assigned to the word-forms in which they occur. In this regard, a connection between both models would be very helpful in order to specify more information about word-forms in an OntoLex dataset.

4.3 Morphology on the Morph Level

Example 8: Representing (bound) morphs.

MMoOn

```
eng_inv:Stem_play_v a eng_schema:Stem ;
     mmoon:belongsTo eng_inv:SimpleLexeme_play_v ,
                     eng_inv:DerivedWord_player_n.
eng_inv:Suffix_s1 a eng_schema:Suffix ;
     mmoon:attachedToStem eng_inv:Stem_play_v ,
                          eng_inv:Stem_call_v ;
     mmoon:hasRepresentation eng_inv:Representation_s1 .
eng_inv:Representation_s1 mmoon:morphemicRepresentation "-s" .
```

Example 9: Representing the meaning of morphs.

MMoOn

```
eng_inv:Stem_play_v a eng_schema:Stem ;
     mmoon:hasSense mmoon:Sense_play_v_s1 ;
     mmoon:senseLink <http://lexvo.org/id/wordnet/30/verb/play_2_33_00> .
eng_inv:Suffix_s1 a eng_schema:Suffix ;
     mmoon:inherentInflectionalMeaning eng_schema:ThirdPerson ,
                                       eng_schema:Singular ,
                                       eng_schema:Present .
eng_inv:Suffix_er1 a eng_schema:Suffix ;
     mmoon:derivationalMeaning eng_schema:AgentNominalizer.
```

Example 10: Representing morphemic homonymy and allomorphy.

MMoOn

```
eng_inv:Suffix_s1 a eng_schema:Suffix ;
     mmoon:isHomonymTo eng_inv:Suffix_s2, eng_inv:Suffix_s3 .
eng_inv:Suffix_er1 a eng_schema:Suffix ;
     mmoon:belongsTo eng_inv:DerivedWord_player_n ,
     mmoon:isAllomorphTo eng_inv:Suffix_or .
```

Morph (and morpheme) resources constitute the morphemic entries of each MMoOn morpheme inventory and are in the center of the MLD domain. In general they correspond to the segmented line within an interlinear morphemic glossed text [8]. A morph is the perceivable side of a morpheme, i.e. it is orthographically and phonemically representable. For representing bound morphs, the ontolex module provides only the `Affix` class which is not further specified. The

only possible statement which can be made, is to make an `Affix` class assignment of some suffix, prefix, infix or circumfix resource. Because of this limitation, the examples 8 to 10 only show MMoOn examples. Nothing that is illustrated can be expressed with the ontolex/decomp modules. The MMoOn vocabulary provides a `Morph` class which contains the following subclasses; `Affix`, `Stem` and `Root`. The `Affix` class is further broken up into the `Prefix`, `Suffix`, `Infix`, `Circumfix`, `Simulfix`, `Transfix`, `EmptyMorph` and `ZeroMorph` subclasses. By that, a precise representation of all morph elements which can be segmented from lexical entries or word-forms shall be enabled. Example 8 shows the representation of the verbal stem *play* and the inflectional suffix *-s*. For stem resources it can be further stated to which word resource they belong. In the example the stem *play* belongs to the simple lexeme *play* and the derived word *player*, but additionally it belongs to all word-forms of the simple lexeme *play*. For affix resources it can be stated to which root or stem resource an affix is attached to. In the example the suffix belongs to two stem resources, indicating that affixes are not only listed but also semantically interconnected to other morphemic or lexical entries in a MMoOn morpheme inventory. Further, the datatype property `morphemicRepresentation` is additionally provided to enable the representation of the morpheme boundary or position of the morph within a word.

By having separate morph resources one can additionally specify which parts of a word encode which meaning. Within word-forms the lexical meaning is usually encoded by the stem resource and the grammatical meaning by the affix(es). This is shown in Example 9. With MMoOn new senses can be defined for stem (and word) resources or one can link already existing senses via the `senseLink` property. Since the sense of a stem is the same as the sense of its corresponding lexeme, lexical sense resources provided in already existing LLD datasets could be used to assign sense information to MMoOn `Stem` instances. Lexical senses are not regarded as part of the MLD domain within the MMoOn Core model, but extensively modelled within OntoLex, which presents a potential interconnection point between both models. The grammatical meanings of inflectional affixes like `eng_inv:Suffix_s1` is stated with the same property as in Example 6. In contrast, however, this assignment to the suffix resource is more precise in terms of morphological segmentation. Moreover, MMoOn provides the property `derivationalMeaning` and a set of derivational meanings which can be used to specify resources such as `eng_inv:Suffix_er1`.

Finally, the MMoOn Core vocabulary contains two properties to state homonymous and allomorph relations between morphs, as is illustrated in Example 10. There are several *-s* suffixes in English which share the same surface form but encode different meanings. The `eng_inv:Suffix_s2` encodes plural in nouns and the `eng_inv:Suffix_s3` marks the genitive case, hence, they are represented as being homonym to `eng_inv:Suffix_s1`. For morphs which have different surface forms but share the same meaning the `isAllomorphTo` property is established. E.g. the two instances `eng_inv:Suffix_er1` and `eng_inv:Suffix_or` both encode the derivational meaning of agent nominalizer but occur in complementary distribution, i.e. they attach to distinct verb stems. As the examples show, the

MMoOn vocabulary enables a fine-grained representation of morphemic language data that is semantically relatable to lexical language data. A connection of the ontolex/decomp modules with MMoOn Core would facilitate a morphological description of lexical data with MMoOn on the one side and a lexical description of morphemic data with ontolex/decomp on the other side.

4.4 Morphology on the Morpheme Level

Example 11: Representing morphemes.

MMoOn

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme .
```

Example 12: Representing the meaning a morpheme resource represents.

MMoOn

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme ;
     mmoon:derivationalMeaning eng_schema:AgentNominalizer ;
     mmoon:hasAbstractIdentity mmoon:MorphemicGloss_AGNR .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme ;
     mmoon:inflectionalMeaning eng_schema:ThirdPerson ,
                               eng_schema:Singular ,
                               eng_schema:Present ;
     mmoon:hasAbstractIdentity mmoon:MorphemicGloss_3P ,
                               mmoon:MorphemicGloss_SG ,
                               mmoon:MorphemicGloss_PRS .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme ;
     mmoon:inflectionalMeaning eng_schema:NoMeaning ;
     mmoon:hasAbstractIdentity mmoon:MorphemicGloss_E .
```

Example 13: Representing the relation between morphemes and morphs.

MMoOn

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme ;
     mmoon:hasRealization eng_inv:Suffix_er1 ,
                          eng_inv:Suffix_or .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme ;
     mmoon:hasRealization eng_inv:Suffix_s1 .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme ;
     mmoon:hasRealization eng_inv: EmptyMorph_u .
```

Next to morphs morphemes are the central resources within the domain of MLD. Morphemes are the smallest meaningful units of language and represent the conceptual side, i.e. the meaning, of morphs. Such data is not part of a lexical dataset and, thus, not modelled in the ontolex/decomp modules. The MMoOn Core vocabulary provides three `Morpheme` subclasses, i.e. `AtomicMorpheme`,

`FusionalMorpheme` and `EmptyMorpheme`, which can be used to represent morpheme resources. For illustration serves Example 11. If exactly one meaning is represented in a language, the morpheme instance is of the type `AtomicMorpheme` If, however, more than one meaning is represented and fused into one morph within a language, the morpheme instance is of the type `FusionalMorpheme`. Some theories of morphology assume morphemes that have no meaning. For representing such elements, e.g. the `EmptyMorph` instance *-u-* in the English adjective *factual*, an `EmptyMorpheme` instance can be created to account for the empty conceptual side of *-u-*[12].

Since morphemes are only meanings, i.e. mental representation of concepts, they are represented by abstract identities in order to be referable. This is done by `MorphemicGloss` instances which are provided for each of the 299 `Meaning` classes in the MMoOn Core model and which also apply to every language-specific schema instance derived from the Core model, e.g. `eng_schema:Singular mmoon:hasAbstractIdentity mmoon:MorphemicGloss_SG`. It has to be noted here, that the `hasMeaning` object subproperties can be used to describe `Word`, `Morph` and `Morpheme` resources in MMoOn Core, as has been shown in the Examples 6, 9 and 12 and seems to over-model the data. While these are just possibilities of describing the meaning of different linguistic elements, within a consistent MMoOn morpheme inventory it is sufficient to model the meaning on the `Morpheme` resources, because these are traceable through the data graph via the corresponding morphs to the word-forms and lexemes in which they occur.

Finally, a morph and its corresponding morpheme must be interrelated because they constitute a unity of a linguistic expression and its conceptualization. Example 13 illustrates the association between morphemes and morphs. The object property `hasRealization` which is inverse of `correspondsToMorpheme` is provided and links a morpheme to all morphs by which it is realized in a given language.

It has to be noted that so far – to my knowledge – no RDF dataset exists which contains morpheme resources as proposed in the MMoOn Core model. However, in linguistic field research and in the general practice of documenting the morphological level of languages, it is common to create interlinear glossed texts, which distinguish morph and morpheme resources in a similar way. While it might be effortful (but not impossible) to create morpheme resources as proposed in MMoOn from scratch or manually, the vocabulary could be useful for representing existing interlinear glossed text resources in MMoOn RDF.

5 Intersections and Issues of an OntoLex - MMoOn Alignment

As the previous sections illustrated, the conceptual overlap of the ontolex/decomp modules and the MMoOn Core model provides an auspicious basis for interconnecting both domain models. In order to align both vocabularies, several

[12]It depends on the choice of the dataset creator if empty morphemes are assumed. One could also assume a suffix *-ual* as being an allomorph to *-al*.

intersections of elements could be used to bring them into mutual agreement. Since ontology alignment and merging might cause "unforeseen implications" [5], this task should be solved together by the OntoLex and MMoOn community groups. Nonetheless, in what follows, elements are proposed which are assumed to be necessary for mapping in order to enable a consistent extension of OntoLex datasets with a MMoOn morpheme inventory and conversely.

1) `ontolex:LexicalEntry` and `mmoon:LexicalEntry`: These two classes are central in both domain models and are regarded as the the most important intersection because they are crucial for the interconnection of lexical entries and morph resources. The OWL property `owl:equivalentClass` could be an appropriate mapping choice, since it would allow to infer that all more specific `mmoon:LexicalEntry` subclasses are also subclasses of `ontolex:LexicalEntry`. With consideration of the use of MMoOn properties which have some of these subclasses in their domain and range restrictions, however, it is debatable if a stated equivalency between these two classes will be sufficient or if a separate mapping of each `mmoon:LexicalEntry` subclass might be required.

2) `ontolex:Affix` and `mmoon:Affix`: These two classes can be also mapped via `owl:equivalentClass`. This would allow to later classify `ontolex:Affix` resources for the more specific `mmoon:Affix` subclass types by remaining of the `ontolex:Affix` type at the same time.

3) `decomp:subterm` and `mmoon:isDerivedFrom`; `mmoon:isComposedOf`: The decomp module clearly favours the representation of compound words. Therefore, an interconnection of `mmoon:isDerivedFrom` and `mmoon:isComposedOf` as being subproperties of `decomp:subterm` would enable more specific interrelations of lexical entries in OntoLex if desired.

4) `ontolex:LexicalSense` and `mmoon:Sense`: A reuse of `ontolex:LexicalSense` resources for `mmoon:Stem` resources would facilitate the assignment of senses to stems a lot. Although, the `owl:equivalentClass` property could be used here as well, a more elegant solution would be the implementation of an axiom that automatically creates a link between the `ontolex:LexicalSense` resource (of a given `ontolex:LexicalEntry` instance) and the `mmoon:Stem` instance of which the lexical entry consists.

Even though more elements could be considered for an alignment, the proposed mappings already bear a significant impact for the use of present OntoLex and MMoOn datasets and advantages for future datasets as well. E.g. the considerable amount of linguistic categories and derivational meanings provided with MMoOn Core could be directly used for OntoLex. Moreover, the morphological segmentation of `ontolex:LexicalEntry` instances is easily describable with an aligned MMoOn Core model. Finally, the `ontolex:Affix` resources as part of a lexicon, would be enriched with information on the specific kind of affix in question, its interrelation to the lexical entries in which it occurs and the inflectional or derivational meaning it carries.

6 Conclusion

At the moment, the OntoLex and the MMoOn model coexist as two separate ontologies, even though both models exhibit a conceptual overlap in the representation of the LLD and MLD domains. This paper motivated an alignment of both models, since it could be shown that the ontolex/decomp modules are not sufficient to describe fine-grained MLD in such an extensive way as the the MMoOn Core model does. Therefore, the undertaken comparison of the capabilities of both models to represent MLD revealed intersecting elements of both vocabularies and proved that the MMoOn model is a suitable candidate for achieving extensibility of OntoLex datasets with MLD. Further, the paper pointed out intersecting elements for which mapping possibilities have been suggested and discussed. The aim of this paper was to propose a unification of both models. Now, it is up to the LLOD and OntoLex community to discuss and to decide whether the proposed alignment of these two linguistic domain models is desired and to work together on the realization of an OntoLex-MMoOn alignment. The author is convinced, that it would indeed enhance the exploitation of linguistic Linked Data in the Semantic Web world and would moreover contribute to the development of more coherent linguistic Linked Data datasets in general.

7 Acknowledgements

This papers research activities were partly supported and funded by grants from the EU's H2020 Programme ALIGNED (GA 644055) and the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) for the SmartDataWeb Project (GA-01MD15010B).

References

1. Booij, G.: The grammar of words: An introduction to linguistic morphology. Oxford University Press (2012)
2. Chavula, C. and Keet, C.M.: Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In: OWLED, pp. 61-72 (2014)
3. Cimiano, P., Buitelaar, P., McCrae, J. and Stintek, M.: LexInfo: A declarative model for the lexicon-ontology interface. In: Web Semantics: Science, Services and Agents on the World Wide Web. Elsevier, pp. 29–51 (2011)
4. Haspelmath, M. and Sims, A.: Uderstanding morphology. Routledge (2013)
5. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In IJCAI-2001 Workshop on ontologies and information sharing. pp. 53–62 (2001)
6. Klimek, B., Arndt, N., Krause, S. and Arndt, T.: Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (2016)
7. Lehmann, C.: Data in linguistics. In: The Linguistic Review. Vol. 21, pp. 175–210 (2004)

8. Lehmann, C.: Interlinear morphemic glossing. In: Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2, pp. 1834–1857 (2004)
9. McCrae, J. et al.: The lemon cookbook. (2010)
10. McCrae, J., Spohr, D. and Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In: The semantic web: research and applications, Springer, 2011, pp. 245–259 (2011)

2.4 Challenges for the Representation of Morphology in Ontology Lexicons

The development of the MMoOn Core ontology and the proposal to align it with the OntoLex-*lemon* model raised the awareness of the limitations of OntoLex-*lemon* for the domain of morphological data. However, instead of working on a standardised interconnection the W3C Ontology-Lexicon community group[1] decided to create a new Morphology Module that addresses the missing possibilities for modelling morphological language data with OntoLex-*lemon*. Due to the acquired knowledge in this field the author has been asked to lead the module creation development on which she agreed.

[P4] presents the interim results of this effort. It describes the challenges that come with the development of a model within a community of researchers that represent multiple sciences working with or creating morphological language data. Whereas structural interoperability of the resulting datasets will be accomplished through the Linked Data-based semantic modelling per se, the issue of "human interoperability" is still underestimated in the context of reaching cross-disciplinary usage goals. Furthermore, the current state of the Morphology Module, including newly created classes and properties as well as their integration into the other OntoLex-*lemon* modules, is presented. The differences between this emerging OntoLex-*lemon* Morphology Module and the MMoOn Core ontology regarding the target user group, scope and coverage are outlined as well.

[1] https://www.w3.org/community/ontolex/

Challenges for the Representation of Morphology in Ontology Lexicons

Bettina Klimek[1], John P. McCrae[2], Julia Bosque-Gil[3], Maxim Ionov[4], James K. Tauber[5], Christian Chiarcos[4]

[1] Institute for Applied Informatics (InfAI), Leipzig University
[2] Data Science Institute, National University of Ireland Galway
[3] Ontology Engineering Group, Universidad Politécnica de Madrid
[4] Goethe-Universität Frankfurt am Main
[5] Open Greek and Latin Project

Abstract

Recent years have experienced a growing trend in the publication of language resources as Linguistic Linked Data (LLD) to enhance their discovery, reuse and the interoperability of tools that consume language data. To this aim, the OntoLex-*lemon* model has emerged as a *de facto* standard to represent lexical data on the Web. However, traditional dictionaries contain a considerable amount of morphological information which is not straightforwardly representable as LLD within the current model. In order to fill this gap a new Morphology Module of OntoLex-*lemon* is currently being developed. This paper presents the results of this model as on-going work as well as the underlying challenges that emerged during the module development. Based on the MMoOn Core ontology, it aims to account for a wide range of morphological information, ranging from endings to derive whole paradigms to the decomposition and generation of lexical entries which is in compliance to other OntoLex-*lemon* modules and facilitates the encoding of complex morphological data in ontology lexicons.

Keywords: morphology; RDF; OntoLex-*lemon*; MmoOn; inflection; derivation

1. Introduction

Morphology is a vital and, in many languages, very sophisticated part of language, and as such it has been an important part of the work of lexicographers. In the traditional print form, morphological information is provided in brief abbreviated terms that can only be deciphered with significant knowledge of the language, however with the transformation of the dictionary to an electronic resource a re-imagining of the morphology information in a dictionary is certainly due. We base our work within the framework of the ontology-lexicon (McCrae et al., 2012; Cimiano et al., 2014) and in particular in that of the OntoLex-*lemon* model. This model has been used not only for the conversion of existing dictionaries (Khan et al., 2017; Borin et al., 2014; Bosque-Gil et al., 2015) but also for the development of new dictionaries (Gracia et al., 2017) as Linked Data (Chiarcos et al., 2013).

In this paper, we present the current modelling as well as the underlying challenges within the development of the Morphology Module for OntoLex-*lemon*, which extends the existing work by providing modelling for representing the morphology that is associated with the entries. In many cases, morphology is an important part of the language, for example in both German and Irish noun plurals are irregular and cannot be predicted from the stem alone, so many dictionaries, especially learners' dictionaries, list these irregular forms for most or all of the entries. Further, for languages such as the Romance ones, verbs may have many forms that are frequently irregularly or semi-irregularly derived, and learners' dictionaries for these languages also list many forms. However, as electronic dictionaries become of use not only to humans but also machines, it is necessary to provide all forms in a manner that can be readily processed by the latter. To this end, the Morphology Module covers not only the description of some forms of a lemma, but also allows the generation of all forms through morphological patterns, which corresponds to the idea of declensions or conjugations of an entry. Further, we base our model on the MMoOn Core ontology (Klimek, 2017), which has been designed to more generally represent morphology as a linguistic domain, and as such this module can handle a wide range of linguistic phenomena including distinctions between derivational and inflectional morphology, allomorphy, suppletion, simulfixes and transfixes among others. Moreover, this module is, as its name suggests, part of the overall model of OntoLex-*lemon* and as such can be integrated well with other parts of OntoLex-*lemon* and is consistent with its other semantic and syntactic modules.

The rest of this paper is structured as follows. In Section 2, we provide an example based illustration of the shortcomings of morphological data representation in traditional dictionaries. In Section 3 we provide background of the OntoLex-*lemon* model for readers, who are not familiar with it, which is followed by an overview of related work in Section 4. We then present the challenges of representing morphology within the OntoLex-*lemon* framework in Section 5 before presenting the current modelling state of our proposed model in Section 6. Finally we look into the further improvements that we plan for the module in Section 7, and present some conclusions in Section 8.

2. Morphological data in dictionaries and lexical databases

The treatment of morphology in dictionaries is a complex topic which is related to the lexicographic selection process (or lemma selection) (Schierholz, 2015), and the definition of the micro-structure of entries, i.e., the data model upon which the description (Hartmann, 2001) and layout (Atkins & Rundell, 2008) of each entry will be based, with different types or 'templates' being also considered, e.g. a typical noun-entry type (Abel, 2012).

Opacity, frequency and predictability of form and meaning in words were aspects that had to be considered when deciding whether a complex lexeme or compound word should be contained in a dictionary or not (De Caluwe & Taeldeman, 2003), but

dictionaries and lexicographic traditions, in general, vary substantially. For example, derivational affixes have often received main entry status, with differences from dictionary to dictionary in their description: from dictionaries that identify them just as suffixes, to dictionaries that also point to their derivational or inflectional use (Alsina & DeCesaris, 1998).

Different approaches to lexicography also play a role in these various representations of morphological data. Linguistics-oriented dictionaries, guided by a linguistic theory for morphology and its terms, contrast with function-theoretic based (or communicative) works which are focused mainly on the morphological information needs of users in specific situations (Swanepoel, 2015; Bergenholtz & Tarp, 2005).

This context leads to a heterogeneous landscape when it comes to analysing the morphological description provided in dictionaries. Most traditional dictionaries do not cover morphological information extensively: usually, the morphological description of the lexical entry is limited to the list of the word forms that allow users to identify the morphological pattern to which the entry adheres, and hence generate the paradigm by themselves. Following this, word-forms that can be formed regularly are not listed. Moreover, the description of these 'reduced' inflection lists is often minimal on the assumption of users being familiar with the lexicographic tradition of the object language. For example, users of a German dictionary familiar with the German language easily interpret the description ***Na · me*** *der; -ns, -n* to refer to the gender of the entry, and its genitive singular and nominative plural endings. Other dictionaries, such as The K Dictionaries Multilingual Global Series[1], provide groups of word-forms inflected for case and number, along with the ending that is displayed in the user interface, as illustrated in Example 1.1.

This is similarly the case for Ancient Greek dictionaries, where noun entries will typically list the nominative singular form, the genitive singular ending, and the article (indicating the gender). This assumes the reader is able to work out the stem by comparing the nominative form with the abbreviated genitive ending. This, in combination with the gender, is then generally enough to produce other forms of the nominal paradigm. Additional forms of the noun are generally not given in the entry unless deemed impossible or non-obvious to produce from the standard information given.

For verbs it also very common to find verbal paradigms as a reference in the appendix of dictionaries. For example, Figure 1 shows the paradigm of the verb *amar* 'to love' as an example of a verb that inflects according to the 1st conjugation pattern in Spanish[2]. Even though such tables contain all forms of a lemma, the underlying morphological

[1] https://www.lexicala.com/resources#dictionaries

[2] http://www.rae.es/diccionario-panhispanico-de-dudas/apendices/modelos-de-conjugacion-verbal#advertencias, last accessed on 05.06.2019.

structure separating the stems from the regular and productive inflectional suffixes remains again implicit.

```
<HeadwordBlock>
  <HeadwordCtn>
    <Headword>Stipendiat</Headword> [...]
    <GrammaticalGender value="masculine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <HeadwordCtn>
    <Headword>Stipendiatin</Headword> [...]
    <GrammaticalGender value="feminine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiatin</Inflection>
        <Display>-</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiatinnen</Inflection
        >
        <Display>-nen</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <PartOfSpeech value="noun" />
</HeadwordBlock>
```

Example 1.1: An extract of the entry *Stipendiat* ‘scholarship holder’ from the K Dictionaries Global Series German Dictionary.

1. AMAR		Verbo modelo de la 1.ª conjugación		
INDICATIVO				
TIEMPOS SIMPLES				
presente	**pret. imperfecto / copretérito**	**pret. perfecto simple / pretérito**	**futuro simple / futuro**	**condicional simple / pospretérito**
amo	amaba	amé	amaré	amaría
amas (amás)	amabas	amaste	amarás	amarías
ama	amaba	amó	amará	amaría
amamos	amábamos	amamos	amaremos	amaríamos
amáis	amabais	amasteis	amaréis	amaríais
aman	amaban	amaron	amarán	amarían
TIEMPOS COMPUESTOS				
pret. perfecto compuesto / antepresente	**pret. pluscuamperfecto / antecopretérito**	**pret. anterior / antepretérito**	**futuro compuesto / antefuturo**	**condicional compuesto / antepospretérito**
he amado	había amado	hube amado	habré amado	habría amado
has amado	habías amado	hubiste amado	habrás amado	habrías amado
ha amado	había amado	hubo amado	habrá amado	habría amado
hemos amado	habíamos amado	hubimos amado	habremos amado	habríamos amado
habéis amado	habíais amado	hubisteis amado	habréis amado	habríais amado
han amado	habían amado	hubieron amado	habrán amado	habrían amado

Figure 1: Table of the inflectional paradigm of the verb *amar* 'to love' from the *Diccionario Panhispánico de Dudas* (Real Academia Española and Asociación de Academias de la Lengua Española, 2005).

From the examples just illustrated, it becomes clear that all the common approaches regarding the representation of morphological data rely highly on the implicit knowledge of the dictionary user about the language. As a consequence, morphological data varies greatly concerning their amount, their way of representation and interconnection to the relevant element they are contained in, i.e. the lemma or a form in a paradigm.

3. Overview of OntoLex-*lemon*

The OntoLex-*lemon* model[3] has been under development for several years and was originally based on the combination of the three pre-existing models (LingInfo (Buitelaar et al., 2006), LexOnto (Cimiano et al., 2007), LIR (Montiel-Ponsoda et al., 2011)) that were combined into a single model (lemon) by the EU project Monnet and later extended into the OntoLex-*lemon* model by the Ontology Lexicon Community

[3] The full specification can be consulted here: https://www.w3.org/2016/05/ontolex/.

Group[4]. This model was developed around five basic principles: 1) it would be an RDF model that used the Web Ontology Language (OWL) (McGuinness, Van Harmelen, et al., 2004) for its semantics; 2) it would support multilinguality and avoid language-specific assumptions that might affect the applicability of the model to other languages; 3) it would use the principle of 'semantics by reference' as a basic semantic model (Cimiano et al., 2013); 4) it would embrace openness in being free of any financial costs or licensing as well as allowing contributions from any interested party, and 5) relevant standards and models would be reused wherever appropriate. This led to the core model that is depicted in Figure 2, which is based around a lexical entry, composed of a number of forms and a number of senses, which can then be linked to either lexical concepts or entities in an ontology.

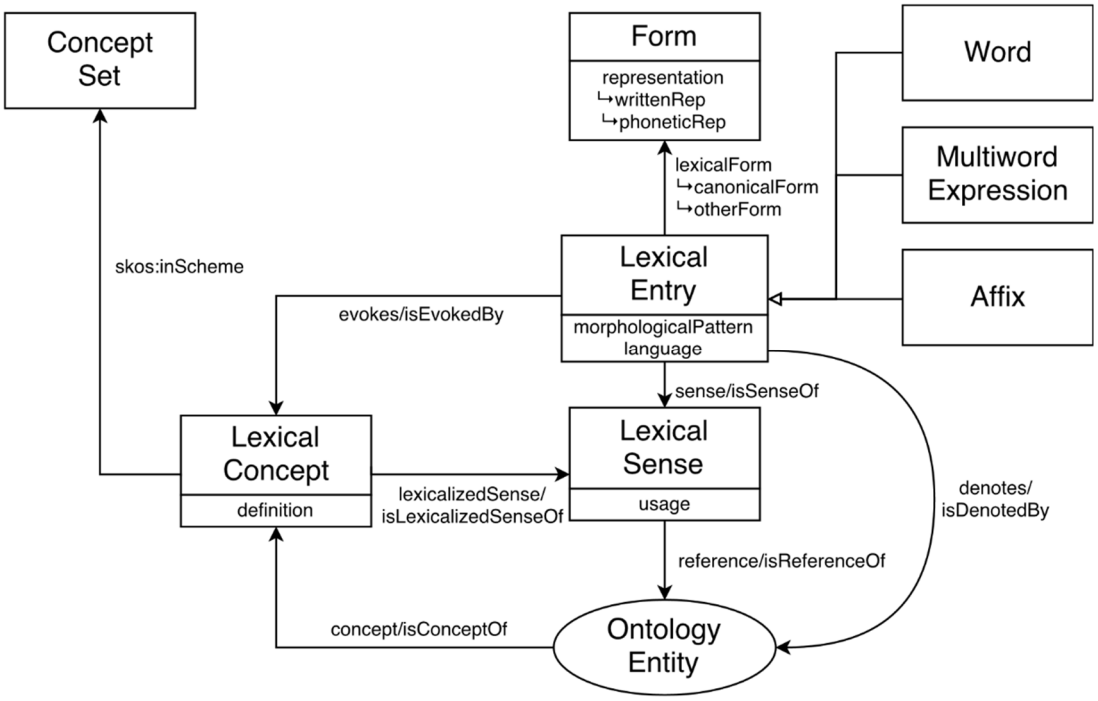


Figure 2: The core model of OntoLex-*lemon*.

In addition to this core, that is often also called "ontolex", there were four further modules developed in the initial release of the model:

Syntax and Semantics (synsem) This module describes how syntactic frames may be modelled and how they can be mapped to ontology structures,

Decomposition (decomp) The decomposition of multiword expressions and compound terms is described by this module,

Variation and Translation (vartrans) Modelling of translations and other kinds of

[4] https://www.w3.org/community/ontolex/

relations are provided by this module,

Linguistic Metadata (lime) This module provides metadata about the lexicon and the ontology and how this may be used to encourage interoperability between resources.

In addition, since then the group has continued to develop modules to extend the usefulness and applications of the model. One such extension, the recently released Lexicography Module (Bosque-Gil et al., 2017), has provided features for representing dictionaries in ways that are more compatible with traditional print dictionary forms. Other modules are in development, in particular this one along with a module for representing frequencies, attestations and corpus information[5], and a module for etymological and diachronic information (Khan, 2018).

Since its development, the OntoLex-*lemon* model has been extensively used for representing a vast amount of different lexical data: In addition to traditional dictionary data mentioned in Section 1, it has been applied to lexical databases like WordNet (McCrae et al., 2014), etymological resources (Chiarcos et al., 2016; Khan, 2018), and domain-specific lexicons (Bellandi et al., 2018).

4. Related work

The emerging OntoLex-*lemon* Morphology Module described in this paper aims to enable the representation of the morphological elements and processes that are involved in the decomposition and generation of lexical data (of both lexemes and their word-forms) by overcoming the representational limitations of traditional dictionaries as outlined in Section 2 and within the technical realm and the design principles of the overall OntoLex*lemon* model introduced in the previous section. Since the emergence of the (multilingual) Semantic Web in the early 2000s, several ontologies emerged from the lexicography, language resource and language documentation communities that already contain the modelling of morphological language data to some extent. Here we briefly describe some of these ontologies that are considered the most relevant with regard to the morphological data they allow to represent, together with an explanation to what extent they could or why they could not be reused within the OntoLex-*lemon* Morphology Module.

In the early development of the OntoLex-*lemon* model, its priorities have been on lexicalizing ontologies and knowledge bases. This was accompanied by a natural focus on lexical semantics, i.e., multilingual labels for the same concept, and, here, the original contribution of Monnet-Lemon, the predecessor of OntoLex-*lemon* has been to complement such labels with morphosyntactic information in order to facilitate context-adequate lexicalization. Morphology was only considered in the form of morphosyntax, i.e. inflectional features as well as the possibility to provide the adequate form for these.

[5] https://acoli-repo.github.io/ontolex-frac/

The current OntoLex-*lemon* representation of morphological information can complement ontology concepts with morphosyntactic categories (part of speech, a property of a lexical entry), and provide different forms with different morphosyntactic features (e.g., gender, case, number, etc.) Neither derivational morphology nor morphological information beyond the specification of grammatical features was expressible with this model, and lexicalizations of the same concept with different parts of speech required independent lexical entries, without being able to represent the systematic relations on the level of form and meaning that hold between them.

OntoLex-*lemon* does not provide any vocabulary of grammatical features, instead, it endorses the reuse of the existing ontologies and vocabularies for linguistic annotations, most notably, ISOcat, GOLD, OLiA, and LexInfo. ISOcat, a shared repository for linguistic concepts, features and data structures, was developed as a successor of the ISO Data Category Registry (DCR), originally designed as an RDF-based knowledge graph (Ide & Romary, 2004) and is built on XML technologies and resolvable URIs (Kemps-Snijders et al., 2009). ISOcat was a semistructured resource populated in a bottom-up process, so that it did not provide formal and consistent vocabulary, but its subsets became an important source of knowledge that more consolidated domain vocabularies described here drew from. GOLD, one of the first attempts in creating a linguistic ontology (Farrar & Langendoen, 2003), and OLiA (Chiarcos & Sukhareva, 2015) were designed primarily as solutions to harmonize linguistic categories and make markup schemes interoperable. In OLiA this is achieved by linking the hierarchy of abstract grammatical categories which constitutes the reference model with specific markup schemas that can vary for resources and languages.

Despite their interoperability and applicability to a vast amount of linguistic data, these ontologies are primarily focused on providing labels for the categories and lack the expressibility to represent morphosyntactic information.

LexInfo is an inventory containing various types, values and properties to describe linguistic categories (Cimiano et al., 2011). It is partially derived from ISOcat and is often used to represent linguistic annotations in Ontolex-*lemon* (however, this is not a requirement). Even though it covers certain aspects of morphology, it has a focus on inflectional morphology whereas it lacks expressiveness in describing derivational morphology.

Finally, the last relevant model is the MMoOn Core ontology[6] (Klimek et al., 2016). It is currently the only existing comprehensive domain ontology for the linguistic area of morphological language data. As such it is highly specialized and far more-fine grained than the desired modelling of the OntoLex-*lemon* Morphology Module requires. It contains, among other aspects, an extensive modelling of linguistic meanings, including derivational meanings in addition to grammatical categories. It also differentiates

[6] https://mmoon.org/core

between morph and morpheme resources and comes with a set of nearly 300 morphemic glosses to provide sufficient expressivity to represent morphological data contained in Flex or Toolbox datasets. At the same time, a specification of lexical data is not provided in MMoOn Core because this ontology was envisaged to be used complementary to OntoLex-*lemon*. Therefore, there is only one existing interconnection of the two domain ontologies so far, i.e. an established subclass relation between the two classes mmoon:LexicalEntry and ontolex:LexicalEntry. A more extensive ontology alignment has been thus far only proposed from the MMoOn Core perspective (Klimek, 2017) and might be considered for future implementation. Once the OntoLex-*lemon* Morphology Module will be officially released, further alignment options might be realized. Even though the MMoOn Core ontology exceeds by far the modelling needs of the Morphology Module, it served as a modelling template since the creation of MMoOn Core was initially motivated to fill the gap of representing morphological language data in OntoLex-*lemon* that still existed back then. So far, certain types of affix classes, e.g. mmoon:Simulfix) as well as the two object properties mmoon:consistsOf and mmoon:meaning have been reused in the OntoLex-*lemon* module, although only in an inspirational manner. These classes and properties are defined and integrated slightly differently within the morphology module and should not be confused as long as no explicit alignment has been implemented.

From this review of relevant existing ontologies it can be concluded that the emerging OntoLex-*lemon* morphology module adheres to the Semantic Web best practice of reusing existing vocabularies. Since none of the presented ontologies sufficiently satisfies the representation needs of morphological data in particular with regard to lexical data so far, the Morphology Module will adequately fill this gap. Furthermore, as a result of the outlined reuse choices, the Morphology Module could be kept user-friendly and manageable by replacing the usually necessary modelling of grammatical categories and morphological meanings of morph resources with the recommendation to use existing vocabularies instead, and also linguistically accurate because it is influenced by the more precise MMoOn Core domain ontology.

5. Challenges in developing a Morphology Module extension

Creating a descriptive modelling foundation for representing lexical data entails several design choices that directly affect the usability of the model. This does not only hold for ontology lexicons, but also for lexicon models in general. In what follows, challenges that arose during the development of the morphology module for OntoLex*lemon* will be outlined. With the ongoing development of modules, these issues gain increasing importance and can serve as orientation points of consideration for future module extension development efforts.

5.1 Scope and coverage

Description: The first question that arises when a new ontology is being created is who should use it for what purpose? As illustrated in Section 2, morphological information is highly implicit in the landscape of traditional dictionaries. However, along with the liberation from the limits of print dictionaries came almost unlimited possibilities of lexicographic data compilation in eLexicography, which are yet again broadened by the possibilities of the Linked Data paradigm. While some lexicographers only like to digitize a printed dictionary into Linked Data using RDF, others aim at transforming their already more fine-grained lexical databases and intend to use the resulting RDF dataset to generate more lexicographic content out of it, e.g. to generate inflectional paradigms including full word-forms together with the underlying morpho-phonological formation rules.

Modelling Choice: In line with OntoLex-*lemon* model, the Morphology Module also aims at being applicable for everyone working with lexicographic content who either focuses on the transformation of traditional dictionary data into RDF or on the conversion of more structured computational lexical data. Accordingly, the scope of the module is divided into two main parts: 1) enabling the representation of elements that are involved in the decomposition of lexical entries and word-forms, and 2) enabling the representation of building patterns that are involved in the formation of lexical entries and word-forms. A fine-grained description of phonological processes that are involved in any kind of stem or word formation on the phoneme level is, however, excluded and not representable with this Morphology Module. Only the elements between the lexical entry and the morph levels will be covered.

5.2 Consistency

Description: The ontolex and decomp modules of OntoLex-*lemon* already contain various classes and properties that can be used to describe morphological data. The ontolex:Affix and decomp:Component classes for instance already exist to represent sub-word units and can be put into relation to the lexical entries in which they are contained via properties like decomp:correspondsTo or decomp:subterm. Due to the widespread usage of OntoLex-*lemon*, the development of the Morphology Module is challenged with creating the necessary missing vocabulary by taking the existing classes and properties into account, while ensuring backwards compatibility at the same time.

Modelling Choice: Due to the incremental approach of developing the module for morphology and also future OntoLex-*lemon* extensions, it is inevitable to deal with overlapping existent vocabulary. Therefore, the OntoLex Community Group agreed to aim for the goal of reaching consistency by reusing as much of the existent vocabulary as possible and minimize duplication that results from creating similar classes and properties. Specifically, this entails that suitable existent vocabulary can be adapted as

long as the changes made are a) only additions to domain and range restrictions of properties or b) adaptions in the rdfs:comment description to broaden the applicability of classes. In this way, existing vocabulary can be coherently integrated into later developed modules while simultaneously preserving already established functionalities.

5.3 Terminological ambiguity

Description: During the module development process it turned out that one of the greatest challenges is to unambiguously define the terminology that is used to label the classes and properties of the new vocabulary. As intended, the widely set scope of the Morphology Module presented in Section 5.1 attracts the use of the module for various user groups which are, however, also coming from different terminological backgrounds. The understanding and usage of linguistic concepts like *morph* or *root* diverge considerably depending on whether the user of the module is, for example, a traditional linguist, a computer linguist or a lexicographer managing data for specific languages. This entails a high risk of an inappropriate usage of the ontological vocabulary that might result in an unintentional wrong data representation the user is generally not even aware of.

Modelling Choice: While the human-readable definition of ontology elements is defined within the rdfs:comment, the underlying machine-processable semantics are determined by implications and restrictions for an element and its relation to other elements of the ontology. For the computational processing of the data the former is not relevant, whereas the latter is formally fixed and unambiguous. What matters is the consistent usage of the vocabulary according to the ontologically defined semantics, notwithstanding that a user would have chosen a different label for an element. Moreover, providing a definition that is interpreted in the same way by all users is almost impossible. Therefore, the rdfs:comment descriptions of classes and properties are discussed and refined until the highest possible consensus is reached. In addition to that, the Morphology Module specification that will be published together with the release of the module contains usage examples and recommendations that support a shared understanding to ensure the consistent application of the module vocabulary.

6. Current state of the Morphology Module

6.1 Summary of the current state

The development of the Morphology Module is an ongoing joint effort by members of the OntoLex Community Group that started in November 2018. This paper presents the intermediate results which have been reached and the state of the module as of May 2019. The documentation creation process reflecting the discussions of the scope, identified representation needs and modelling steps can be consulted on the respective

OntoLex Wiki page[7]. It contains the outcomes as well as the links to the minutes of the regular calls that have been held.

So far, half of the defined scope for the Morphology Module (cf. Section 5.1) could be modelled. In particular this includes the fist scope, i.e. the representation of the decomposition of ontolex:LexicalEntry and ontolex:Form resources. An overview illustrating the resulting model structure is shown in Figure 3. The second scope of representing the automatic generation of entries and forms from morph resources is still in an early development stage and, hence, will not be addressed in detail in this paper. The model in Figure 3 displays how the Morphology Module is embedded within the existing OntoLex-*lemon* vocabulary it relates to. Classes and properties written in blue indicate the new vocabulary that is specified with the prefix morph with the class morph:Morph building the centre of the module. The two object properties decomp:subterm and decomp:correspondsTo are also represented in blue, thus, highlighting that these are vocabulary elements that will have to be adjusted by extending their ranges (as explained in Section 5.2) to arrive at an overall OntoLex-*lemon* model consistency. It has to be noted that the presented Morphology Module is not officially published yet and, therefore, not usable at this current stage. However, it can be assumed that the vocabulary elements that are described in the next Section will remain very close to their final published module specification.

6.2 New classes and properties

In order to solve the presented challenges outlined in Section 5, new classes and properties had to be developed for the Morphology Module. Altogether eleven new classes and seven object properties have been implemented into the modelling so far. In doing so, central concepts of the domain of morphological data could be reused from the OntoLex-*lemon* vocabulary, and a considerable reduction of overlap between the new and the existing vocabulary could be reached. The ontolex:Form class, for instance, was already appropriate to represent all forms of a lexical entry, which are crucial elements for the description of the segmentation of words. Table 1 and Table 2 present an overview of the module vocabulary with the definitions and restrictions that have been defined for all new classes and properties.

The morph:Morph class builds the centre of the module and is divided into six subclasses. As a result it will be possible to specify root, stem and certain affix types. The prominent affixes, i.e. prefix, suffix, infix and circumfix, are, however not part of the vocabulary because these can be reused from other ontologies such as LexInfo. The treatment and function of the ontolex:Affix class was highly debated for its potential re-usability. Since this class is a subclass of ontolex:LexicalEntry it cannot be used to represent bound morphs that are inflectional, because those are usually not described

[7] https://www.w3.org/community/ontolex/wiki/Morphology

as headwords in lexical databases or dictionaries. In order to avoid uncertainty within the classification of inflectional and derivational affixes, the morph:AffixMorph class has been created. Affixes that should be represented as lexical entries can be described with ontolex:Affix, whereas those that cannot should be described with the morph:AffixMorph class, regardless of their derivational or inflectional nature. Moreover, an explicit declaration for these two morphological functions has been enabled by providing the object property morph:hasMorphStatus and the class morph:MorphValue that already contains the two individuals morph:inflectional and morph:derivational ready for use.

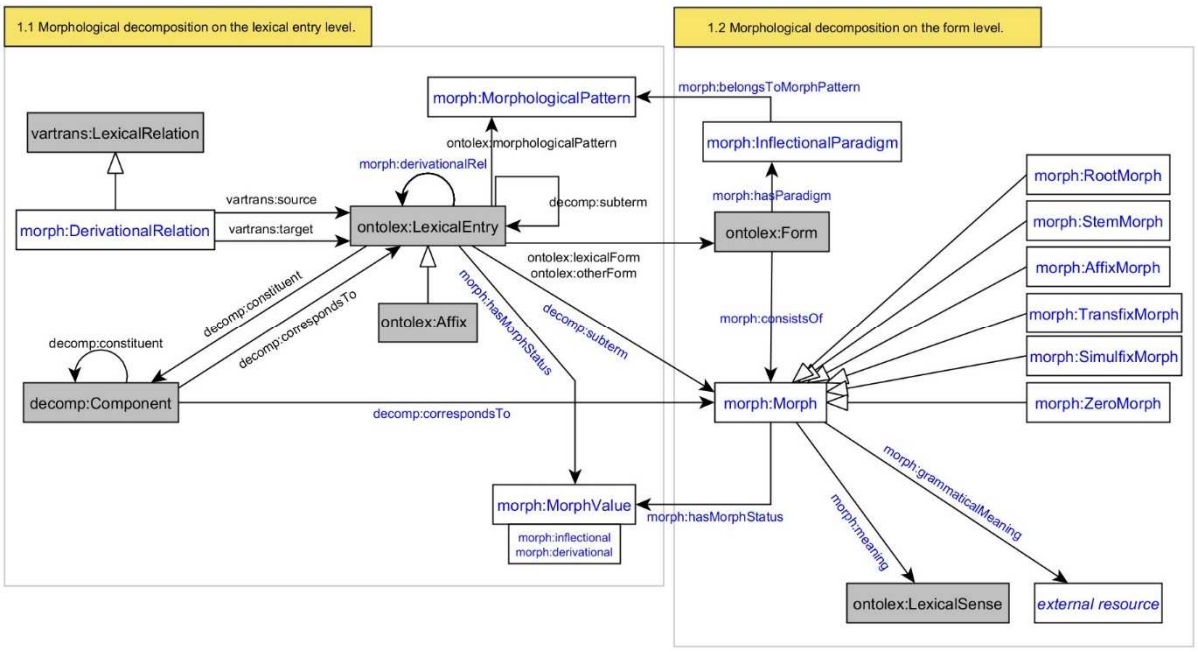


Figure 3: Current proposal of the Ontolex-*lemon* morphology module.

Since the derivational morphs of a derived lexical entry are now explicitly representable within the Morphology Module, a possibility to state that one derived lexical entry is derived from another lexical entry should be provided. This has been achieved by creating the class morph:DerivationalRelation that is defined as a subclass of vartrans:LexicalRelation. Therefore, it inherits the same domain and range restrictions which mean it can represent the direction of the derivational relation between two lexical entries, i.e. one can explicate that one derived lexical entry is derived by a specific derivational relation from another lexical entry. Furthermore, more generically all lexical entries that can be created through a derivational relation from another lexical entry can be expressed by using the object property morph:derivationalRel. Examples illustrating the use of this class and this property will be provided in Section 6.3.1.

Class Name	**Definition**	**Class Relation**
Morph	A morph is a concrete primitive element of morphological analysis.	owl:disjointWith ontolex:LexicalEntry
RootMorph	A morph that constitutes the semantic nucleus of a stem. It cannot be further segmented and is often not specified for a part of speech.	rdfs:subclassOf morph:Morph
StemMorph	The stem is the morph to which inflectional marking applies.	rdfs:subclassOf morph:Morph
AffixMorph	An affix is a bound segmental morph.	rdfs:subclassOf morph:Morph
TransfixMorph	A transfix is a discontinuous affix.	rdfs:subclassOf morph:Morph
SimulfixMorph	A simulfix is a bound morph that entails a change or replacement of vowels or consonants (usually vowels) which changes the meaning of a word, e.g. *eat* in past tense becomes *ate*.	rdfs:subclassOf morph:Morph
ZeroMorph	A morph that that corresponds to no overt form, i.e. orthographic or phonetic representation.	rdfs:subclassOf morph:Morph
MorphValue	The value of a morph states the relationship that holds between the morph and the forms or lexical entries in which it can occur.	class instances: morph:inflectional morph:derivational
DerivationalRelation	A 'derivational relation' is a lexical relation that relates two lexical entries by means of a derivational affix.	rdfs:subclassOf vartrans:LexicalRelation
MorphologicalPattern	The morphological pattern states the inflectional, derivational or compositional building pattern that applies to a lexical entry.	none
InflectionalParadigm	A structured set of inflected forms according to specific grammatical parameters.	none

Table 1: Overview of new classes of the Morphology Module.

With the foresight to enable also the automatic generation of ontolex:LexicalEntry resources from given morph:Morph and ontolex:Affix resources, the necessary conceptual frame has been modelled already. Figure 3 shows that the existing ontolex:morphologicalPattern object property was an initial proposal but remained under specified due to the non-existent Morphology Module at the point of its creation. This lack of expressivity has been now resolved by creating the two classes morph:MorphologicalPattern and morph:InflectionalParadigm which interrelate

ontolex:LexicalEntry and ontolex:Form within the graph structure of the module via the two established object properties morph:hasParadigm and morph:belongsToMorphPattern. Even though the specific usage of this part of the module is not sufficiently attested yet, the example for it provided in Section 6.3 illustrates the intended utilization.

As a central component of the morphological data domain the representation of the meaning of morph:Morph resources had to be modelled as well. Therefore, the two object properties morph:meaning and morph:grammaticalMeaning have been implemented in the module. The underlying concepts of morph:StemMorph and morph:RootMorph resources can be expressed by the former property by pointing to a ontolex:LexicalSense resource and the grammatical categories that are encoded in resources that represent grammatical morphs, usually bound affixes, can be expressed by pointing to an external resource. As already mentioned, the creation of an extensive modelling of possible linguistic categories has been considered to be out of scope for this module, and it is recommended to reuse existing vocabulary elements, e.g. from LexInfo, instead. The possible lack of a grammatical catogory in any existing ontology can be then compensated by using the morph:grammaticalMeaning property alternatively together with a newly created vocabulary.

Property Name	**Definition**	**Restrictions**
derivationalRel	The property relates two lexical entries that stand in some derivational relation.	domain: ontolex:LexicalEntry ontolex:LexicalEntry
consistsOf	This property states into which Morph resources a Form resource can be segmented.	domain: ontolex:Form morph:Morph
hasMorphStatus	The property states whether a morphological element functions as inflectional or derivational.	domain: morph:Morph, ontolex:Affix morph:MorphValue
hasParadigm	This property assigns a form to an inflectional paradigm.	domain: ontolex:Form morph:InflectionalParadigm
belongsToMorphPattern	This property assigns an inflectional pattern of a form as belonging to a morphological pattern of a lexical entry.	domain: morph:InflectionalParadigm morph:MorphologicalPattern
meaning	This property assigns a lexical sense to a morph resource.	domain: morph:Morph ontolex:LexicalSense
grammaticalMeaning	This property assigns a grammatical meaning to a morph resource.	domain: morph:Morph

Table 2: Overview of new object properties of the Morphology Module.

Finally, a relation was needed that states that an ontolex:Form resource consists of morph:Morph resources analogously to the ontolex:constituent object property that interrelates ontolex:LexicalEntry resources and decomp:Component resources. This relation manifests itself in the object property morph:consistsOf which is used to identify the segmentable morphs of inflected words, whereas ontolex:constitutent can identify the lexical parts of derived or compounded words. By further extending the range of ontolex:correspondsTo and ontolex:subterm for the class morph:Morph it is even possible to identify inflectional affixes within complex lexical entries. This is a particularly useful functionality of the morphology module for many languages that involve the expression of an inflectional morph in the process of word-formation. German nominal compounds, for example, can consist of some linking morph that can be identified as a case marking morph (or depending on the underlying linguistic theory as a zero morph), e.g. as in *Haushalt-s-kasse*, 'household-GEN-budget'.

6.3 Representing morphological decomposition

In what follows the usage of the introduced vocabulary of the Morphology Module will be illustrated by the example displayed in Figure 4. It shows the graph modelling evolving around the English noun *speaker*, including all the properties, classes and instances that are involved. For better understandability the graph is reduced to the representation of only one derived lexical entry, i.e. the adjective *speakerless* and only two word-forms of *speaker*, assuming that there are more. All boxes highlighted in yellow represent the new classes of the Morphology Module vocabulary.

6.3.1 On the lexical entry level

Looking at the resource :lex_speaker_n as the subject of this graph clarifies which morphological information can be explicated by creating the following statements:

1) It consists of two constituents which are decomp:Component resources which again can be said to correspond to another ontolex:LexicalEntry and a morph:AffixMorph resource, i.e. the verb :lex_speak_v and the derivational suffix :suffix_er. This suffix has been specified with the value morph:derivational and the ontolex:LexicalSense :agentNominalizer. This modelling indicates that in this example dataset this derivational suffix *-er* is explicitly not a lexical entry but could, however, be easily turned into one by changing its type assertion to ontolex:Affix.

2) It can be created with the morphological pattern :pattern_CommonNouns. As mentioned already, this is technically not implemented yet but it is intended to use the two decomp:Component resources :component_speak and :component_er for this purpose.

3) It can be linked to other lexical entries by using the morph:derivationalRel property in order to state which other derived words can be derived from :lex_speaker_n. This is, however, only a very generic statement but one that is often found in lexical or dictionary data.

Finally, the statement in 3) can be specified in a fourth statement by turning :lex_speaker_n into an object of a statement that describes it as the target of the derivational relation :derivRel_speaker_AgentNoun. While the property in statement 3) just states that there is some derivational relation between two ontolex:LexicalEntry resources, triples with a morph:DerivationalRelation instance in the subject position explicitly interlink the source lexical entry and the target lexical entry for which a unique derivational relation holds.

6.3.2 On the form level

The interconnection between lexical entries and the forms that can be built from them has been already established within OntoLex-*lemon* with the ontolex:otherForm property and has been, therefore, used in this example accordingly to relate the two forms :form_speakers1 and :form_speakers2 to the lexical entry :lex_speaker_n.

Considering these two instances as the subjects when consulting Figure 4 makes it possible to create the following statements about them:

1) They are both specified to belong to the inflectional paradigm :paradigm_NounInflecion. This paradigm defines the grammatical form variants of the ontolex:Form resources, i.e. case and number, and is itself assigned to the overall building pattern :pattern_CommonNouns for ontolex:LexicalEntry resources that are nouns like :lex_speaker_n.

2) They are both segmentable into morph:Morph resources that are stated with the morph:consistsOf property. As it is clear from Figure 4, they both share the same morph:StemMorph resource but consist of two different morph:SuffixMorph resources.

In addition to that, the three morphs :stem_speaker_n, :suffix_s1 and :suffix_s2 can be further specified for their meanings by pointing to ontolex:LexicalSense instances and grammatical values for the linguistic category case reused from the LexInfo vocabulary. It is essentially due to this enabled decomposition chain that makes it possible to not only identify, specify and interrelate all meaningful sub-word units but also the lexical entries and forms contained in lexical data, that all these elements can be disambiguated and described within a dataset modelled with the Morphology Module and OntoLex-*lemon*.

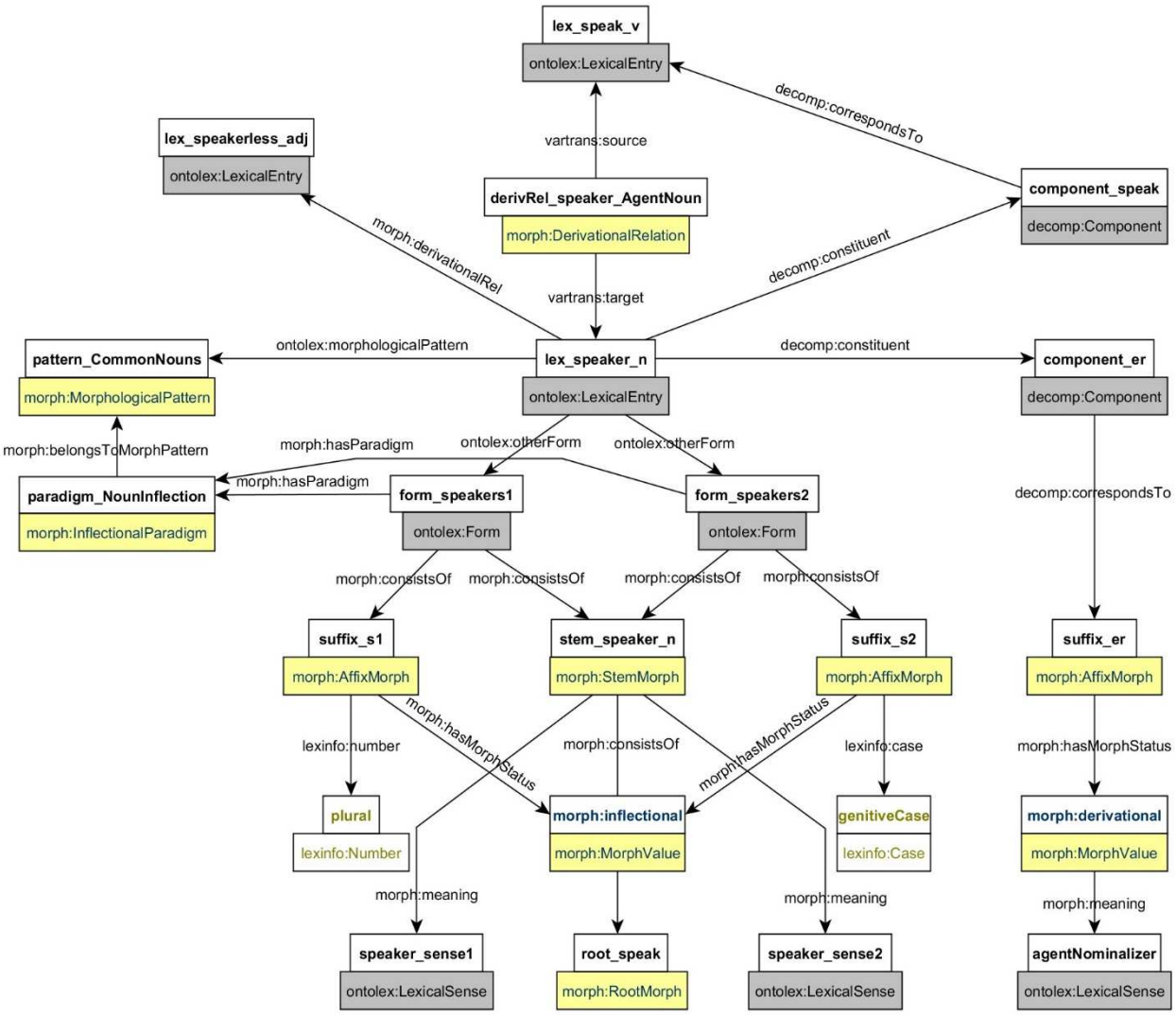


Figure 4: Graph representation for the example entry :lex_speaker_n.

7. Future work

Even though the modelling outcomes presented here have been largely agreed upon, several issues remain open for future work. Due to the various linguistic backgrounds of the OntoLex Community Group members some desired implementation options have been raised that might be still realized and included within the final Morphology Module specification. The following three features have been proposed for additional realization and are still under discussion:

1) **Morphemic glosses:** Since interlinear glossed text language data is an emerging source of lexical data that can be also represented in RDF, interest has been indicated to include the representation of morphemic glosses. So far it has been discussed if a modelling of glosses would exceed the scope of the Morphology Module, while the option to provide a shallow modelling with an

alignment to the MMoOn Core vocabulary that already provides a representation of glosses is also considered.

2) **Ordering:** For some highly polysynthetic and morphology-rich languages it is desirable to have a more precise representation of the internal morphological structure of lexical entries and forms. Therefore, it has been decided that a more expressive possibility for representing the position and ordering of morphs should be implemented to be available next to the currently used but very inexpressive rdfs:list object property. Proposals for that have been already made, but no agreement has been reached yet.

3) **Multiple segmentations:** Taking into account that a lexical dataset created based on the Morphology Module could be also applied in the context of computational linguistics, the processability of this data for machines might require the representation of more than one possible segmentation strategy. Allowing for the explication of that would be also interesting for linguists who want to document and analyse competing segmentations of words in their research.

In addition to these yet unrealized features it is necessary to focus on the refinement of the definitions of the newly created vocabulary elements. The exchanges within the community group have revealed that some of the presented rdfs:comment information is not precise enough and might lead to misunderstandings. In order to avoid misunderstandings in the usage of the vocabulary, time and attention will be invested again to resolve currently ambiguous or unclear definitions.

Furthermore, the second part of the Morphology Module that will enable the generation of forms with existing productive morphs in a dataset is also a part of the future work. However, the modelling is envisaged to produce lexical entries and forms based on patterns and paradigms, including also discontinuous morphs like transfixes and infixes. As it turned out in previous discussions such a formal representation is not trivial to model, especially with regard to the aim to be language-independently applicable.

8. Conclusion

To summarize, the current state of the Ontolex-*lemon* Morphology Module has been presented. The created vocabulary has been introduced and its usage illustrated. From that it becomes clear that the new module overcomes the limitations of the current representation of morphological data contained in traditional dictionaries by enabling the explication of formerly implicit information. With the Morphology Module modelled so far it is possible to represent the decomposition of lexical entries and forms with regard to both their derivational and inflectional morphs and underlying building patterns.

Furthermore, the challenges that arose from integrating the module into the existing

Ontolex-*lemon* model have been explained and design choices have been supported. It has been also shown that the module applies to existing Semantic Web standards by reusing relevant existing ontologies within its framework.

The remaining open issues have been presented and will be addressed in future work in order to arrive at the release of the final Morphology Module specification.

9. Acknowledgements

John McCrae is supported in part by a research grant from Science Foundation Ireland, cofunded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289, as well as by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure) and 825182 (Prêt-à-LLOD).

Julia Bosque-Gil is supported by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program.

Maxim Ionov and Christian Chiarcos are supported by the German Ministry for Education and Research (BMBF) through a project Linked Open Dictionaries (LiODi, 2015-2020) as a part of an Early Career Research Group on eHumanities.

10. References

Abel, A. (2012). Dictionary Writing Systems and Beyond. In S. Granger and M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. Chap. 5, pp. 86–106.

Alsina, V. & DeCesaris, J. (1998). *Morphological structure and lexicographic definitions: The case of -ful and -like*.

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

Bellandi, A., Giovannetti, E. & Weingart, A. (2018). Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* 9.3, p. 52.

Bergenholtz, H. & Tarp, S. (2005). Dictionaries and inflectional morphology. In *Encyclopedia of Language and Linguistics*. Pergamon Press, pp. 577–580.

Borin, L. et al. (2014). Representing Swedish Lexical Resources in RDF with lemon. In: *International Semantic Web Conference (Posters & Demos)*. Citeseer, pp. 329– 332.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a module for lexicography in OntoLex. In: *DICTIONARY News* 7.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. et al. (2015). Applying the ontolex model to a multilingual terminological resource. In *European Semantic Web Conference*. Springer, pp. 283–294.

Buitelaar, P. et al. (2006). LingInfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the OntoLex Workshop at LREC*.

Chiarcos, C. & Sukhareva, M. (2015). Olia – ontologies of linguistic annotation. *Semantic Web* 6.4, pp. 379–386.

Chiarcos, C., Abromeit, F. et al. (2016). Etymology Meets Linked Data. A Case Study In Turkic. In *Digital Humanities 2016, DH 2016, Conference Abstracts*. Krakow, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 458– 460. ISBN: 978-83-942760-3-4.

Chiarcos, C., McCrae, J. et al. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.

Cimiano, P., McCrae, J. et al. (2013). "On the role of senses in the ontology lexicon". In *New trends of research in ontologies and Lexical resources*. Springer, pp. 43–62.

Cimiano, P., McCrae, J. & Buitelaar, P. (2014). *Lexicon Model for Ontologies: Community Report*. W3C Community Group Final Report.

Cimiano, P., Buitelaar, P. et al. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* 9.1, pp. 29–51.

Cimiano, P., Haase, P. et al. (2007). "LexOnto: A model for ontology lexicons for ontology-based NLP". In *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07*.

De Caluwe, J. & Taeldeman, J. (2003). 2.5 Morphology in dictionaries. In P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Vol. 6. John Benjamins Publishing, pp. 114–126.

Farrar, S. & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international* 7.3, pp. 97–100.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem et al. (eds.) *Proceedings of eLex 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.

Hartmann, R. R. K. (2001). *Teaching and researching lexicography*. Routledge.

Ide, N. & Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *4th International Conference on Language Resources and Evaluation-LREC'04*, pp. 135–138.

Kemps-Snijders, M. et al. (2009). ISOcat: Remodeling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)* 4.4, pp. 261–276.

Khan, F. (2018). Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In J. P. McCrae et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-19-1.

Khan, F. et al. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In: *LDK Workshops*, pp. 43–50.

Klimek, B. et al. (2016). Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Klimek, B. (2017). Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*.

McCrae, J., .Fellbaum, C. & Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

McCrae, J., Aguado-de-Cea, G. et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation* 46.6, pp. 701–709.

McGuinness, D. L., Van Harmelen, F. et al. (2004). *OWL: Web Ontology Language overview*. W3C recommendation.

Montiel-Ponsoda, E. et al. (2011). Enriching ontologies with multilingual information. *Natural language engineering* 17.3, pp. 283–309.

Real Academia Española and Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Santillana Ediciones Generales.

Schierholz, S. J. (2015). Methods in Lexicography and Dictionary Research. *Lexikos* 25, pp. 323–352.

Swanepoel, P. H. (2015). The design of morphological/linguistic data in L1 and L2 monolingual, explanatory dictionaries: a functional and/or linguistic approach? *Lexikos* 25, pp. 353–386.

2.5 Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory

This publication is the result of an effort to evaluate and prove the technical correctness and usability of the MMoOn Core ontology together with its cross-disciplinary applicability. An individual researcher compiling Hebrew language data became aware of the MMoOn Core ontology and its possibilities for representing morphological language data. As many language experts that lack a comprehensive representation tool for documenting their manually accumulated and often fine-grained language data, this researcher worked with a custom made table that grew in complexity and size over time. Consequently, the manual management and data documentation came to the turning point where adding more data without any consistency or duplication errors posed a serious issue threatening the data quality.

[P5] presents the Hebrew Morpheme Inventory as the result of the conversion of the source data table into Linked Data and explains how its prevalent issues could be solved. Therefore, this publication describes the transformation of this tabular dataset into MMoOn-RDF. This first MMoOn morpheme inventory serves as a verification of the applicability of the MMoOn Core ontology in general. The illustration of the suitability of MMoOn Core for more complex and non-concatenative morphological languages such as Hebrew implies that the ontology is equally applicable to other languages exhibiting a high degree of inflection and, thus, also to languages involving less complex morphological elements and processes such as agglutinative languages.

Creating Linked Data Morphological Language Resources with MMoOn
The Hebrew Morpheme Inventory

Bettina Klimek[1,2,a], Natanael Arndt[1,2,a], Sebastian Krause[2,b], Timotheus Arndt[3,c]
[1]Agile Knowledge Engineering and Semantic Web
[2]Faculty of Mathematics and Computer Science
[3]Research Centre Judaism, Faculty of Theology
Leipzig University, Germany
[a]{klimek, arndt}@informatik.uni-leipzig.de
[b]sebastian.krause@studserv.uni-leipzig.de
[c]tarndt@uni-leipzig.de

Abstract

The development of standard models for describing general lexical resources has led to the emergence of numerous lexical datasets of various languages in the Semantic Web. However, there are no models that describe the domain of morphology in a similar manner. As a result, there are hardly any language resources of morphemic data available in RDF to date. This paper presents the creation of the Hebrew Morpheme Inventory from a manually compiled tabular dataset comprising around 52.000 entries. It is an ongoing effort of representing the lexemes, word-forms and morphologigal patterns together with their underlying relations based on the newly created Multilingual Morpheme Ontology (MMoOn). It will be shown how segmented Hebrew language data can be granularly described in a Linked Data format, thus, serving as an exemplary case for creating morpheme inventories of any inflectional language with MMoOn. The resulting dataset is described a) according to the structure of the underlying data format, b) with respect to the Hebrew language characteristic of building word-forms directly from roots, c) by exemplifying how inflectional information is realized and d) with regard to its enrichment with external links to sense resources.

Keywords: morpheme ontology, Hebrew, language data, morphology, linguistic linked open data, MMoOn

1. Introduction

Since the development of the Linguistic Linked Open Data Cloud[1] in 2010 more than one hundred datasets have been created. They represent linguistic data such as lexicographic and phonological resources, terminological data, but also corpora and etymological language resources (Chiarcos et al., 2012). However, they lack the morphological layer. In addition, a Linked Data model dedicated to the domain of morphology has not been created so far. Nonetheless, there are Linked Data vocabularies which describe morphological features to some extent, e.g. GOLD[2] (Farrar and Langendoen, 2010), *lemon*[3] (McCrae et al., 2011), and Lexinfo[4] (Cimiano et al., 2011), even though these have not yet been used to create morphological data[5]. In order to fill the gap of missing morphological resources published as Linked Data, we created the Multilingual Morpheme Ontology (MMoOn) which is designed to describe morphemic data of any language at the word and sub-word level. In this paper we introduce an exemplary dataset, the Hebrew Morpheme Inventory, which is built with the MMoOn Core model, that shall encourage the construction of further MMoOn morpheme inventories for different languages. The data is freely available under the MMoOn project website[6].

Throughout the paper we are using QNames[7] for better readability of RDF terms. The prefixes are defined as in Figure 4.

The paper proceeds with a short overview of related work in Section 2., followed by a description of the MMoOn Core model for describing morphemic data in Section 3. Section 4. describes the development of the Hebrew Morpheme Inventory with MMoOn, including an outline of the data basis in Section 4.1. Sections 4.2. and 4.3. illustrate the specific language data according to the applicability of the MMoOn ontology for fine-grained morphological data description. This involves the representation of root derivation and verb inflection. Additionally, Section 5. describes the enrichment of the data with external resources. An overview of the resulting dateset will be given in Section 6. The paper closes in Section 7. with concluding remarks and a prospect of the future work.

2. Related Work

The examination of the related works in the domain of morphological data revealed five types of language resources. The resources were investigated for 1) the data format in which they are provided, 2) the extent of morphological data they contain, e.g. morphemes, morphs, lemmas, and 3) reusability. The findings are described as follows:

1) Unstructured data: A great amount of (free of charge) morphemic data is available only in human-readable formats. These comprise mostly html websites such as Wiktionary[8] and Canoo[9] for German, but also interlinear glossed text examples which can be found in numerous published

[1]http://linguistic-lod.org/llod-cloud
[2]http://linguistics-ontology.org/gold/2010
[3]http://lemon-model.net
[4]http://www.lexinfo.net
[5]http://lov.okfn.org/dataset/lov/terms?q=Morpheme
[6]https://github.com/aksw/mmoon
[7]https://www.w3.org/TR/2009/REC-xml-names-20091208/
[8]https://www.wiktionary.org
[9]http://www.canoo.net

linguistic PDF documents. These resources are mostly produced manually by domain experts and contain high quality data including segmented inflectional and derivational morphemes even for under-resourced languages. However, this kind of morpheme data is not machine-processable and, therefore, hardly reusable and hence remains isolated on the Web.

2) Structured data: In recent years, efforts have been undertaken to convert unstructured language data into XML datasets (ODIN[10]) or to encode morphological data directly in XML. Examples are the Alexina[11] (Sagot, 2010) and TypeCraft[12] (Beermann and Mihaylov, 2014) projects. These datasets also contain fine-grained morphemic data but are also machine-processable and, thus, easier to reuse.

3) Segmentation tools: Next to the existing language resources providing and describing morphological data, morphological segmentation tools have been developed which derive morphemic segments from language-independent word list input, e.g. Morfessor[13] (Creutz and Lagus, 2005a; Creutz and Lagus, 2005b), and language specific text or word list inputs, e.g. Morphisto[14] (Zielinski and Simon, 2009) and TAGH[15].
All of the tools we examined used their proprietary output formats and representation for the morphemic output, which is not directly reusable or convertible to Linked Data without further ado. Also, due to the variety of morphological realizations, the resulting segmentations are error-prone and require further post-editing. What is more, these tools handle mainly languages with concatenative morphology. For the particular case of Modern Hebrew, the state of the art is the morphological analyzer available on MILA[16] (Itai and Wintner, 2008), based on a morphological grammar implemented previously using finite-state technology (Yona and Wintner, 2008). This analyzer provides fine-grained data about the morphological information, which is available also in XML format. This tool, however, provides information about roots and patterns only for verbs, but not for other word classes e.g., nouns and adjectives, for which word-formation also involves association of roots and patterns.

4) Linked Data vocabularies: Within the research area of the Semantic Web, ontological models covering linguistic data –in general– have been created. Ontologies such as GOLD[17] (Farrar and Langendoen, 2010), OLiA[18] (Chiarcos, 2008), Lexinfo[19] (Cimiano et al., 2011) and *lemon*[20] (McCrae et al., 2011) provide very broad, multi-domain vocabularies that only partially cover concepts and relations of the morphological domain. The advantage of these vocabularies lies in the highly interoperable data format which allows for direct reuse and extension. None of these vocabularies was designed to exhaustively describe the domain of morphology in the first place, thus leaving a gap, which motivated the creation of MMoOn.

5) Linked Data datasets: So far Dbnary[21] (Sérasset, 2012) extracts Wiktionary inflection tables for German, French and Serbo-Croatian in RDF[22]. These Dbnary "morpho" datasets are based on the *lemon* and OliA vocabularies and are hence interoperable in a non-specific manner. Nonetheless, the data provided does not contain morphs but only a set of grammatical meanings attached to unsegmented word-forms. Similar or even more fine-grained morphological datasets in RDF are not available yet.

This overview of morphological resources reveals a gap between the existing non-Linked Data resources and the available Linked Data models. As a result, the current landscape of morphology consists of isolated but extensive non-RDF resources on the one side, and interoperable Linked Data vocabularies which are insufficiently expressive to model morphology, on the other side. In particular, the fact that concrete segmented morpheme data could not be identified in RDF resources reduces the applicability of the mentioned models, dictionary resources, and tools on language-specific textual datasets or corpora. Consequently, this general lack of Linked Data language resources in the domain of morphology reveals the demand for morphological data that applies both to the language-specific morphological domain needs and to cross-lingual interoperable data modelling standards.

3. The Multilingual Morpheme Ontology

In order to bridge the gaps that currently separate the various existing morphological data resources and models described above, we developed the MMoOn Core model[23]. In particular, it focuses on the description of the necessary concepts and their relations involved in the domain of morphology. The ontology is freely available for reuse, and can be downloaded from: https://github.com/AKSW/MMoOn/blob/master/core/mmoon.ttl[24].
MMoOn enables the documentation of the morphological data of any inflectional language in RDF. Figure 1 shows how language-specific morpheme inventories are designed. The ontological foundation of each morpheme inventory builds the MMoOn Core model which covers eight main classes (dark blue and orange) and serves as the **language-independent schema level**. The largest classes are `mmoon:Meaning` and `mmoon:MorphemicGloss` providing

[10] http://odin.linguistlist.org
[11] http://alexina.gforge.inria.fr
[12] http://typecraft.org
[13] http://www.cis.hut.fi/projects/morpho
[14] http://www1.ids-mannheim.de/lexik/home/lexikprojekte/lexiktextgrid/morphisto.html
[15] http://www.tagh.de/index.php
[16] http://yeda.cs.technion.ac.il/
[17] http://linguistics-ontology.org/gold/2010
[18] http://acoli.cs.uni-frankfurt.de/resources/olia
[19] http://www.lexinfo.net
[20] http://lemon-model.net
[21] http://kaiko.getalp.org/about-dbnary
[22] http://kaiko.getalp.org/about-dbnary/download
[23] see also http://mmoon.org/publications.
[24] An overview of the MMoOn Core vocabulary is displayed here: http://mmoon.org/mmoon-core-model.

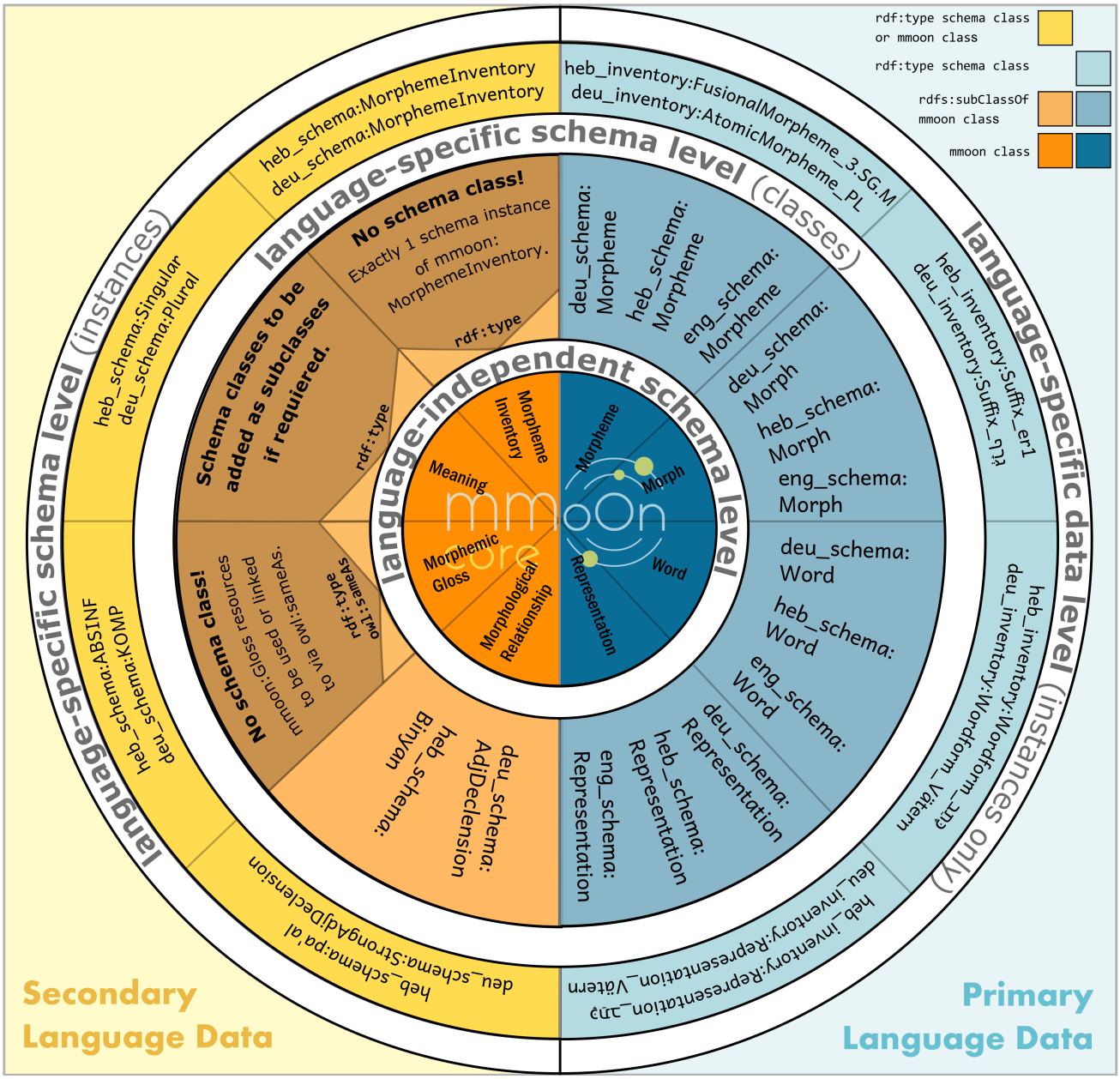


Figure 1: Architectural setup of MMoOn morpheme inventories.

a substantial set of grammatical features and categories together with their respective gloss representations. Specific properties explicate granular relations between the word and sub-word, i.e. the `mmoon:Morph` elements. Also the `mmoon:Word`, `mmoon:Morph` and `mmoon:Morpheme` classes are further devided into subclasses which enable more fine-grained specifications of the language data, e.g. lexemes, word-forms, prefixes, suffixes, transfixes, atomic and fused morphemes. Additionally, various relations are established via object properties between the MMoOn Core classes which are not depicted in Figure 1. Properties such as `mmoon:consistsOfMorph`, `mmoon:belongsToWord`, `mmoon:correspondsToMorpheme`, or `mmoon:isAllomorphTo` ensure that precise statements about the morphological segmentation of words can be asserted.

In order to build the morpheme inventory of a certain language, the MMoOn Core model needs to be extended with a **language-specific schema level** for all required classes describing primary language data[25]. On this level, language-specific subclasses and subproperties are defined according to the descriptive needs of the language, and all language-relevant secondary language data must be directly added as instances to corresponding MMoOn Core classes (see middle circle in Figure 1) whenever possible or to the newly created language-specific schema classes. As such, the schema data evolving on this level constitutes the language-specific terminological foundation for describing the primary lan-

[25] The distinction between primary and secondary linguistic data is based on Lehmann (2004). Even though, it has to be stressed, that the notion of *primary data* in the context of the MMoOn model basically refers to the meta-instances under which all primary language data instances can be subsumed.

guage data. A schema ontology extension of the MMoOn Core has been set up for the Hebrew Morpheme Inventory. Together with the schema ontologies of future inventories to come, this layer in the MMoOn architecture will enable a multilingual comparative access to the language data due to their shared conceptual basis of the MMoOn Core ontology. Finally, the MMoOn morpheme inventory is created as instance data on the **language-specific data level** by using the language-specific schema vocabulary and the MMoOn Core properties.

To sum up, the MMoOn Core model enables the creation of language-specific, extensive and fine-grained morphological datasets in RDF. What is more, by sharing the conceptual core of the MMoOn ontology, all MMoOn morpheme inventories to come will add to the formation of a multilingual dataset, which can be used not only as a data basis for specific NLP tasks but also as an empirical foundation for comparative linguistic research.

4. The Hebrew Morpheme Inventory

In accordance with the procedure outlined above, we created the Hebrew Morpheme Inventory. It is a dataset which consists of two ontologies resp. models and one file containing only primary language instance data: 1) the MMoOn Core ontology (http://mmoon.org/mmoon/, `mmoon.ttl`), 2) the Hebrew schema ontology (http://mmoon.org/lang/heb/schema/oh/, `heb_schema.ttl`) and 3) the Hebrew morpheme inventory[26] (http://mmoon.org/lang/heb/inventory/oh/, `heb_inventory.ttl`). This dataset is an ongoing effort of compiling lexical and morphological Hebrew language data in RDF and shall serve as the knowledge base for an Open Hebrew online dictionary in the future. This initial release and all future versions will be provided at http://mmoon.org/.

4.1. Data Basis

The basis for the inventory data is a handcrafted vocabulary table containing vocalized and unvocalized Hebrew content words, suffixes and non-inflecting words annotated with their roots, word-class information and English, German and Russian translations. This data has been compiled by a Hebrew speaker and, therefore, assures a significant quality of the data. The data has been analyzed, integrated and transformed to the MMoOn Core and the specific Hebrew schema using a custom data integration pipeline. Therefore, the data has been cleaned according to formal criteria. Lexical data entries containing invalid syntax have been removed, e. g. invalid braces, multiple entries in one column, or entries with missing word-class information. This step has been undertaken to achieve a sufficient data quality. After this mostly syntactic cleaning process, from the initial 52.000 lexical entries 11.600 remained for which morphological information is of relevance. These have been mapped onto the established schema ontology and then further processed and transformed to RDF.

[26]The most recent versions of file 2) and 3) are available here: https://github.com/AKSW/MMoOn/tree/master/lang/heb.

4.2. Hebrew Root Derivation

Hebrew is characterized by a highly fusional morphology. However, in contrast to the Indo-European languages Hebrew exhibits a prominent discontinuous morphological relationship called introflection as Semitic languages typically do. That means that morphs do not appear as linearly segmentable units in terms of concatenative stems and affixes. Rather, words in Hebrew consist of a consonantal root tier, which is inserted into a specific pattern tier, consisting of vowels and possibley also consonants (McCarthy, 1981), as depicted in Figure 2. A root is primarily composed of three consonants, called *radicals*, and it carries the core semantic of every lexical expression derived from it. The pattern carries the morpho-syntactic features of the word-form.

Figure 2 shows the root כתב (*k.t.b*), having a general meaning around the concept 'write' and is given as illustrative case. Often, a more complex meaning can be directly derived from roots by adding affixes. Here, the secondary root שכתב (*š.k.t.b*) is formed from the primary root כתב and the prefix ש, resulting in a combinatory meaning of both elements yielding the concept 'rewrite'. At the root level no grammatical meanings are involved and hence roots do not have any word-class affiliation. A word-form is then created by applying a specific vowel pattern to the root. In morphological terms, these patterns can be classified as transfixes, given that they have some (grammatical) meaning on the one side but a discontinuous representation which leaves slots (cf. the dotted circles shown in the `heb_schema:Transfix` instances) for the consonantal letters on the other side. Hence, roots and transfixes in Hebrew have very abstract representations, which make them unpronouncable in isolation. Only when both are combined a word-form[27] evolves. Figure 2 displays eight word-forms, four of which are built with the simple (primary) root and four with the complex (secondary) root. This kind of word-formation through root derivation is very productive in Hebrew and many more word-forms can be constructed from one root. The meaning of word-forms can be predicted through the combination of the root sense and the grammatical function of the transfix. E.g. word-form five is a noun with the underlying concept of 'write' plus an agent nominalization, resulting in the lexical meaning 'reporter, correspondent, journalist' – "a person whose profession is writing (news)". Similarly, the meanings of the other seven word-forms can be deduced[28].

4.3. Verb Inflection

Due to the high productivity of the transfixal patterns in Hebrew, linguists and dictionary writers created a high amount of inflectional tables linked to specific groups defining the underlying morphological building patterns for hundreds of Hebrew roots (Even-Shoshan, 2003; Barkali, 1962). The knowledge contained in these works is very valuable, but

[27]Note that the `heb_schema:Transfix` resources contain also inflectional meanings, i.e. gender, number, person, tense, which are not displayed in Figure 2.

[28]These meanings are also included in the data, however, not shown in Figure 2.

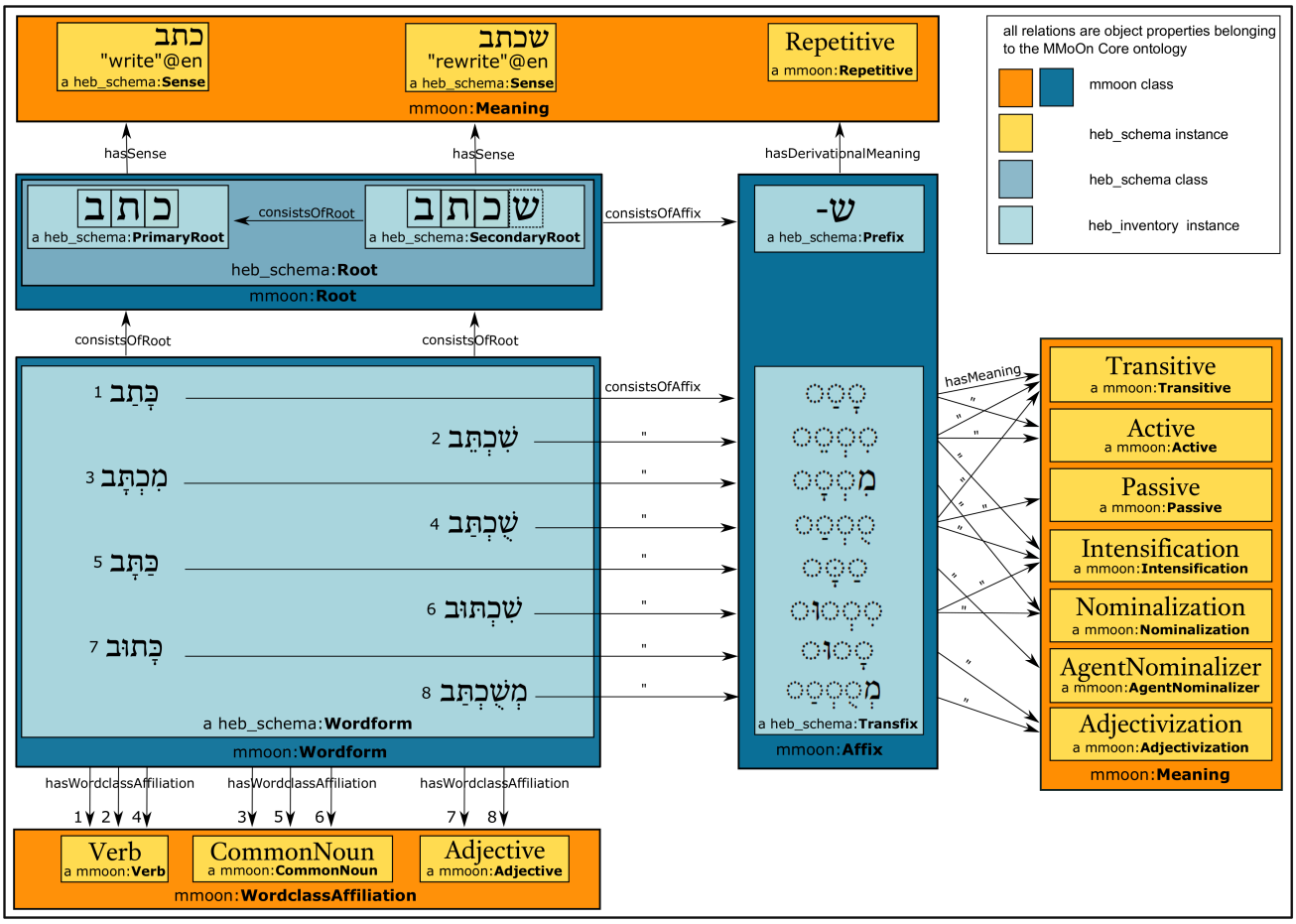


Figure 2: An example for Hebrew root derivation modelled with MMoOn.

non-existent in a digital format. Therefore, retrieving information about words that share the same root, word-forms that belong to one lexeme or which morphological patterns are used in word-formation is bound to tedious and time-consuming manual search through books.

The Hebrew Morpheme Inventory presents the first step towards an interlinked and machine-readable representation of roots, lexemes, word-forms and morphs. Similarly to the root derivation process, the inflection of lexemes relies on combining a consonantal root with a transfix pattern. Figure 3 shows four word-forms for each of the two lexemes לִמֵּד *limmed* 'teach.3SG.M.PST' and בִּשֵּׁל *biššel* 'cook.3SG.M.PST' as they are represented within the data graph of the Hebrew Morpheme Inventory. Crucial to the formation of the word-forms is the morphological relationship, i.e. the assigned Binyan, that holds for the lexeme, and which depends on its root. Traditionally, Barkali (1962) has set up verb conjugation tables that are classified according to Binyan groups which apply to certain roots, and which list all associated word-forms with an exemplary root. Due to the fine-grained vocabulary of the dataset, all lexical and morphological relevant information can be explicated in the specific resources. Both lexemes in Figure 3 consist of roots from which the word-forms of the Barkali Pi'el group 1 have been built. Consequently, all of these word-forms are related to the same set of transfixes, since they are in the same Binyan group. Given that Hebrew verbs inflect for the categories of person, number, tense and gender Barkali lists altogether 32 word-forms (of which four are shown in Figure 3 only), including five infinitives and four imperatives. As can be seen, each word-form is related to the lexeme it belongs to and to the morphs, i.e. the root and the pattern resources, of which it consists. The meanings of the transfixes are further specified by relating them to their corresponding morphemes. This is illustrated only for one transfix in Figure 3. That way the structural components of the word-form as well as the fusional meaning they convey are stated. This separation of the various kinds of resources involved in the Hebrew verb conjugation enables precise extraction of morphological information from the dataset. For instance, it is possible to find all roots which can build word-forms according to a specific Binyan. Also all distinct word-forms that consist of a specific transfix can be retrieved, e.g. all verb-forms that are inflected for first person, singular, feminine, past tense. By searching for all realizations that a specific morpheme is linked to, even allomorphs can be obtained.

Conforming to the example given in Figure 3, the verb-forms have been generated via a script that takes the roots as parameters and returns the list of word-forms according to the transfix patterns of the Barkali Pi'el 1 group. In a similar fashion, the script associates other roots with different Binyan groups to create word-forms. Similarly to the Binyan determining the verb-form patterns, the so called

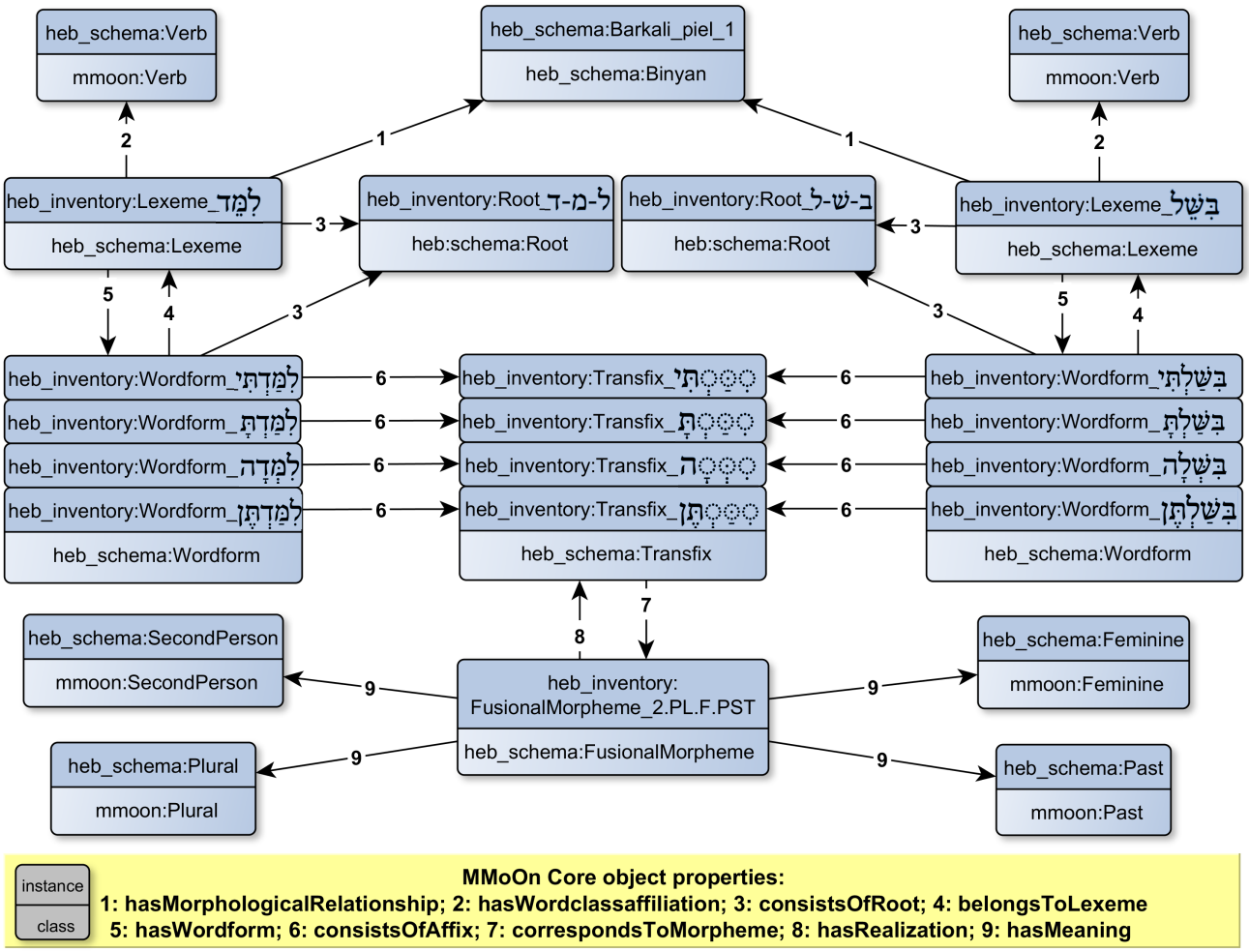


Figure 3: Morphological data of verbs in the Hebrew Morpheme Inventory.

Mishkal determines the word-form patterns for certain noun classes. At this point word-forms for verbs have been generated for the first seven Barkali Binyan groups and subgroups (Pa'al 1-5, Nif'al 1-3, Pi'el 1, Pu'al 1, Hitpa'el 1, Hif'il 1-3 and Huf'al 1-3) and for four Mishkal groups (Barkali nb. 91, 118, 144 and 274). These cover the most frequent inflectional patterns in Hebrew.

5. Interlinking the Hebrew Morpheme Inventory

In order to comply to the five star Linked Data principles (Berners-Lee, 2009) the Hebrew Morpheme Inventory needs to be interlinked with other resources on the Semantic Web. As already mentioned before, morphological Linked Data resources for the Hebrew language are not available to date. Lexical data, however, is present in BabelNet[29], which is the largest multilingual Linked Data dataset and semantic network (Navigli and Ponzetto, 2012). It contains around half a million lexical entries for Hebrew together with their canonical forms, part of speech information and senses. Since BabelNet is very well maintained and of high quality we decided to enrich the `heb_schema:Lexeme` resources of the Hebrew Morpheme Inventory with external sense links from BabelNet, for example `http://babelnet.org/rdf/אֲוִירוֹן_HE/s00001697n`. The sense links in BabelNet in turn also refer to Wordnet senses such as http://wordnet-rdf.princeton.edu/wn31/201203727-v wich are very granular and accurate. Firstly, the `heb_schema:Lexeme` instances have been looked up for their equivalent existance as BabelNet Hebrew lexical entries. For every obtained match the `heb_schema:Lexeme` instances have been linked to the lexical BabelNet instances via the `rdfs:seeAlso` property. Secondly, the corresponding BabelNet sense instances have then been linked by using the `lemon:sense` property. The integration of these links is exemplified in Figure 4. Currently the dataset contains 1848 links to BabelNet lexical entries and 2520 links to BabelNet senses. This interlinking is seen as a valuable enrichment for the Hebrew Morpheme Inventory.

6. The Dataset

From the given examples in Sections 4.2. and 4.3. it becomes clear that describing morphological data requires a highly specialized and fine-grained data model that can cap-

[29] http://babelnet.org

```
@prefix mmoon: <http://mmoon.org/mmoon/> .
@prefix heb_schema: <http://mmoon.org/lang/heb/schema/oh/> .
@prefix heb_inventory: <http://mmoon.org/lang/heb/inventory/oh/> .
@prefix lemon: <http://www.lemon-model.net/lemon#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

heb_inventory:Lexeme_אֲוִירוֹן a heb_schema:Lexeme ;
                          rdfs:label "אֲוִירוֹן"@he ;
                          mmoon:hasWordclassAffiliation heb_schema:CommonNoun ;
                          mmoon:hasInflectionalCategory heb_schema:Masculine ;
                          rdfs:seeAlso <http://babelnet.org/rdf/אֲוִירוֹן_n_HE> ;
                          mmoon:hasRepresentation heb_inventory:Representation_אֲוִירוֹן ,
                                                  heb_inventory:Representation_אווירון ;
                          mmoon:hasSense heb_inventory:Sense_de_Flugzeug ,
                                  heb_inventory:Sense_en_aircraft , heb_inventory:Sense_en_airplane ,
                                  heb_inventory:Sense_ru_аэроплан , heb_inventory:Sense_ru_самолёт ;
                          lemon:sense <http://babelnet.org/rdf/אֲוִירוֹן_HE/s00001697n> ,
                                      <http://babelnet.org/rdf/אֲוִירוֹן_HE/s16750414n> .
```

Figure 4: Interlinking of `heb_schema:Lexeme` resources with BabelNet lexical entries and senses.

ture all the morphological elements together with their various meanings and relations. For the Hebrew Morpheme Inventory this is achieved by using and extending the MMoOn Core ontology as shown in the Figures 2 and 3. Therewith, the dataset constitutes a language resource which applies both to the granularity of the morphology domain needs and to recent data modelling standards. Being created in RDF enables the explicit reference to morphemic elements together with their various interrelations to other linguistic units.

Overall the Hebrew Morpheme Inventory currently consists of the following resources that have been converted to RDF from the original cleaned tabular data basis:

- 2923 words which have another word-class than verb, noun or adjective,
- 8714 lexemes which are either verbs, nouns or adjectives,
- 21030 representations of the vocalized and unvocalized lexeme and word resources,
- 17892 senses which are the English, German and Russian translations of the table,
- 1795 roots (1769 primary and 36 secondary),
- 98824 word-forms which have been additionally generated for 1568 lexemes from ca. 400 roots,
- 619 tranfixes,
- 13 suffixes, and
- 2520 links to external BabelNet senses.

At the moment this is only one fifth of the original data basis. Since this data shall serve as the foundation for an open online dictionary, however, the dataset will be constantly growing and maintained.

7. Conclusion and Future Work

In this paper, we introduced the MMoOn Core ontology for describing morphemic data with different levels of granularity. Such an effort is –to the best of our knowledge– unique and fills the gap among existing coarse-grained RDF vocabularies as described in the Related Work section. We presented the development of the Hebrew Morpheme Inventory as a showcase for the creation of language-specific morphemic data with MMoOn. We showed that MMoOn is suitable for describing complex morphemic elements and their relations even for languages, such as Hebrew, which deviate from traditional Indo-European word-form analysis. Consequently, the Hebrew Morpheme Inventory represents a novelty among the current language resource landscape by expressing fine-grained morphemic language data in conformity with Linked Data modelling standards.

Future work includes: (1) the transformation of the remaining tabular data basis to RDF, (2) the constant enrichment of the Hebrew morpheme inventory with further language data, (3) the interlinking of this dataset to further resources in the Linguistic Linked Open Data cloud (4) the publication of the Hebrew Morpheme Inventory on the Web together with a SPARQL endpoint. Also, a paper presenting the MMoOn Core ontology is currently written and will be submitted to the Semantic Web Journal soon. This paper can then be found at: http://mmoon.org/publications.

8. Acknowledgements

This paper's research activities were partly supported and funded by grants from the FREME FP7 European project (ref.GA-644771), the European Union's Horizon 2020 research and innovation programme for the SlideWiki Project under grant agreement No 688095 and the German Federal Ministry of Education and Research (BMBF) for the LEDS Project under grant agreement No 03WKCG11C.

The authors want to thank Amit Kirschenbaum at Leipzig University for supporting this work with his expertise on the Hebrew language and his insightful advice.

9. Bibliographical References

Barkali, S. (1962). *Luax HaP'alim HaShalem (the complete verbs table)*. Reuven Mass, Jerusalem. In Hebrew.

Beermann, D. and Mihaylov, P. (2014). Typecraft collaborative databasing and resource sharing for linguists. *Language Resources and Evaluation*, 48(2).

Berners-Lee, T. (2009). Linked Data. Design issues, W3C, June. http://www.w3.org/DesignIssues/LinkedData.html.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, pages 1–136.

Cimiano, P., McCrae, J., Buitelaar, P., and Stintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 29–51.

Creutz, M. and Lagus, K. (2005a). Inducing the morphological lexicon of a natural language from unannotated text. *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, 1:106–113.

Creutz, M. and Lagus, K. (2005b). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. *Helsinki University of Technology*.

Even-Shoshan, A. e. a. (2003). *Even Shoshan Dictionary*. Jerusalem: Qiryat-Sefer Publishing.(Hebrew).

Farrar, S. and Langendoen, D. T. (2010). An owl-dl implementation of gold. *Linguistic Modeling of Information and Markup Languages*, pages 45–66.

Itai, A. and Wintner, S. (2008). Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Lehmann, C. (2004). Data in linguistics. *The Linguistic Review*, 21:175–210.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, pages 373–418.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *The semantic web: research and applications*, pages 245–259.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. *7th international conference on Language Resources and Evaluation (LREC)*.

Sérasset, G. (2012). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web Journal-Special issue on Multilingual Linked Open Data*.

Yona, S. and Wintner, S. (2008). A finite-state morphological grammar of hebrew. *Natural Language Engineering*, 14:173–190, 4.

Zielinski, A. and Simon, C. (2009). Morphisto – an open source morphological analyzer for German. *Proceedings of the 2009 conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008*.

2.6 Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

Finally, the last publication presents another dataset description of a MMoOn morpheme inventory. It focuses on the under-resourced languages of the Bantu language family, in particular on the language Xhosa, which are often accompanied with data scarcity. Until a consistent dictionary and grammar for such a language evolves many restructuring and transformation steps of initial raw data take place. Efforts to arrive at such resources for creating multilingual language material for language revitalisation and education purposes are pursued for the Bantu languages by the South African Centre for Digital Language Resources (SADiLaR). The development of a suitable data model that enables the preparation and usage of this data within a multilingual and federated environment posed the main focus of this application of the MMoOn Core ontology. In this respect, the desired technical foundation of all available language resources of SADiLaR should a) account for the representability of the language specific peculiarities of the Bantu languages and b) allow for a homogeneous interconnection of all, hitherto, unrelated datasets which differ with regard to their underlying data format but also content-wise in terms of their coverage and granularity.

As a result of addressing these data representation needs, the Xhosa RDF dataset has been created. It is the first dataset that transformed a SADiLaR language resource adhering to these two requirements. The source data was manually compiled on index cards before being digitised into a table and contained mostly morphological information and translations of Xhosa. [P6] describes how the tabular data could be used to derive a MMoOn-based model, i.e. the Bantu Language Model (BLM) that enabled the representation of the Xhosa source data. Moreover, the BLM is explained in detail and illustrates the flexibility that the MMoOn Core model allows in representing not only the morphological data of single but also multiple languages of the same language family.

Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, Uwe Quasthoff
Department of African Languages, University of South Africa;
Natural Language Processing Group, Agile Knowledge Engineering and Semantic Web Group, Leipzig University
boschse@unisa.ac.za, {teckart, klimek, dgoldhahn, quasthoff}@informatik.uni-leipzig.de

Abstract

The South African linguistic landscape is characterised by multilingualism and the influence between their eleven official and some local languages. Unfortunately, for most of the languages the amount and quality of available lexicographical data is suboptimal, even though its availability is essential for all educational institutions and for the development of state-of-the-art language technology. In this paper we present a new source of lexicographical data for Xhosa, a language spoken by more than eight million speakers. For its utilisation in a multilingual and federated environment it is modelled using a dedicated OWL ontology for Bantu languages and possesses all features that are currently considered integral for the promotion of resource reuse as well as long-term usage. In the future, the introduced ontology may be used for other Bantu languages as well and may ease their combination to achieve more extensive, multilingual data stocks.

Keywords: Xhosa, lexicography, research infrastructures, linked data

1. Introduction

A basic requirement for the language processing capability for any language is the availability of lexicographical data, ideally open source data which is often hard to find for less resourced languages. This includes many members of the Bantu language family. For enhancing the usability of this kind of data this paper presents a Bantu Language Model for describing lexicographical data in RDF. Furthermore, it presents a new resource of lexicographical data for the Xhosa language based on this new model. The data to be presented in this model is a representative sample of raw data for a Xhosa-English dictionary, containing approximately 6,800 lexical entries. In its final state, the data set should contain approximately 10,000 lexical entries. Whereas the available data enables us to use Xhosa as the language of instantiation, the method and model are extensible and applicable to many other Bantu languages, in particular those belonging to the same group.

Together with this paper, both the Bantu Language Model and the current state of the lexicographical data set are freely available for download and for querying via a dedicated SPARQL endpoint. It should be noted that the research reported on in this paper is work in progress. Its final version will be provided via SADiLaR, the South African Centre for Digital Language Resources. Moreover, the current version of the data is already available via CLARIN-D (see section 4.).

The remainder of this paper is structured as follows: Section 2 describes the origin of the Xhosa language material and explains essential features of the Xhosa language with a focus on its morphology. Section 3 explains the new Bantu Language Model that is based on the established MMoOn ontology[1]. Section 4 gives detailed information about the structure of the Xhosa RDF data set using a concrete example. Furthermore, information about its current extent are provided. Section 5 demonstrates the relationship of the described work in the context of federated research infrastructures and how they can simplify access to and enhance usability of modern lexicographical data. The paper closes with a short summary and an outlook to planned further work.

2. The Xhosa Source Data

The data used for this case is based on Xhosa [xho][2], one of the official languages of South Africa belonging to the so-called Bantu language family. It is spoken predominantly in the Eastern Cape and Western Cape regions. There are approximately 8.1 million Xhosa speakers[3], adding up to about 16% of the South African population. Xhosa, as member of the Nguni language group, shares many linguistic features with other Nguni languages, which include Zulu [zul], Swati [ssw], Southern Ndebele [nbl] and Northern Ndebele [nbe][4]. Xhosa, like the other Bantu languages, is structurally agglutinating and is therefore characterised by words usually consisting of more than one morpheme. Each morpheme corresponds to a single lexical meaning or grammatical function. This particular Xhosa lexicographical data set is accompanied by English translations and was compiled and made available for purposes of further developing Xhosa language resources[5]. The process involved digitisation into CSV tables and various iterations of quality control in order to make the

[1] http://mmoon.org/core/

[2] Each language is followed by its ISO 639-3 code http://www.loc.gov/standards/iso639-2/php/code_list.php in order to distinguish one language from other languages with the same or similar names and to identify the names of cross-border languages.

[3] http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf

[4] The names of the Nguni languages in the languages themselves are respectively: isiXhosa, isiZulu, Siswati and isiNdebele.

[5] Bilingual (Xhosa-English) word lists were compiled by JA Louw after his retirement with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of among others botanical, animal and bird names, grammar terms, modern forms etc., as well as lexicalisations of verbs with extensions.

data reusable and shareable. In this paper, we concentrate on nouns and verbs. The excerpt of the lexicographical data set is a representative sample of Xhosa nouns and verbs. Nouns of all possible regular and irregular combinations of noun classes, and verbs with a variety of verbal extensions (leading to lexicalisations in meaning) are represented. Nouns are listed alphabetically according to noun stems, followed by the POS, the surface form of the singular and plural class prefixes (if applicable) as well as the number(s) of the class prefixes, and finally the English translations, e.g.

Noun stem	*POS*	*Class pref sg*	*Class no.*
phathi	noun	um	1
Class pref pl	*Class no.*	*English translation*	
aba	2	superintendent	

Verbs are listed alphabetically according to verb stem, i.e. the basic verb root followed by the inflection suffix -a, or sometimes -i, e.g.

Verb stem	*POS*	*English translation*
mi	verb	be standing
tyalisa	verb	help to plant

The lexicographic data is by no means based on corpus frequencies of nouns and verb stems as for instance the Oxford School Dictionary (De Schryver, 2014) but rather on complementation of existing, established dictionaries.

2.1. Xhosa Morphology

The noun is made up of two main parts, namely a noun prefix and a noun stem. All nouns are assigned to a particular class, as reflected in the class prefix. For practical and comparative purposes, noun classes have been given numbers by scholars working in the field of Bantu linguistics. Although 23 such classes have been reconstructed in Proto-Bantu, most Bantu languages have fewer than 20 classes (Nurse and Philippson, 2003). In Xhosa for instance, the class numbers end with class 17, while classes 12 and 13 do not occur at all (Pahl, 1967). It should be added that class 15 represents the infinitive class, while class 16 is no longer used productively to form nouns in Xhosa, but rather has an adverbial significance. Each noun class is characterised by a distinct prefix, which also includes a pre-prefix, and a particular singular/plural pairing with uneven numbers signifying singular and even numbers signifying plural. These class prefixes may show agreement with other constituents in a sentence. The only class pair with specific semantic contents is class 1/2 which contains personal nouns only. This does not, however, mean that all personal nouns occur in this class pair. For the rest of the noun classes, semantic arbitrariness is observed, although certain semantic generalisations do occur, e.g. classes 9 and 10 are generally referred to as the "animal classes" since they contain many animal names, but also many other miscellaneous terms. Noun stems may also be suffixed with morphemes such those indicating diminutive, augmentative, derogatory or feminine modifications to the basic meaning of the noun. A verb consists of a series of prefixes and suffixes that are built around a basic verb root carrying the basic meaning. A final inflection suffix completes the verb stem, to which pre-stem inflection is added in the form of, for example, the following morphemes: subject agreement, object agreement, negation, tense and aspect. Verbal suffixes may include morphemes such as: negation and derivational extension. The verb therefore carries much information and is pivotal in the sentence.

2.2. Discussion of the Xhosa Data

In the Xhosa data set under discussion, noun stems are separated from their class prefixes as is the case in traditional Bantu language dictionaries. In each instance the class prefix modifies the meaning of the basic noun stem e.g.

balo (isi 7; izi 8) "arithmetic"
balo (u 11) "census"

Although noun stems can be sub-divided into a root plus suffixes, any suffixes that occur, e.g. the feminine suffix *–kazi* and the diminutive suffix *–ana*, are not identified separately in our data. This is illustrated in the following examples where the modification of the basic meaning only appears in the English translations:

caka (isi 7; izi 8) "servant"
cakakazi (isi 7; izi 8) "servant girl"
cakazana (isi 7; izi 8) "young servant girl"

Noun class pairs normally signify singular/plural that correspond to the odd and even class numbers respectively, e.g.

khwenyana (um 1; aba 2) "son-in-law"
kroti (i 5; ama 6) "hero"

There are exceptions, however, for instance the singular class 11 takes its plural in class 10 (instead of 12, which does not exist in Xhosa), e.g.

diza (u 11; iin 10) "straw"

Also, the distinction between singular and plural does not apply to nouns that denote, for example, mass or abstract concepts, as in the case of:

bisi (u 11) "milk"
ophu (um 3) "vapour"

The following examples demonstrate that phonetically and phonologically conditioned allomorphs of class prefixes 1/2 (um-/aba- versus um-/ab-); 7/8 (isi-/izi- versus is-/iz-) and 11 (u- versus ulu-) appear in the data, e.g. in the case of vowel initial noun stems or monosyllabic noun stems (Kosch, 2006):

biki (um 1; aba 2) "reporter" vs. ongi (um 1; ab 2) "nurse"
kolo (isi 7; izi 8) "school" vs. enzo (is 7; iz 8) "deed, act"
patho (u 11) "school" vs. bi (ulu 11)

"misfortune, calamity"

These examples, therefore, demonstrate that for some noun classes more than one prefix member exists, resulting in allomorphs that occur in complementary distribution.
Verb stems are listed according to their infinitive form minus the infinitive prefix, i.e. the basic verb root followed by the inflection suffix *–a*. In some few cases, the final suffix presents as *–i* or *-e*. The latter only occurs in the case of stative verbs such as *–krekrelele* "stand in line". In the data there is no morphological differentiation between basic verb stems and verb stems with suffixed extension morphemes. The modification of the basic meaning of the verb stem, however, appears in the English translation, as in:

tenda "entertain"
tendana "entertain one another"
tendeka "be able to be entertained"
tendela "entertain at or for"
tendisa "help to entertain"

3. The Bantu Language Model

For the representation of the tabular Xhosa dictionary data and their translations we chose to convert the data into the RDF (Resource Description Framework) format. The mapping of the source data to RDF, however, requires a specific vocabulary which can be some existing or newly created ontology. While the lexicon model for ontologies (Lemon) (McCrae et al., 2011) was designed to represent lexical language data, its usage has been proven to be problematic for Bantu languages (Chavula and Keet, 2014). This is mainly due to the lack of the conceptualisation for morphological language data. Even though the Lemon model evolved to become a W3C recommendation published as the OntoLex-Lemon model that is split into five specified modules[6] (McCrae et al., 2017), the necessary modelling of morphological data has not been worked into this refined model.
Therefore, we created the Bantu Language Model[7] (in short BantuLM) as illustrated in Figure 1. This ontology is fully based on the reuse of and alignment to already existing vocabularies[8]. The largest part is based on the Multilingual Morpheme Core Ontology (MMoOn Core)[9] because it provides fine-grained classes and properties for representing morphological data and, moreover, already shares a considerable amount of overlap to the ontolex module for lexical data (Klimek, 2017). By taking the Xhosa verb and noun source data as an orientation point we identified three major linguistic subdomains that were to be modelled: 1) lexicographic data which is based on the OntoLex lime module[10] and MMoOn Core, 2) morphological data which is solely based on MMoOn Core, and 3) translational data which is based on the OntoLex vartrans module[11]. Despite the best practice recommendation to make direct reuse of existing vocabularies if they appropriately fit the modelling domain in question, a different approach of vocabulary reuse has been taken. In order to represent the BantuLM under a single namespace, all classes and properties have been newly created, however, corresponding to the reused external vocabularies. I.e. all classes that are based on MMoOn Core have identical labels and are aligned by usage of the `rdfs:subClassOf` object property in accordance to the creation procedure for MMoOn Core-based data sets. Otherwise, all classes that are based on the ontolex and lime vocabulary are interconnected with their derived counterparts via the `owl:equivalentClass` object property. The equivalence of all object properties within the BantuLM vocabulary is created with the `owl:equivalentProperty` property. Consequently, all definitions of the classes and properties need to be obtained from the interconnected original vocabularies. This poses, however, no disadvantage since the naming of the classes and properties is quite self-explanatory. While this kind of duplication of vocabularies is rather unusual it is formally valid in terms of ontology creation. This modelling of the BantuLM vocabulary has been chosen in preference of user-friendliness given that the data creators are mainly linguists that have only little or no expertise in creating language resources in the RDF format or within the Linked Data framework. It is assumed that a vocabulary that is applicable to all Bantu languages is easier to use and query for non-experts if it is built on a single namespace instead of a variety of vocabularies that need to be studied before they can be actually used for language data representation.
To conclude, the BantuLM is an aggregation of those classes and properties from the mentioned vocabularies that are necessary or useful to represent not only the Xhosa source data but also other Bantu languages in general, e.g. we had no data for the class `blm:Wordform`, but other Bantu language resources might well have and can then use this class accordingly. In contrast to the reused models the BantuLM is a language-specific model and, hence, specified for its affiliation to the Bantu language family. That means in particular, that grammatical meanings such as wordclass, number or nominal classifier are newly created and consequently specific to and shared by all Bantu language resources that will be based on the BantuLM ontology.
For the creation of the Xhosa RDF inventory data set the BantuLM proved not only to be fully suitable but also contributed to an explicit semantic interrelation between the lexical and morphological elements which is rather implicit in the tabular source data[12].

[6] `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[7] The ontology is available under the URI: `http://mmoon.org/bnt/schema/bantulm/`.

[8] Please consult the ontology URI for more information on how to use the ontology for creating other Bantu language data.

[9] Cf. `http://mmoon.org/` and `http://mmoon.org/core/` for more information.

[10] `http://www.w3.org/ns/lemon/lime#`

Model for lexical and morphological data of Bantu languages.

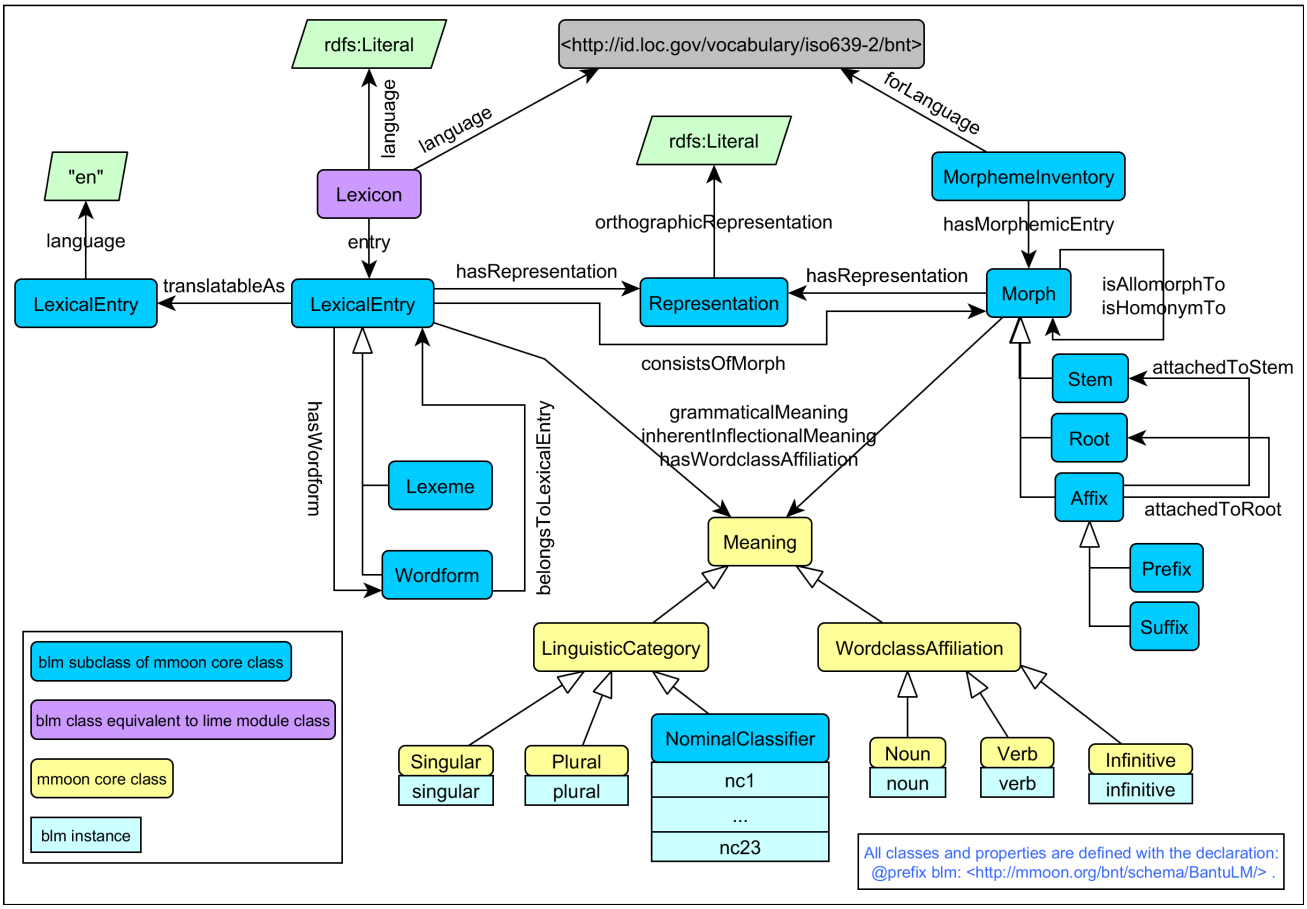


Figure 1: Ontology for the Bantu Language Model.

4. The Xhosa RDF Data set

The creation of the BantuLM ontology enabled the conversion from the Xhosa tabular source data into the Xhosa RDF data graph without any data loss. Necessary meta data is explicitly stated within the data set declaring information such as the data set creator, version and the underlying license.

In addition to the source data, the ontology-based representation of the Xhosa language data allowed for an explication of indirectly contained linguistic information. This is exemplified in Figure 2 which illustrates the graph representation of the lexical and morphological data.

With regard to the lexical data it can be seen that unique lexeme resources, like `xho_inv:lexeme_umbiki_n`[13] and `xho_inv:lexeme_ababiki_n`[14], have been created which were formerly separated as root and affix entries within the tabular data. As for the morphological data, the relationship that holds between affixes could be further specified by making use of the two object properties `blm:isAllomorphTo` and `blm:isHomonymTo`. That is, Figure 2 shows that the prefixes *aba-* and *ab-* are allomorphs to each other since they share the same meaning (noun class 2 and plural) but differ in their orthographic representation. Not illustrated, but included in the data set, are the homonymous relations that hold between affixes that share the same orthographic and/or phonological representation but differ in meaning. Such detailed linguistic information might be very useful for linguistic research investigating Bantu noun class systems.

Next to this internal enrichment of the tabular source data, the Xhosa RDF data set has been also externally enriched by linking the English translations, e.g. `xho_inv:trans_reporter_n` to lexical entries of the WordNet RDF data set[15] (McCrae et al., 2014). The object property `owl:sameAs` has been used to automatically create appropriate links. The full equivalence between the Xhosa RDF and WordNet RDF lexical resources is assured because only those lexemes have been interlinked that consisted of exactly one and the same word and also agreed in their part of speech. Figure 2 shows an example linking

[11] `http://www.w3.org/ns/lemon/vartrans#`

[12] To examine the increased expressivity, please compare an example of the source and RDF data here: `http://mmoon.org/lrec2018figures/`

[13] `http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/lexeme_umbiki_n`

[14] `http://rdf.corpora.uni-leipzig.de/resources/xho/inventory/lexeme_ababiki_n`

[15] Please cf. `http://wordnet-rdf.princeton.edu/about` for more information. The data set can be found here: `http://wordnet-rdf.princeton.edu/static/wordnet.nt.gz`.

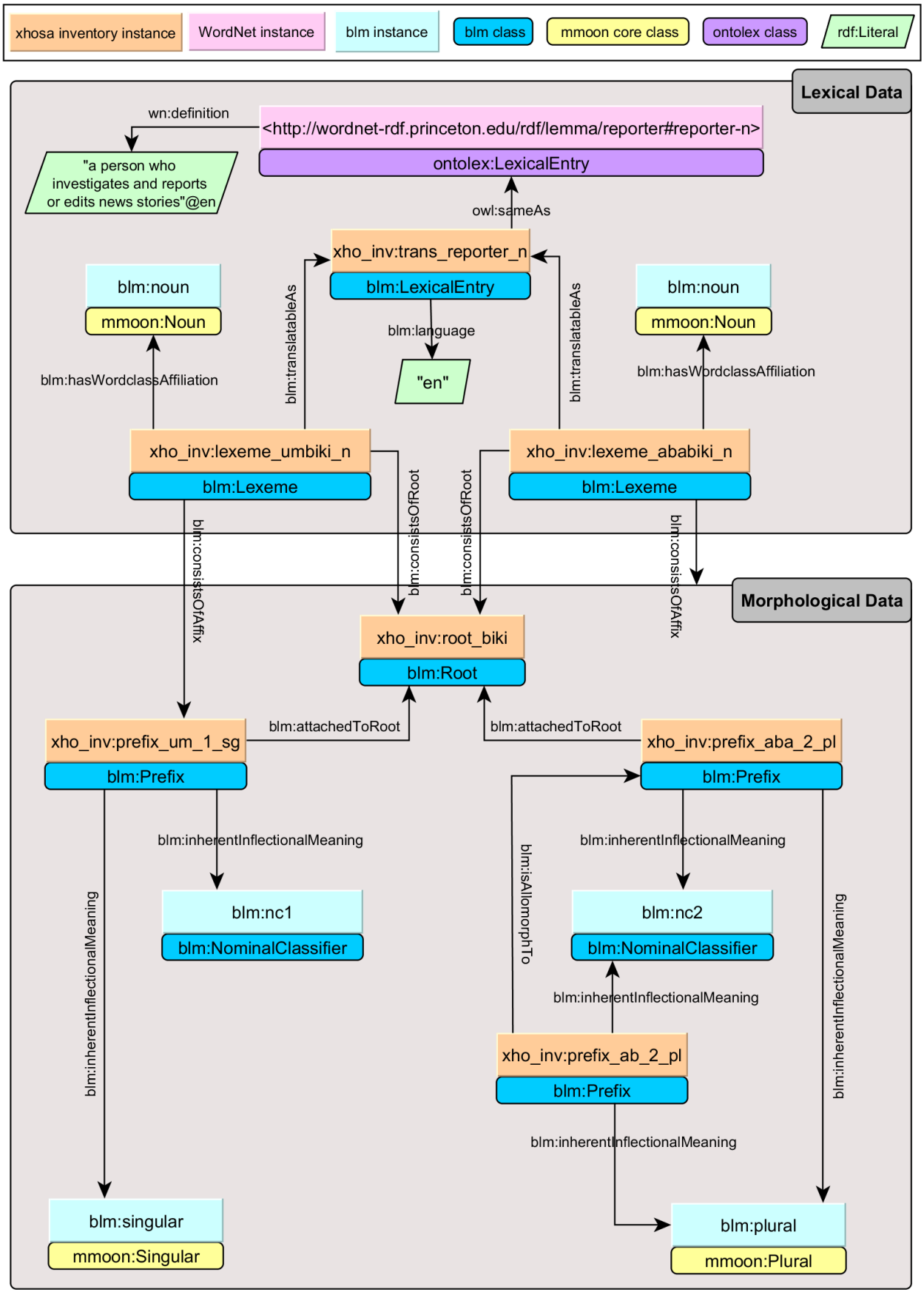


Figure 2: Excerpt from the Xhosa RDF data graph.

for the English translation *reporter* of the Xhosa nouns *umbiki* (singular) and *ababiki* (plural) to the corresponding WordNet lexical entry. Further, it can be seen that this WordNet entry ultimately[16] leads to a sense definition of the lexeme *reporter*. As a result, the interlinking of the Xhosa English translations with the WordNet RDF lexical entries, consequently, leads to an enrichment of the Xhosa noun and verb lexemes with corresponding lexical senses. Senses or sense definitions have not been part of the source data but are now accessible for all Xhosa lexemes whose translations are linked to WordNet and can be obtained by traversing through the interconnected data graph. While this enrichment with lexical senses already leads to a more coherent lexical data set for Xhosa, the linking to WordNet RDF entails an additional value in the context of the multilingual Bantu language landscape. Provided that more Bantu language data sets will be similarly converted into RDF and interlinked with WordNet, an interconnection of different Bantu language data sets could be realised by using the WordNet RDF as the pivot data basis for a multilingual Bantu language data graph.

Finally, the Xhosa RDF data set has been validated by using the RDF Unit[17] (Kontokostas et al., 2014) which conducts syntactic and semantic data quality tests of RDF data, which have been all passed by the Xhosa RDF data set.

In summary, the presented Xhosa RDF data set generates an added-value in comparison to its underlying tabular source data due to the successful internal and external data enrichment just explained. The Xhosa RDF data set in its current state contains 4,014 noun and 2,763 verb lexemes, 66 affixes as well as 2,818 links from the English translations to WordNet RDF. The Xhosa RDF data set is available within the LLOD Cloud and also accessible here: `https://github.com/MMoOn-Project/OpenBantu/blob/master/xho/inventory/ob_xho.ttl`. Moreover, the SPARQL endpoint provided at the URL `http://rdf.corpora.uni-leipzig.de/sparql` enables the querying of the data set to obtain deeper insights into the Xhosa language data.

5. Lexicographical Infrastructures in a Federated Environment

Despite strong efforts and significant progress towards open access to linguistic resources over the last years, many languages still lack those resources or their uncomplicated availability for larger user groups. Therefore, the presented work should not only be seen as another building block for a more complete landscape of linguistic resources, but in the context of federated and distributed infrastructures in a sometimes complex political and administrative environment.

Many countries with heterogeneous linguistic environments have decided to promote joint efforts for documenting their native languages for the benefit of education — primary, secondary, and academic — or the promotion of language technology, which currently is often only available for a highly resourced subset. This is especially problematic in a larger context where rights on relevant resources are held by different institutions with a varying degree of openness and each providing their own proprietary access interfaces.

As a consequence of this rather typical situation many large-scale infrastructures in the field of linguistic resources promote the usage of service-oriented architectures (SOAs) that provide data and services via standardised Web interfaces and data models. One of the benefits of this approach is that data can still be hosted by the publishing institution — being the main authority for the specific resource — and still allow access for the broader (or academic) public while promoting use and re-use in an active research environment.

In the South African context, the recently established Centre for Digital Language Resources[18] (SADiLaR) is a new research infrastructure with a focus on the creation, management and distribution of digital language resources of all official languages of the country. The ultimate aim is to provide a central repository for reusable language resources as well as applicable software tools that will be made freely available for research purposes (cf. Roux (2016)).

In the European context, CLARIN-D (cf. Hinrichs and Krauwer (2014)) is a long-term digital research infrastructure for language resources in the Humanities and Social Sciences. This includes language data bases, highly interoperable language technology tools as well as web-based language processing services. Researchers and students of Humanities and Social Sciences can use resources and technologies easily and in a standard way, without having to deal with technical complexities. The CLARIN-D infrastructure is built upon a network of centres, each of which with its own established competence and international reputation. For the time being, the described resource is hosted via CLARIN-D's infrastructure.

In our work we utilize this approach of making data available based on a standardised data model, i.e. the MMoOn Core ontology as the main basis of the BantuLM ontology, that has already proven to be adequate for describing morphological and lexical data (Klimek et al., 2016) and that is especially suitable to be used for other members of the Bantu language family as well.

The strict separation of data model, technical interface and end-user applications in a service-oriented environment opens the data for innovative applications. Among others, this is especially relevant for the field of meta-lexicography in the context of a multilingual environment. Besides the benefit of combining resources hosted and administered in different locations by different institutions, a SOA is a suitable backbone for enhancing usability with the major aim of addressing and reaching new user groups. This can be established by creating specific portals for different target audiences with varying and partially incompatible requirements. The specific demand may range from looking up simple words for language learners

[16] Please note, that there are several nodes in the WordNet RDF graph between the lexical entries and the sense definitions which are omitted in the Figure.

[17] `http://aksw.org/Projects/RDFUnit.html`

[18] http://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/p-ctext/documents/Graphics.RMA.Newsletter.1.0.3.LvdB.2016-11-23.pdf

to concrete usage examples for dictionary enrichment or highly specific information of different linguistic fields for academic studies. Naturally, aspects such as necessary functions, form and content aspects and intended use are playing a vital role here (Gouws et al., 2007).

6. Summary and Outlook

The presentation of a new Xhosa lexicographical resource for a multilingual federated environment is an example for the transformation of isolated and unpublished dictionary data to the digital age. However, the data set used to develop the BantuLM ontology is only a snapshot of a resource in development. Currently, more lexemes are curated and quality assurance methods will be used to improve the already available data constantly. The publication date of the final data set is expected to be within the next 15 months.

The Bantu Language Model described in this paper can be used for many more languages. Dictionary data is available in a variety of formats, see, for instance, *http://www.cbold.ish-lyon.cnrs.fr/Dico.asp* with dictionaries for about 70 Bantu languages with 5,000 to 10,000 entries per dictionary.

A next logical step is the construction of a user interface to use this data as an actual online dictionary. For comfortable dictionary look-up an additional morphological analysis would be helpful. Again, a unified approach for many Bantu languages seems possible here. As most existing dictionaries translate to English or French, the transitive connection of several dictionaries can be used to interconnect different Bantu languages and allow their combination to a joined "virtual" resource for the whole language family in the future.

Acknowledgment The various phases of research activities related to this paper were funded by grants from the: H2020 EU projects ALIGNED (GA-644055); Smart Data Web BMWi project (GA-01MD15010B); BMBF project CLARIN-D (01UG1620C); South African Centre for Digital Language Resources (SADiLaR); Erasmus+ Programme; Scientific eLexicography for Africa project; and the South African National Research Foundation. The late JA Louw is acknowledged for making the Xhosa data available for purposes of further developing Xhosa language resources.

7. Bibliographical References

Chavula, C. and Keet, C. (2014). Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In *OWLED*, pages 61–72.

De Schryver, G.-M. (2014). Oxford school dictionary: Xhosa-english. Cape Town. Oxford University Press Southern Africa.

Gouws, R. H., Heid, U., Schweickard, W., and Wiegand, H. E. (2007). Dictionaries. an international encyclopedia of lexicography. supplementary volume: Recent developments with special focus on computational lexicography. an outline of the project. *Edited by Fredric FM Dolezal et al.*, page 262.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, May.

Klimek, B., Arndt, N., Krause, S., and Arndt, T. (2016). Creating Linked Data Morphological Language Resources with MMoOn – The Hebrew Morpheme Inventory. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz.*

Klimek, B. (2017). Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model*, pages 68–73.

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 747–758, New York, NY, USA. ACM.

Kosch, I. M. (2006). Topics in morphology in the african language context. Unisa Press.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259, Berlin, Heidelberg. Springer.

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: Development and applications. In *Electronic lexicography in the 21st century: Lexicography from Scratch. Proceedings of the eLex 2017 conference.*

Nurse, D. and Philippson, G. (2003). *The Bantu languages*. London: Routledge.

Pahl, H. (1967). *isiXhosa*. Johannesburg: Educum.

Roux, J. C. (2016). South African Centre for Digital Language Resources. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), 23-28 May 2016, Slovenia, Portoroz.*

Chapter 3

Synopsis

The presented six publications can be regarded as individual studies of differing research topics centred around the creation and usage of morphological language data. As indicated in Section 1.1 the outcomes of these works establish the three stated and required prerequisites to generally enhance the cross-disciplinary usage of language data by semantically modelled and represented morphological data. Within this synopsis it will be explained in detail to what extent they are initiating and contributing to the induction of cross-disciplinary morphological data usage. Therefore, all publications will be summarised below and interrelated in terms of their content according to the thesis' topic. First, the two central areas of morphological data and semantic modelling will be introduced and located within the field of digital humanities. Beyond that, their scientific relevance for intersecting disciplines will be illustrated (Chapter 3.1). Second, the outcomes of each individual publication will be summarised and placed into the larger context of related works (Chapter 3.2). Subsequently, the implications and limitations underlying these results with regard to their impact on further research will be deduced (Chapter 3.3).

3.1 Cross-Disciplinary Relevance

3.1.1 Morphological Language Data

The selected publications contributing to this thesis are based on the traditional linguistic perspective on the study field of morphology. Morphology constitutes one of the core sub-fields of linguistic studies next to phonetics and phonology, syntax, semantics and pragmatics. A clear delimitation of these fields is not possible. As intersecting fields like morpho-phonology or the syntax-semantics interface have shown, there exists already a highly intradisciplinary relevance affecting the studies of these sub-fields of linguistics among each other. This coincides with the difficulties described for the creation but also alignment of the MMoOn Core ontology (cf. [P2] and [P3]) and the OntoLex-*lemon* modules (cf. [P4]) for representing the data

of these fields in domain ontologies which are not only troubled with setting the respective domain boundaries but also with establishing an adequate interconnection between them. Additionally, every linguistic sub-field is divided into two study areas, i.e. the theoretical and methodological foundation of the field itself and the concrete application of it describing a specific language. The kind of data attributed to the former kind of study area shall be defined as **linguistic data** which is characterised as being language-independently applicable to natural language as such. The latter kind of study area of each linguistic sub-field will yield datasets which represent the instantiation of the theoretical foundation for a specific language and is, thus, defined as **language data**. For the field of morphology this entails resulting data describing the underlying phenomenology of words, sub-word units and their meanings which are mainly documented in unstructured formats such as textual documents (for example in works like Haspelmath & Sims (2013) or Booij (2012)). The results of identifying and describing the morphological inventory of a single language or language family are again mostly textual documents such as print dictionaries and grammars but also increasingly structured datasets such as tables or lexical databases. Consequently, the research area of morphological language data, being rooted in the linguistic sub-field of morphology, is characterised by mutually dependent intradisciplinary aspects. Studying and creating morphological data, therefore, entails the consideration of further overlapping linguistic sub-fields if necessary. It moreover requires the representation of language-specific morphological data in conjunction with their underlying linguistic theories which are in turn distributed across the fields of lexicography and grammaticography and differ strongly across different disciplines.

During the long history of the study of language the scientific field of linguistics gave rise to a multitude of complementary study areas taking up linguistic and language data (which per definitionem always includes morphological data) to differing extents. In this thesis, the term **cross-disciplinary usage**, therefore, is applied by its definition as "knowledge acquisition gain in one discipline that is achieved by the reuse of language data that was originally produced within another discipline". In this broad sense it subsumes **multidisciplinary**, **interdisciplinary** and **transdisciplinary usage** as sub-terms thereof. Figure 3.1 gives an overview of exemplary research fields according to these disciplinary interrelations to morphological language data. In the centre of this diagram is the domain of morphological language data to which exemplary disciplines are attached with different degrees of overlap. The usage of morphological data in linguistic fieldwork can be regarded as interdisciplinary since their creation and analysis are synthesised with the ongoing research of morphology. The data basis and outcomes of both disciplines contribute to insights in both of them (Payne, 1997). A transdisciplinary usage is illustrated for the discipline of computer linguistics. Its boundary to morphology is stronger

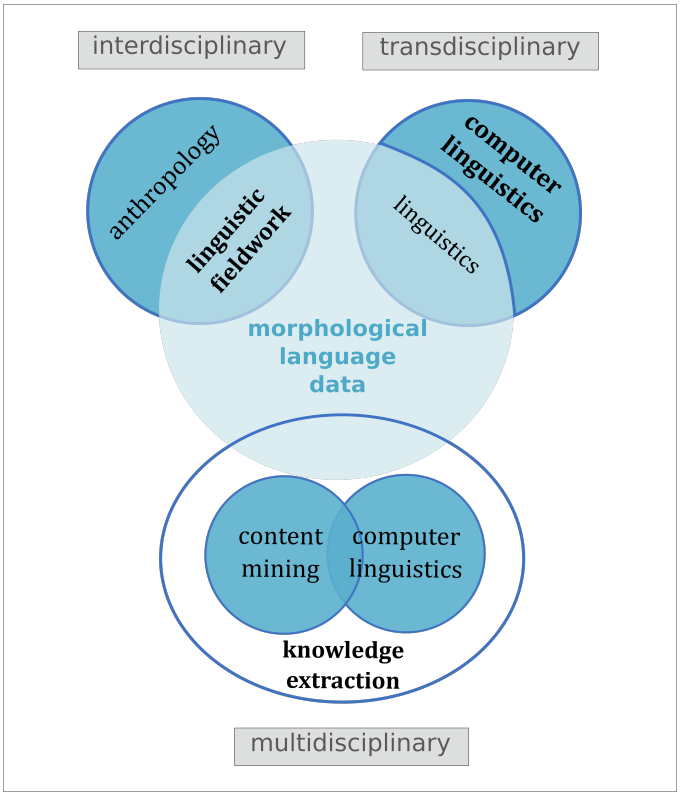


Figure 3.1: Cross-disciplinary usage of morphological language data.

with regard to the scope of the language data usage. While the linguistic research involved aims at understanding natural language for human consumers, computer linguistics is merely focused on machine processing of natural language. Moreover, both disciplines reuse the results of each other, however, without contributing new insights to the other discipline. The focus is on the application of the outcomes with a mainly basic understanding of the theoretic foundations. This means that the linguistic knowledge or data which is also part of morphological language data is not crucial for the methodologies applied in computational linguistics. Finally, a multidisciplinary relation is displayed for the very specific task of knowledge extraction. Here, the three fields of (morphological) language data, content mining and computer linguistics remain mostly autonomous in that they synthesise only very specific results with a very low reusability of the outcomes for the individual disciplines. Further, it has to be noted that not only autonomous disciplines are involved but also disciplines which in themselves are already interdisciplinary. In this example content mining, when dealing with language resources, overlaps with computer linguistics which in turn overlaps with linguistics in a transdisciplinary way. The linguistic foundations of the morphological data used are not considered at all anymore. What is more, the language data already could have undergone several processing steps and deviate largely from the original data which usually came with more detailed linguistic information. The boundaries of

individual disciplines are dissolving in as far as knowledge acquisition gain in multidisciplinary research is directed towards single problems instead of an entire discipline. From this overview it can be concluded that every discipline that relies on or uses morphological language data benefits from a knowledge acquisition gain. Therefore, a significant cross-disciplinary relevance can be stated for a shared reuse of morphological data. However, this differs with regard to the degree of understanding the inherent linguistic basis and the size of its application area. While interdisciplinary scholars have a sound knowledge of the linguistic and morphological language data used and directly contribute to the field of morphology, the transdisciplinary scholars have a basic understanding of it and in multidisciplinary research scholars from individual fields work together on solving a single complex problem. The two latter disciplinary interrelations also show that the smaller the research impact, i.e. on only a part of the discipline or even a single research problem, the likelier it is that only fractions of original morphological data are reused and that new insights and datasets are less reusable for morphological research in turn. A considerable potential to increase the reusability rate of the morphological data emerging across such disciplines does, nonetheless, exist even if not exploited at present due to other reasons (cf. Chapter 3.1.2).

In consideration of the inherent intradisciplinarity of morphological data indicated above in conjunction with the outlined kinds of its reuse across other disciplines, it can be observed that morphological data seems to be not only highly relevant for entire disciplines which rely on natural language data. Additionally, it is also of importance for specific tasks and problems which require a high degree of multidisciplinary methods and data in which it represents an additional research piece. As a consequence, this situation yields fuzzy discipline boundaries in which a relevance can be certainly attributed to the morphological data used within another discipline, however, leaving the identification of the concrete knowledge acquisition gain due to the morphological data more or less specific.

Similar findings are described in Van Leeuwen (2005) which approaches the topic of cross-disciplinarity with three, usually distinct models of interdisciplinarity (“centralist”, “pluralist” and “integrationist”) which have become co-existent to date. Especially the latter and youngest model reflects the future trend of cross-disciplinary usage of linguistic and language data. The definition of the integrationist model contains two noteworthy aspects. First, it regards disciplines as independent and team-work-based research projects that are problem- rather than method-focused and second, it recognises “that no single discipline can satisfactorily address any given problem on its own” (cf. Van Leeuwen, 2005, p. 8). Both appear even more appropriate in the prospect of an expectable increase of digitisation within science and research.

In fact, an entire academic field, i.e. the digital humanities, can be re-

garded as the integrationist model in practice. In contrast to other inter- or transdisciplinary fields that are often a combination of two autonomous disciplines, the digital humanities are not only embracing multiple autonomous disciplines of the humanities but also their already transdisciplinary successors, e.g. linguistics and computational linguistics. Against this background the circumstance that this field has difficulties to define itself and to clearly demarcate discipline boundaries (Vanhoutte, 2013) can as such be regarded as a defining characteristic of the digital humanities. In this way, as a reaction to the increasing diffusion of knowledge and technology in science, the identity of this field manifests itself by various scholars of different scientific backgrounds who together are shaping the field by detecting the multidisciplinary potential of cross-related fields and by constantly adapting in terms of methodology and innovation.

Embedded into this scientific setting of the digital humanities is morphological language data as the central topic of this thesis. It ranges from traditional linguistics, including all affected fields from the humanities, to cross-related disciplines from the area of computer science, such as natural language processing, knowledge extraction and artificial intelligence. The outcomes of this development can be observed in a rising degree of cross-disciplinary relevance through morphological and general language data that is incorporated in more and more specific research tasks. Nonetheless, the accompanying impact this integrationist research practice has on the knowledge acquisition gain of the individual disciplines using morphological language data should not be underestimated. Thus, morphological datasets emerging from non-linguistic fields that have been tailored to a particular problem solution often exhibit a reduced linguistic adequacy. With the progressing opening of discipline boundaries knowledge gain is no longer bound to a single discipline, however, "at the cost of limitations to understanding and expertise" (cf. Frodeman, 2013, p.3) which directly affects the data quality and limits the reusability of this data in its originating field, i.e. linguistic research. Therefore, the six presented works in this thesis pursue a sustainable data-driven and linguistic-centred approach aiming at providing a semantic data representation basis that enables every cross-related discipline to create linguistically profound morphological datasets in order to increase the quality and reusability for all disciplines relying on such language data.

3.1.2 Semantic Data Representation

As already indicated, morphological language data is not reused in accordance to its persisting relevance across different disciplines. Reasons for this situation can be attributed to the still ongoing implementation of the FAIR Guiding Principles for scientific data in general (Wilkinson et al., 2016) which state that data needs to be findable, accessible, interoperable and reusable. For the documentation of language data in particular

the seven dimensions of content, format, discovery, access, citation, preservation and rights (Bird & Simons, 2003) are additionally impeding their portability across different disciplines. From the identified principles and dimensions this thesis focuses primarily on the aspect of the format which is summarised as semantic data representation and indirectly also affects the other aspects of content, interoperability, reusability, discovery and citation (for details on the effects of the data format on these aspects cf. Chapter 3.3.1).
This comparably young kind of data representation emerged within the so called Semantic Web envisioned by Tim Berners-Lee. It overcomes the Web of documents which is only understandable by humans in that it explicates the underlying content and the data of these within the infrastructure of the Web in order to render it comprehensible for machines (Berners-Lee et al., 2001). Semantic data representation according to the Semantic Web builds on two main formats. The first one is the Resource Description Framework (RDF). Within RDF every possible datum is regarded as a unique resource which is identifiable through the Web-inherent Uniform Resource Identifiers (URIs). All information, i.e. other data and knowledge, about a resource is described with RDF in the form of statements or so-called triples, each consisting of a subject, predicate and object. This way of data representation enables not only machine-processability but also allows for a direct data distribution across the Web in a structurally homogeneous way. The second central format of the Semantic Web is the Web Ontology Language (OWL). Every resource is formally described for its general semantics with OWL. It consists of a hierarchical class structure, instances which are the resources as members assigned to these classes and properties that establish relations between the instances. Numerous ontologies have been created based on OWL which represent a specific knowledge domain and are semantically richer. The ontologies serve as a machine-readable vocabulary that allows to reason over data since the data itself as well as its underlying semantics are represented in RDF and OWL. Moreover, as a result of these two formats, data can be interconnected across multiple datasets and thereby enriched with external information. This kind of data is then called Linked Data. All Linked Data datasets published under an open license constitute the semantically interrelated Web of Data which is visualised by the Linked Open Data (LOD) Cloud diagram[1]. To date it displays 1239 datasets of nine large knowledge domains, each of which is interlinked with at least one other dataset.

Consequently, the semantic data representation in terms of Linked Data enables the interoperability of data not only format-wise but also content-wise. These two aspects are especially advantageous in the given context of the cross-disciplinary usage of morphological language data. Ideally, such a reuse would manifest itself in that morphological data originating within

[1]https://lod-cloud.net/

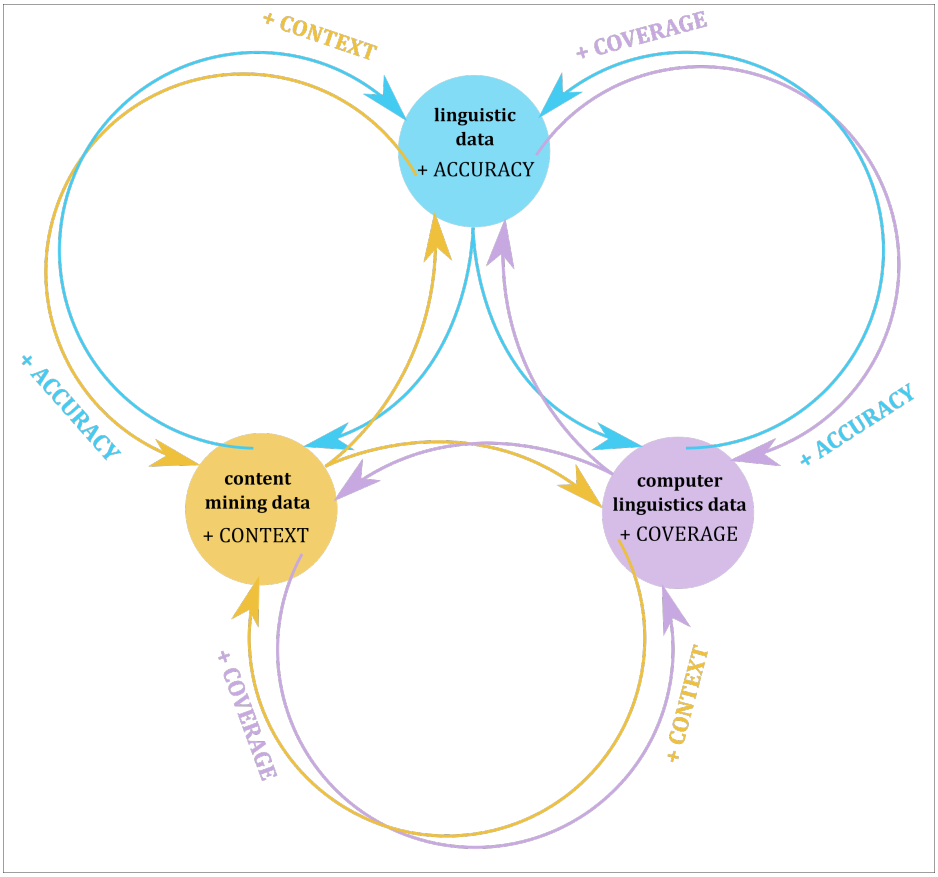


Figure 3.2: Reuse cycle of interoperable morphological language data.

one field is reused by other disciplines to reach better results from this data integration. This envisaged data exchange is illustrated with a reuse cycle for morphological language data in Figure 3.2. Research areas using and producing language data are for the reason of simplification reduced to three fields in this diagram (including possible interdisciplinary overlaps as outlined in 3.1) which stand for different user groups. Linguistic data is mainly produced by linguists for linguists with the focus on investigating natural language for human understanding. Less linguistically accurate data is created and used by and for computer linguists in order to enable large-scale machine-processing of natural language. For the field of content mining language data is used and produced by non-linguists only as a means to exploit knowledge which is encoded in natural language. It has neither a primary interest in improving machine-processability nor in a human understandable investigation of natural language. Thus, each research field in this diagram accounts predominantly for one aspect that is significant for both the human and machine usage and creation of language data. Data emerging in the area of linguistics is highly accurate and elaborate but since compiled mostly manually covers only small amounts of language resources. The coverage is highly increased by computer linguis-

tics' datasets which in turn render linguistically less accurate results. For the area of content mining multiple linguistic and computer linguistic resources and tools are applied simultaneously to explore a specific thematic context. The obtained results are linguistically even less accurate and in terms of coverage only applicable to the selected, often non-standardised, textual basis. Still, they pose a valuable source for the field of linguistics in that they provide the necessary empirical data foundation for investigating the linguistic encoding of certain semantic fields, e.g. possession, sentiment, time, motion, toponymy or evidentiality within cultural, social and political contexts. This enables the identification of new linguistic constructions or elements and also contributes attestations for the verification of prevalent assumptions. Conversely, linguistic datasets including these results can in turn improve the outcomes of content mining tasks since a large amount of knowledge is still encoded in natural language. A cyclic reuse of external disciplinary morphological data as indicated from the inner to the outer circles for each discipline in Figure 3.2 would, therefore, increase the knowledge acquisition gain for each discipline individually by taking data from other disciplines into account that compensates for two of the missing aspects of accuracy, coverage and context, respectively.

This kind of data reuse has been prevented so far due to the different data formats used by the user groups. Usually, data reused from another discipline is reproduced in other formats, often accompanied with data loss and remains unrecognised by external researchers hidden in publications or data silos, e.g. linguistic examples taken from a text corpus reappearing in PDF documents of publications. Linked Data and semantic modelling overcome this issue and enable an efficient data reuse that preserves the original data and makes it both machine processable and understandable for humans in the envisaged cross-disciplinary manner. The RDF data format does not only realise an interoperable data usage, but also enhances the accessibility of the data through its underlying Web integration via URIs. By interlinking resources of different datasets a redundant data reproduction becomes unnecessary and formerly isolated data contributes not only to the enrichment of but also benefits from already existing data. While RDF and Linked Data already ensure a format-wise machine interoperability, the semantic modelling with the use of ontologies additionally provides a shared semantics underlying all datasets that renders the data also content-wise interoperable for machines. Consequently, a portability of morphological or linguistically relevant data originating from one discipline is achieved to be directly integrated into datasets of another discipline's research application scenario. This ultimately provides the foundation to realise the illustrated cross-disciplinary data usage cycle.

As illustrated so far, semantic data representation constitutes a promising approach in order to close the existing gap in the cross-disciplinary reuse of morphological language data. In fact, Linked Data is already successfully

put into practice within the field of digital humanities and partially also in (cross-)linguistic research areas. Early on it has been discovered that a better understanding and new insights can be gained from the data accumulated across the fields of the humanities by overcoming format barriers and enriching existing datasets with information compiled in other fields by interlinking them (Blanke et al., 2012). Emerging RDF-based data collections of biographical, cultural, historical, geographic, and bibliographic data are valuable contributions that are increasingly integrated into digital humanities research (cf. for example Nurmikko-Fuller et al. (2016), Baierer et al. (2017), Hyvönen et al. (2019), Ciotti et al. (2014)). As a consequence, efforts to transform existing data into RDF and to provide ontology-based knowledge foundations that facilitate a shared data usage are regarded as an integral part of the digital humanities' technology stack (Berry & Fagerjord, 2017).

Linked Data is also implemented in the field of linguistics. Especially the areas of language documentation and typology make use of the interoperability gained through the RDF format. The General Ontology for Linguistic Description (GOLD[2]) has been one of the earliest efforts to establish best practices for encoding linguistic data (Farrar & Langendoen, 2003). Another initiative applying Linked Data is the Cross-Linguistic Linked Data (CLLD[3]) project (Forkel, 2014) which provides an infrastructure for facilitated data publication and interconnection for data providers and consumers. To date sixteen datasets, including large and renowned datasets like the World Atlas of Language Structures[4] and Glottolog[5], have been converted into RDF and are maintained by the CLLD project. In this area of comparative linguistics the numerous resources such as word lists, dictionaries and other structured datasets exhibit an increased reuse due to cross-linguistic data formats (Forkel et al., 2018) coming with a basic ontology, a software package for validation and manipulation and links to more general frameworks which also enable a tool-based reuse. This outlined adoption of the Linked Data framework is, however, centrally focused on the format-wise structural interoperability that is realised with Linked Data and enhances linguistic and language data publication, discovery and accessibility. At the same time as the CLLD project the LLOD community evolved dealing with language-specific content of linguistic datasets. Semantic data representation has been taken up as a means to interconnect all kinds of language data produced across various research fields. Linked Data is regarded as a possibility to increase cross-disciplinary reuse and data enrichment but also to gain new insights into single datasets by also linking them to non-linguistic domain data of the LOD cloud, e.g. to DB-

[2]http://linguistics-ontology.org/
[3]https://clld.org/
[4]https://wals.info/
[5]https://glottolog.org/

pedia[6] (Auer et al., 2007) as the central knowledge graph of the the LOD cloud. Numerous highly specific domain ontologies have been created by members of the LLOD community which cover the linguistic domains of lexicography, translation, terminology, fieldwork, grammatical categories, syntax, metadata and others. The mentioning of many of them within the related works sections of the accompanying publications in Chapter 2 in conjunction with the introduction of the LLOD community in Chapter 1.1 shall suffice to illustrate the high relevance and ongoing innovation of Linked Data and semantic data representation for the linguistic data domain.

However, for morphological language data comparable progress is still lacking. As presented in Section 2 of [P5] the large amount of resources containing morphological data prevalently exists in unstructured non-RDF formats provided by linguists. So far, efforts to create Linked Data vocabularies failed to reach the required granularity to describe morphemic elements. The LLOD cloud as renowned accumulation point for linguistic and language resources can be regarded as the current state of the implementation of the aspired reuse cycle described above. In order to extend this cloud with morphological datasets and providing these concomitantly for cross-disciplinary reuse, an adequate semantic data representation model is needed as the foundation for transforming the existing morphological datasets into RDF resources.

3.2 Summary of the Publication Outcomes

The following summary of the publication outcomes groups the six publications into three subchapters in accordance to the three prerequisites defined in Chapter 1.1. Each chapter can be regarded as a realisation of these requirements in order to support the central hypothesis of this thesis that semantically modelled and represented morphological data will enhance the cross-disciplinary usage of language data in general.

3.2.1 Publication 1

Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge

[P1] is motivated by the question if linguistic knowledge about the morphological complexity of a language and morphological data, respectively, can enhance the results of NER tasks. In fact, a large part of computational linguistics takes the position that knowledge about the morphological and syntactic fundamentals of linguistics enhance NLP approaches (Beesley, 2003; Bender, 2013). Supported by this viewpoint the conducted

[6] https://wiki.dbpedia.org/

investigation presented in this publication takes a strong linguistic perspective under the assumption that the computational linguistic concept of 'named entity' corresponds to the linguistic category of proper nouns. This implies that they are classified as nouns and as such likewise affected by productive morphological formation processes which can generate new and undocumented variations. The established data bases and methods in computational linguistic do, however, only account insufficiently for this aspect. Lexicon, rules and machine learning are regarded as the three major approaches to NER (cf. Gudivada, 2018, pp. 414-415.). Lexicons or gazetteers in this context are lists of named entities which are extracted from external knowledge sources. Usually, they do not contain detailed linguistic information and no morphological form variants but mostly their lexical canonical form. Additionally, such sources are limited in their scope depending on the external data source they have been derived from and fail to account for the infinite number of forms that can be created and could appear when the lexicon is applied to a new textual resource. The disadvantage of the rule-based approach lies in the challenge to create a comprehensive set of morpho-phonological word-formation and inflectional rules, which are also highly language-specific and difficult to reuse. Finally, machine-learning methods rely on a large amount of training data in order to enable a machine to identify unknown named entity tokens. Concrete corpus data, such as the investigated GermEval NER challenge corpus in [P1], reveals the actual impact of morphological processes applied to proper nouns in natural language production and their degree of variation. In order to determine if morphological data for proper nouns can complement and, therefore, improve the current lexical data bases as well as the methods and results in NER tasks, a thorough linguistic examination had to be conducted. Hence, the outcomes of [P1] can be summarised as follows:

Morphological complexity analysis: The results of the linguistic exploration of morphological alternations of proper nouns in German are presented in Appendix A of [P1]. It provides the significant linguistic features that constitute the scale of complexity that is involved in the recognition of a proper noun contained in a text token. These features, i.e. inflection, inner modification, compounding and derivation, point out to what extent machines need to perform in order to automatically recognise named entities for German. In the realm of the three annotated subsets TPi, FNi and EN ExB the table in Appendix A represents only a fraction attesting the actual occurrences of morphological complex named entities. Taking the combinatory potential of the four chosen linguistic features into account the identified complexities indicate that additional variants are likely to occur in other text sources. Therefore, it can be assumed that the full scale of morphological complexity is not displayed by this data alone. To this extent, for every source named entity that occurred in the investigated data numerous further morphological variants can be linguistically formed

as well. For example, the features of inflection and inner modification were not examined in detail but increase the number of possible variations according to the inflectional paradigms of the token's word-class and the number of morpho-phonological alternations entailed in the inner modifications (e.g. the deletion, change or addition of letters in a token which impede the machines' performance of recognising the target named entity). Moreover, the conducted analysis reveals that the degree of morphological synthesis of a proper noun within a given token correlates with a lower machine performance. This does not only explain why systems perform better for English, which is a less inflectional language than German, but also indicates that languages exhibiting more synthetic or even polysynthetic morphology could benefit from taking morphological information or data for the task of NER into account.

Data-based assessment of morphological relevance and complexity: A main outcome of [P1] constitutes a comprehensive measurement of morphological relevance and complexity calculated based on empirical data and a manual linguistic segmentation procedure. From this the insight could be obtained that 30,7% of the named entities unrecognised by all systems are morphologically relevant and 23,4% are even morphologically complex (cf. Figure 2 in [P1]). The best performing system could improve its results by 19,2% if morphological complexity would be targeted. Even though this amount seems considerably insignificant compared to the 69,3% of named entities that were recognised by all systems, the obtained data can be regarded as a significant contribution for further improvements of feature design and error analysis in the field of NER for German.

Identification of significant annotation issues: The documented annotation issues do not directly affect the analysis of morphological complex variants of proper nouns. However, they influence the evaluation of the system performances with regard to their ability to recognise morphologically complex named entities. Moreover, the identified annotation issues expose the general difficulties in establishing a shared understanding of a linguistic task such as NER. Therefore, they pose a valuable contribution to the progressing improvement of the training and test data provided by challenges such as the GermEval NER.

To conclude, the in-depth examination of the morphological complexity of German proper nouns is relevant for the performance of NER systems. What is more, it was shown that the linguistic segmentation steps and features that are necessary to identify proper nouns as derivational or compounding components of other lexical items supports the assumption that existing NER systems for German have difficulties to detect named entities that are morphologically complex. These insights promote the application of both a segmentational linguistic analysis and more morphological language data in cases where proper nouns strongly deviate from their lexical

canonical forms. These findings are supported by the fact that existing approaches incorporating morphology into NER for other morphology rich languages, e.g., Farber et al. (2008) for Arabic and Marcinczuk et al. (2013) for Polish, also record a remaining amount of unrecognised entities. This might be attributed to the well-known issues of the computational segmentational approach (in contrast to a human analysis). These include difficulties with processing stem alternation, sound changes, zero morphs and words or word parts of foreign origin (cf. Janicki, 2019, pp. 6-9), which are additionally challenging in the context of recognising often unknown proper nouns in morphological variations. Therefore, the obtained results of [P1] indicate that the current computational approaches seem to reach a performance limit that might be overcome by considering a hybrid method which also includes manually segmented morphological data in order to enable machines to cope with the most complex text tokens containing proper nouns. Consequently, the outcomes of [P1] could be used to improve the rule-based methods but also to enhance existing named entity lexicons and machine-learning approaches with a larger data-driven basis to identify remaining unrecognised named entities.

The envisaged morphological data that should result from the conducted linguistic complexity analysis encompasses data describing identified proper nouns more extensively, i.e. including their inflectional word-form variants, lexemes of which they are the derivational bases or a compound element as well as derivational and inflectional affixes together with their meaning. Such data would constitute a valuable extension of the existing lexicons that are created for NER and the closely related task of entity linking. An example for an RDF representation of the source token *Skialpinistinnen* is given in Listing 3.4 in Chapter 3.2.2, where the target named entity *Alpen* is explicitly contained in the data. The DBpedia Spotlight[7] tool, for instance, automatically annotates text input for (named) entities based on a lexicon derived from labels of Wikipedia article names (Mendes et al., 2011), i.e. the lexicon approach. Additional morphological data added to these labels would instantly increase the number of identified and linked named entities, as for example *Alpen* in *Skialpinistinnen*, since embedded proper nouns in derived or compound words would become an inherent part of the lexicon which is already disambiguated and interlinked.

The presented outcomes of [P1] can be regarded as evidence that results of cross-disciplinary tasks can be improved with this kind of language data. For the specific task of NER it could be shown that a complementary use of morphological data in conjunction with the commonly applied computer linguistic lexical data and methods can lead to a significant increase in knowledge acquisition gain, i.e. better results in NER. This includes also the outcomes obtained regarding the identified annotation issues since better annotated corpora lead to better systems trained on them. What is

[7]https://www.dbpedia-spotlight.org/

more, the development toward the application of Linked Data and knowledge bases for named entity linking would also benefit from morphological data usage if this data would be described with the interoperable RDF format.

3.2.2 Publication 2, 3, and 4

Motivated by the findings just outlined for [P1] the developed semantic model that allows to create morphological data in RDF will be presented in this chapter. The three publications [P2], [P3] and [P4] will be summarised with regard to the second requirement defined in Chapter 1.1 that an adequate ontology is available and enables the creation of semantically represented and interoperable morphological data. It has to be noted that even though the OntoLex-*lemon* Morphology Module extension emerged under the influence of MMoOn Core and the supervision of the author, the main focus for the evaluation of this requirement rests on the MMoOn Core ontology.

Publication 2: MMoOn Core – The Multilingual Morpheme Ontology

The MMoOn Core ontology as the proposed semantic representation model has been developed with the aim to yield morphological datasets which induce the cross-disciplinary usage of morphological but also linguistic data in general. Therefore, these two fields of information science and linguistics have been combined and the substantial approaches and insights applied to each other. On the one side, MMoOn Core is based on the computer scientific foundation of ontologies defined as "a level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals" (Gruber, 2016). On the other side, it emerged from the structuralist linguistic view on the field of morphology which defines it as "the study of systematic covariation in the form and meaning of words" (cf. Haspelmath & Sims, 2013, p. 2) together with the central concept of 'morpheme' as "the smallest meaningful part of a linguistic expression that can be identified by segmentation; a frequently occurring subtype of morphological pattern" (ibid., p. 335).

In theory a similar approach has been modelled already over three decades ago in Hudson's "Word Grammar" which is a theory of language structure treating "language [as] a network of entities related by propositions" (cf. Hudson, 1984, p. 1). This network is formalised and explicated under the assumption "that 'linguistic' knowledge is knowledge of 'linguistic' entities, and linguistic entities are words or their parts" (cf. Hudson, 1984, p. 36). Against this background the MMoOn Core ontology presented in [P2] can be regarded as one possible corresponding practical im-

plementation taking up this fundamental idea and realising it by the means of ontological modelling provided by the Linked Data principles. [P2] constitutes its specific application to the knowledge domain of morphological data in accordance with the definitions of *ontology* and *morphology* given above. The main outcomes of this publication are given below:

Linguistic documentation: As indicated in [P2], the MMoOn Core ontology presents the first consistent semantic representation model for the domain of morphological data following a thorough domain analysis (cf. Section 4 in [P2]). The naming and definitions of the ontological elements, i.e. classes, properties and instances, have been chosen to be conceptually and terminologically grounded in the field of linguistics. By doing so, the representation needs of linguists in documenting languages have been deliberately favoured under the assumption that data originating from linguistics is of high quality and usually more fine-grained. Therefore, MMoOn Core encompasses concepts like 'morph', 'morpheme', 'derivation', 'inflection', 'allomorphy' or 'morphemic gloss' which are not covered in other existing linguistic vocabularies to this extent. Overall, consisting of a comparably high number of classes, properties and individuals the MMoOn Core ontology can be regarded as especially suitable for morphological language data documentation for linguists. Linguistic elements are available to a very fine-grained extent encompassing a large variety of linguistic and grammatical categories. This considerable descriptive range facilitates the language documentation process since necessary linguistic concepts are provided within a single vocabulary and do not have to be obtained from external ontologies like LexInfo[8]. Because all ontological elements are not only conceptually defined, but also by their semantic and formal interrelation to each other, the MMoOn Core ontology inherently constitutes a practical environment and template for a morphological data compilation. In this respect, MMoOn Core is not only a data representation but also a documentation means. It enables a unified description and representation of morphological data that is yet predominantly and separately contained in lexicons and grammars. Ultimately, the usage of MMoOn Core as a semantic representation model leads to the creation of morphological datasets from which linguistic resources like morphemicons, full-form lexicons, root lists or customised data extractions can be easily derived.

Semantic interrelatedness: In addition to the MMoOn Core ontology its specified architectural setup ensures the semantic homogeneity of different datasets by enabling maximal descriptive granularity at the same time. It proposes to construct MMoOn morpheme inventories with a schema ontology and an inventory file that are based on an import of the MMoOn Core ontology in order to account for the requirement that the "descrip-

[8]https://lexinfo.net/ontology/3.0/lexinfo.owl

tion of a language has two opposite tasks: to bring out the uniqueness of this language and to render it comparable with other languages" (cf. Booij et al., 2004, p. 1859). Resulting from this, the full range of semantic interrelatedness between multiple morpheme inventories is covered even if they differ in terms of data creator, granularity and language. Figure 3.3 illustrates the two extremes of the spectrum which varies between a generic, language-independent and conceptual extreme and a language-specific instantiation of the conceptual framework. The former is realised

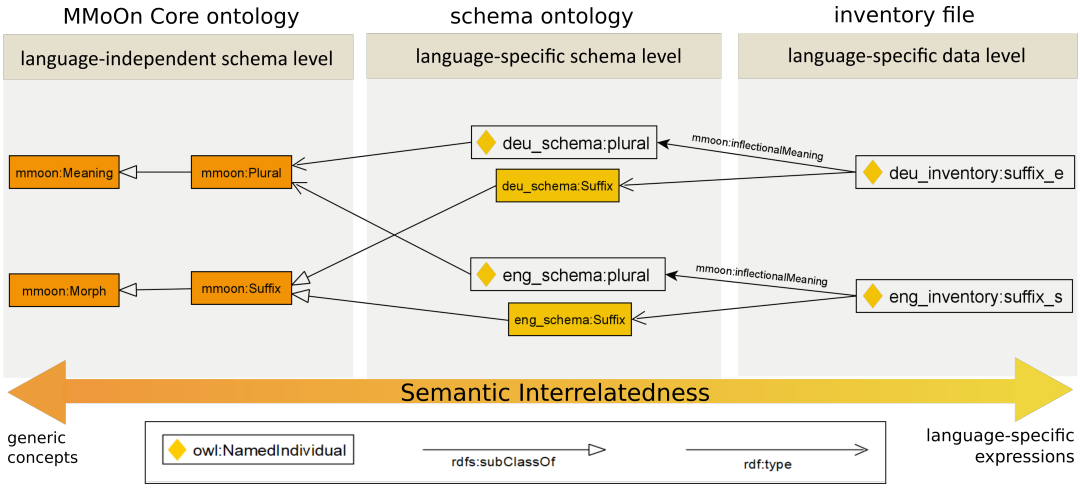


Figure 3.3: Spectrum of semantic interrelatedness between MMoOn morpheme inventories.

by the MMoOn Core ontology that ensures a cross-linguistic interconnection. Due to the provided subclass relations in the schema template file the instance data of independently created datasets is interconnected without requiring any alignment or explicit declarations, e.g. the German and English suffix subclasses both share the same `mmoon:Suffix` superclass. The other end of the spectrum is realised by the inventory data which is connected with the MMoOn Core ontology via the language-specific schema ontologies. These allow for a very specific and granular description of language-specific instances, e.g. the German and English nominal plural suffixes `deu_inventory:suffix_e` and `eng_inventory:suffix_s`, respectively. To that extent, means are provided to describe both the semantics of linguistic data, i.e. `deu/eng_schema:Suffix` and `deu/eng_schema:plural`, and language data, i.e. `deu_inventory:suffix_e` and `eng_inventory:-suffix_s` (as introduced in Chapter 3.1.1). By doing so the existence of a variety of different language-specific definitions of linguistic categories (as they are provided in grammars) is acknowledged and can be included as secondary language data descriptions in addition to the representation of the primary language data. As a consequence, linguists have the possibility to directly incorporate them on their on terms into their dataset while they were hitherto required to search for them in external vocabularies which conceptualise them as language-independent categories only. Overall the

semantic interconnection throughout the spectrum between cross-lingual language-independent concepts and language-specific instance data is enabled through the MMoOn Core ontology and inherently provided for every emerging dataset due to the ontology import into every language-specific schema ontology. Consequently, loading multiple MMoOn morpheme inventories into one triple store automatically creates a multilingual data graph with underlying consistent and shared semantics that allows to retrieve, as in Figure 3.3 for example, the German and English morphs encoding the meaning of 'plural' simultaneously.

Framework integrity: A crucial aspect for a semantic representation model that aims at cross-disciplinary usability is its capacity to enable the representation of morphological data irrespective of its underlying theoretical foundation. This aspect is addressed by the MMoOn Core ontology inasmuch as the fine-grained class structures and coverage of central domain concepts provide the necessary flexibility to enable the representation of different conceptions of the domain of morphological data. As mentioned at the beginning of this chapter, a structuralist view on morphology influenced the creation of the MMoOn Core ontology. In particular, it takes up the notion of the linguistic sign as defined by Ferdinand de Saussure, to be an inseparable unit of form and meaning, i.e. the signifier and the signified (De Saussure, 1989). Transferred to the ontology these concepts are manifested as the `mmoon:Morph` and `mmoon:Morpheme` classes as the perceivable sequence of phonemes or graphemes and its corresponding mental representation, respectively. However, it shall be stressed that the MMoOn Core ontology does not follow a prescriptivist approach. To the contrary, the provided ontological elements shall equally acknowledge and enable the representation of morphological data of different theoretical foundations without taking any preference or position regarding their appropriateness or validity. In the following, examples of selected theories illustrate the MMoOn Core ontology's range of representation possibilities.

Listing 3.1 shows an example of the Item and Arrangement model (Hockett, 1954; Bauer, 2004, p.60; Aronoff & Fudeman, 2011, pp. 46-52) that regards morphology as the concatenation of morphemes corresponding to morphs and accounts best for agglutinative languages and regular forms. As can be seen in the example the process of suffixation is applied to the verbal stem and results in a sequence of morphs. The morphs are regarded as lexical components following arrangement rules, i.e. $[\text{jag}]_{\text{JAG}}+[\text{t}]_{\text{PST}}+[\text{en}]_{\text{3P.PL}}$. It further illustrates the more advanced variant of this model called Distributed Morphology (Halle & Marantz, 1992). This theory takes up a more hierarchical approach similar to syntax but applied to the word level in that the mere linear ordering is separated into a verbal stem in the past tense and a person ending, i.e. yielding the structure $[[\text{jag}]_{\text{JAG}}+[\text{t}]_{\text{PST}}]_{\text{JAG-PST}}+[\text{en}]_{\text{3P.PL}}$. Further, Example 1 shows the representation of a fusional morph as an abstraction to account for phono-

logical elements expressing an inseparable set of morphosyntactic features, e.g. plural and person for the suffix *-en*. The concept of 'fused morphs' as well as that of 'zero morphs' are both required within this model to assure a morph-to-morpheme-correspondence. Both can be represented with the MMoOn Core ontology, even though the latter is not shown in the example.

```
@prefix mmoon: <http://mmoon.org/core/> .
@prefix deu_schema: <http://mmoon.org/lang/deu/schema/og/> .
@prefix : <http://mmoon.org/lang/deu/inventory/og/> .

:syntheticWordform_jagten a deu_schema:SyntheticWordform ;
  mmoon:belongsToLexeme :lexeme_jagen_v ;
  mmoon:consistsOfMorph :stem_jagt_v_PST , :suffix_en_1 .

:stem_jagt_v_PST  mmoon:hasRepresentation :Rep_jagt ;
    mmoon:consistsOfMorph :stem_jag_v_PRS , :suffix_t .

:stem_jag_v_PRS mmoon:hasRepresentation :Rep_jag ;
    mmoon:correspondsToMorpheme :atomicMorpheme_JAGEN ;
    :mmoon:hasMeaning :sense_jagen .

:suffix_t mmoon:hasRepresentation :Rep_t ;
    mmoon:correspondsToMorpheme :atomicMorpheme_PST ;
    :mmoon:hasMeaning deu_schema:past .

:suffix_en_1 mmoon:hasRepresentation :Rep_en ;
    mmoon:correspondsToMorpheme :fusionalMorpheme_3P_PL ;
    mmoon:hasMeaning deu_schema:thirdPerson , deu_schema:plural .

:Rep_jagt mmoon:orthographicRepresentation "jagt"@de .
:Rep_jag mmoon:orthographicRepresentation "jag"@de .
:Rep_t mmoon:orthographicRepresentation "t"@de .
:Rep_en mmoon:orthographicRepresentation "en"@de .
```

Listing 3.1: Example 1: Morphological representation according to the Item and Arrangement and Distributed Morphology models.

The second model of interest is the Item and Process Morphology (Hockett, 1954; Aronoff & Fudeman, 2011, pp. 46-52). Instead of a list of morphs it considers lexemes as items to which operations like affixation, vowel change or reduplication are applied to yield a word-form. Listing 3.2 illustrates three different English nouns, their plural word-forms and the contained allomorphs. In this model the word-form *drivers* results from the lexeme *driver* by the process of pluralisation and the operation of suffixation. The main focus lies on the process and the involved elements that create the word-form and not on the identification of the segment that is aligned to a meaning. Therefore, the concept of 'allomorphy' is applied to account for the different phonological variants of the plural suffix *-s* occurring within the three exemplary word-forms. This yields the process of pluralisation which is formalised by the rule $[\text{driver}]_{\text{noun stem}}$ –> $[\text{drivers}]_{\text{N[+PL]}}$. This rule specifies that plural nouns are formed by suffixation with [s]. In compliance with the respective phonological operation that applies to this suffix in each word-form one of the three allomorphs /z/, /s/ or /ɪz/ is realised. As can be seen in Example 2, the morphological elements, e.g. the word-from, morph and meaning resources can be represented. The formalisation of the pluralisation process is, however, not

explicitly stated, i.e. there is no "Process" class or the like provided in the MMoOn Core vocabulary. However, these processes can be deduced from the data and the dataset vocabulary can be extended in order to represent processes as required.

```
@prefix mmoon: <http://mmoon.org/core/> .
@prefix eng_schema: <http://mmoon.org/lang/eng/schema/og/> .
@prefix : <http://mmoon.org/lang/eng/inventory/og/> .

:Wordform_cats_n a eng_schema:Wordform ;
    mmoon:belongsToLexeme :lexeme_cat_n ;
    mmoon:consistsOfMorph :stem_cat_n , :suffix_s_1 .

:Wordform_drivers_n a eng_schema:Wordform ;
    mmoon:belongsToLexeme :lexeme_driver_n ;
    mmoon:consistsOfMorph :stem_driver_n , :suffix_s_2 .

:Wordform_houses_n a eng_schema:Wordform ;
    mmoon:belongsToLexeme :lexeme_house_n ;
    mmoon:consistsOfMorph :stem_house_n , :suffix_s_3 .

:suffix_s_1  mmoon:isAllomorphTo :suffix_s_2 , :suffix_s_3 .
    mmoon:hasMeaning eng_schema:plural ;
    mmoon:hasRepresentation :Rep_s_1 .

:Rep_s_1 mmoon:orthographicRepresentation "s"@de ;
    mmoon:phoneticRepresentation "[z]".

:suffix_s_2  mmoon:isAllomorphTo :suffix_s_2 , :suffix_s_3 .
    mmoon:hasMeaning eng_schema:plural ;
    mmoon:hasRepresentation :Rep_s_1 .

:Rep_s_2 mmoon:orthographicRepresentation "s"@de ;
    mmoon:phoneticRepresentation "[s]".

:suffix_s_3  mmoon:isAllomorphTo :suffix_s_1 , :suffix_s_2 .
    mmoon:hasMeaning eng_schema:plural ;
    mmoon:hasRepresentation :Rep_s_3 .

:Rep_s_3 mmoon:orthographicRepresentation "s"@de ;
    mmoon:phoneticRepresentation "[Iz]".
```

Listing 3.2: Example 2: Morphological representation according to Item and Process model.

The third approach that is addressed is called the Word and Paradigm model (Matthews, 1972) or word-based model (Aronoff, 1976; cf. Haspelmath & Sims, 2013, pp. 46-53). In contrast to the aforementioned models, here, the smallest morphological operating units are words which are not split up into morphs. Further, no declarations about concatenations are postulated. Instead, word-schemas are created that spell out realisation rules with features that are common to morphologically related words. Morphology, hence, is regarded as morphological correspondence between concrete words, i.e. word-forms. In this regard, word-forms are subsumed as words of a certain morphological feature that match an abstract schema, e.g. pluralised nouns. This theory requires the creation of subclasses for the `Lexeme` and `Wordform` schema ontology classes. Listing 3.3 illustrates the representation of the formulated word-schema $[X_N]$ <-> $[Xs]_N$ for the realisation of plural word-forms from singular nouns for an English MMoOn morpheme inventory. Since an ontology is usually meant to represent and

not generate data the subset of noun lexemes and their corresponding word-forms need to be already provided to instantiate the two classes accordingly. However, the lines 7 and 15 illustrate how the features expressed in such a morphological rule can be ontologically formalised by creating two subclasses. One specific lexeme class and one specific word-form class which contain all lexemes and word-form instances to which these features apply.

```
1  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
2  @prefix owl: <http://www.w3.org/2002/07/owl#> .
3  @prefix mmoon: <http://mmoon.org/core/> .
4  @prefix eng_schema: <http://mmoon.org/lang/eng/schema/og/> .
5  @prefix : <http://mmoon.org/lang/eng/inventory/og/> .
6
7  eng_schema:lexemeClass_nominal_pluralization rdfs:subClassOf
       eng_schema:Lexeme ;
8  [ rdf:type owl:Restriction ;
9   owl:onProperty :wordclassAffiliation ;
10  owl:hasValue eng_schema:noun] ;
11 [ rdf:type owl:Restriction ;
12  owl:onProperty :hasMeaning ;
13  owl:hasValue eng_schema:singular] .
14
15 eng_schema:wordformClass_nominal_pluralization rdfs:subClassOf
       eng_schema:Wordform ;
16 [ rdf:type owl:Restriction ;
17  owl:onProperty :wordclassAffiliation ;
18  owl:hasValue eng_schema:noun] ;
19 [ rdf:type owl:Restriction ;
20  owl:onProperty :hasMeaning ;
21  owl:hasValue eng_schema:plural] .
22
23 eng_schema:WordSchema rdfs:subClassOf
       eng_schema:MorphologicalRelationship .
24
25 :wordSchema_noun_pluralization a eng_schema:WordSchema .
26
27 :lexeme_cat_n a eng_schema:lexemeClass_nominal_pluralization ;
28     :hasMorphologicalRelationship :WordSchema_noun_pluralization ;
29      mmoon:hasWordform :wordform_cats_n .
30
31 :wordform_cats_n a eng_schema:wordformClass_nominal_pluralization ;
32      mmoon:hasMorphologicalRelationship :WordSchema_noun_pluralization .
```

Listing 3.3: Example 3: Morphological representation according to Word and Paradigm model.

In this example it is expressed that every instance of the class `eng_schema:lexemeClass_nominal_pluralization` is a noun with the grammatical meaning of singular, and that every instance of the class `eng_schema:wordformClass_nominal_pluralization` is a noun with the grammatical meaning of plural. From this it becomes clear that meanings which are usually assigned to morphs are encoded on the word level only. The two elements of the rule, i.e. the lexeme and its resulting word-form, are explicitly interrelated with `mmoon:hasWordform` and both are connected to the instance `:wordSchema_noun_pluralization` which can be regarded as an identifier for the rule. In accordance to this modelling no further statements regarding any kind of meaning are created for `:lexeme_cat_n` and `:wordform_cats_n`. Even though the elements

of morphs and morphemes do not have any significant status in the Word and Paradigm model, they emerge inherently from the creation of the rules and could be, therefore, also explicitly represented by their phonological or orthographical representations and their corresponding meanings. That way, a valuable contribution for other researchers who are interested in reusing the data would be created.

Another theoretical view that is representable with the MMoOn Core ontology is the split-morphology hypothesis (Perlmutter, 1988; Scalise, 1988). In this theory it is assumed that two grammatical subsystems exist, one operating pre-syntactically to form new lexemes and one post-syntactically providing the grammatical features to yield a word-form in accordance with its syntactic environment. These are known as the distinction between word-formation encompassing derivation and compounding on the one side, and inflection or word-form-formation on the other side. In Listing 3.4 the representation of these two systems is illustrated with the example of the German word-form *Skialpinistinnen*.

```
@prefix mmoon: <http://mmoon.org/core/> .
@prefix deu_schema: <http://mmoon.org/lang/deu/schema/og/> .
@prefix : <http://mmoon.org/lang/deu/inventory/og/> .

:syntheticWordform_Skialpinistinnen_n a deu_schema:SyntheticWordform ;
    mmoon:inflectionalRelation deu_schema:declension ;
    mmoon:belongsToLexeme :derivedWord_Skialpinistin_n ;
    mmoon:consistsOf :stem_Skialpinistin_n , :suffix_en_2 .

:derivedWord_Skialpinistin_n mmoon:isDerivedFrom
    :compoundLexeme_Skialpinist_n ;
    mmoon:derivationalRelation deu_schema:femaleNoun ;
    mmoon:consistsOf :stem_Skialpinist_n , :suffix_in_1 .

:compoundLexeme_Skialpinist_n mmoon:isComposedOf ; :simpleLexeme_Ski_n ,
    :derivedWord_Alpinist_n ;
    mmoon:compoundinglRelation deu_schema:nominalCompound .

:derivedWord_Alpinist_n mmoon:isDerivedFrom :derivedleLexeme_alpin_adj;
    mmoon:derivationalRelation deu_schema:deadjectivalNoun ;
    mmoon:consistsOf :stem_alpin_adj , :suffix_ist .

:derivedleLexeme_alpin_adj mmoon:isDerivedFrom :simpleLexeme_Alpen_npr ;
    mmoon:derivationalRelation deu_schema:denominalAdjective ;
    mmoon:consistsOf :stem_Alpen_n , :suffix_in_2 .

:suffix_en_2 mmoon:inflectionalMeaning deu_schema:nominative ,
    deu_schema:plural , deu_schema:feminine .

:suffix_in_1 mmoon:derivationalMeaning deu_schema:femaleNominalization .

:suffix_ist mmoon:derivationalMeaning deu_schema:personNominalizer .

:suffix_in_2 mmoon:derivationalMeaning deu:schema:relational .
```

Listing 3.4: Example 4: Morphological representation accounting for the split-morphology hypothesis.

As can be seen, the vocabulary allows to express not only the derivational and inflectional meanings of the respective morph resources but also to state the explicit derivational and inflectional relations that apply. The morphosyntactic feature values are assigned to the morph that yields the

word-form. Moreover, all word-formation processes are analysed on the lexeme level, thus, interconnecting all lexemes that are involved in the formation of *Skilapinistin* which could be further specified for their lexical meanings. MMoOn Core does not only provide a wide range of morphosyntactic features but also supplies a set of derivational meanings which are ready-to-use and can be extended accordingly.

Finally, the semantic representation provided by the MMoOn Core ontology accounts for the actual morphological language data available in addition to the theoretical frameworks just outlined. In particular this encompasses datasets which do not contain explicit resources for morphemes, morphemic glosses, the interrelations between the morphemic and word elements or distinguish between inflectional and derivational morphs. This applies, for example, to the output of morphological analyser tools used in NLP that are based on the so-called two-level morphology (Karttunen & Beesley, 2001; Koskenniemi, 1983). These systems take the surface representation of each token, i.e. the exact spelling of a word-form in a text corpus, as an input, apply specific rules to it and output a morph representation of these forms in a linear way. In Listing 3.5 the word-form *Skialpinistinnen* from Example 4 is taken up again to illustrate the different representations.

```
@prefix mmoon: <http://mmoon.org/core/> .
@prefix deu_schema: <http://mmoon.org/lang/deu/schema/og/> .
@prefix : <http://mmoon.org/lang/deu/inventory/og/> .

:wordform_Skialpinistinnen_n a deu_schema:wordform :
    mmoon:consistsOf :morph_ski , :morph_alp , :morph_in , :morph_ist ,
    :morph_inn , :morph_en .
```

Listing 3.5: Example 5: Morphological representation accounting for two-level morphology.

Usually, the segmented output elements are further specified for linguistic information, such as part of speech and morpho-syntactic feature values. These are, however, omitted in Example 5 since their representation has been shown already in the examples before. Noteworthy is that all further lexical elements involved, from the surface representation to the segmented morphs, are not explicitly provided and, therefore, lost in the output. Moreover, the vocabulary is used on a very general level and, thus, does not further specify for a synthetic word-form, stems or affixes. Also the analyser outputs the grapheme sequence <inn> for the suffix *-in* which is an orthographic variation exhibiting gemination within the given word-form. Still, it is represented as a morph resource because orthographic data can be only represented as strings which cannot appear as instance data autonomously. Nonetheless, this presents valuable morphological data because it serves as attestation for the orthographic realisation of morphs which is also part of the domain of morphology. Furthermore, a more detailed description of two-level morphology data can be realised by extending and specifying the data illustrated in Example 5 with more resources and statements from within the ontology.

To conclude, [P2] introduced the MMoOn Core ontology as the proposed foundation for semantically represented and interoperable morphological language data. The three main outcomes outlined above emphasise its adequacy for a consistent documentation of linguistic and morphological data for single languages. It could be also explained how semantic interrelatedness is established within the architectural setup of MMoOn morpheme inventories to ensure cross-linguistic comparability as well as an inherent multilingual data graph by uniting multiple morpheme inventories. Moreover, it could be demonstrated how the MMoOn Core ontology reaches a large coverage in representing morphological language data by taking four central theoretical frameworks of morphology with a varying degree of descriptive granularity as well as segmentation tool outputs into account. As a result, the semantic representation model offered with the MMoOn Core ontology can be regarded as a suitable foundation for realising cross-disciplinary morphological data usage.

Publication 3: Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models

Before the MMoOn Core ontology emerged, the OntoLex-*lemon* model was the only existing means to describe morphological data as RDF resources. One part of [P3] describes the limitations of this model and simultaneously motivates MMoOn Core as the ontology overcoming these limitations and providing a more accurate domain representation. The other part of [P3] discusses a possible alignment of both models, thus, interconnecting OntoLex-*lemon* as domain ontology for lexical data with MMoOn Core as domain ontology for morphological data. This proposal is motivated by two circumstances: 1) Linked Data best practices suggest to reuse already existing vocabulary whenever possible, and 2) many language datasets contain lexical as well as morphological data. A default alignment of both models facilitates the transformation of such datasets into RDF and standardises the interoperability between datasets using both models. The following three main outcomes are summarised for [P3]:

Model comparison: A large part of this publication presents the comparison of the expressivity of the ontolex and decomp submodules of OntoLex-*lemon* and the MMoOn Core ontology. On the basis of thirteen representation examples the different modelling capabilities are illustrated according to the lexeme, word-form, morph and morpheme levels of the morphology domain. As it is common for linguistic domains to have no clear domain delimitation the results of this comparison revealed an overlap of both models, mainly with regard to the word level. The ontolex module provides the classes `ontolex:LexicalEntry` and `ontolex:Form` similar to `mmoon:Lexeme` and `mmoon:Wordform`. However, this seeming similarity

also revealed how carefully ontological elements have to be studied in their practical application, since both models obviously declare different underlying semantics for them. Likewise, the two classes `ontolex:Affix` and `mmoon:Affix` are treated differently and cannot be used interchangeably.

OntoLex-*lemon* limitations: Next to the similarities, the comparison also exposed the divergence between the two models and, thus, the limitations of the OntoLex-*lemon* model in serving as an appropriate morphology domain vocabulary. It was shown that a semantic description of morphemes is not possible at all which is also due to the difficulty of representing the meaning side of linguistic signs in general. Therefore, all lexical senses can be represented with the ontolex module and all grammatical meanings are intended to be reused from the LexInfo vocabulary. On the morph level only the `ontolex:Affix` class exists, however, without any further granularity (for prefixes or suffixes for example) or useful properties for representing segmentations. Moreover, `ontolex:Affix` is conceptualised as a subclass of `ontolex:LexicalEntry` restricting the usage of this class to describing morphs which bear some lexical, i.e. derivational, but no inflectional, meaning. The provided vocabulary on the word level revealed more possibilities for explicitly stating sub-word elements. Anyhow, the object properties provided are clearly advocating the segmentation of compound words. A morphological segmentation according to a linguistic theory as described in the summary of [P2] above is not intended. Instead, a more general view of morphological segmentation as the decomposition of compound words (hence, the name "decomp" module) is realised with the goal to mainly identify other lexical items as sub-terms of an `ontolex:LexicalEntry` resource. An analogous decomposition of word-forms, i.e. `ontolex:Form` resources, is neither envisioned nor possible due to the semantic restriction of the `ontolex:Affix` class described above. Furthermore, there is also the promising object property `ontolex:morphologicalPattern`. However, this is more of a relict from an earlier model version than an actually integrated part of the modules, since its usage is not clearly defined and this property, therefore, not used in existing datasets. It now remains as a "dead" element in the vocabulary.

The in-depth analysis of both models did not only point out the limitations of the OntoLex-*lemon* model for representing morphological language data but also simultaneously illustrated the applicability of the MMoOn Core ontology to solve these representation shortcomings. As a result, the comparison confirmed OntoLex-*lemon* as suitable domain model for lexical data and validates MMoOn Core as the more appropriate modelling alternative for describing the morphological data domain. Consequently, it is obvious to take up the identified conceptual overlap to create an ontology alignment that provides an efficient usage of both models in parallel.

Alignment proposal: Accordingly, four concrete proposals for interconnecting both models are given in [P3]. Even though, such an alignment would reduce the need to create custom vocabularies and enhance a seamless extension of lexical data with morphological resources, it has not been further pursued from the side of the OntoLex-*lemon* community. Therefore, the MMoOn Core ontology has been adjusted onesidedly to provide at least the creators of MMoOn morpheme inventories with a possibility to easily integrate their OntoLex-*lemon*-based lexical data. Thus, the statement declaring `mmoon:LexicalEntry` to be a subclass of `ontolex:LexicalEntry` as well as the object property `mmoon:senseLink` have been added to MMoOn Core as an outcome of [P3]. Consequently, existing `ontolex:LexicalEntry` resources can be further described for their morphological information by using the MMoOn Core vocabulary. Their respective `ontolex:LexicalSense` resources, in turn, can be reused for describing the senses of `mmoon:Stem` instances.

Overall, [P3] can be considered as an evaluation of the state of research for the semantic modelling of morphological data provided by the two respective ontolex and decomp modules of the OntoLex-*lemon* model at the time of publication. It introduced the MMoOn Core ontology to the LLOD community and raised awareness for the lack of semantic representation means for the domain of morphology and for the presented MMoOn Core ontology as an adequate solution to fill this gap in alignment with the existing OntoLex-*lemon* model.

Publication 4: Challenges for the Representation of Morphology in Ontology Lexicons

The response of the Ontology-Lexicon community group to the publication of [P3] resulted in the decision against the reuse of the MMoOn Core ontology by aligning it with the decomp and ontolex modules of the OntoLex-*lemon* model. This is motivated by two main circumstances. First, OntoLex-*lemon* is used as a defacto standard by now and numerous datasets are created based on it already. Researchers that are familiar with the model shall have the possibility to easily extend their datasets with morphological data by staying within the modular ontology framework they are already used to. Second, the MMoOn Core ontology, even though well received by the community members, is considered too fine-grained and too complex in contrast to the underlying modelling principles of OntoLex-*lemon*[9]. Consequently, the development of the Morphology

[9]The main difference being referred to here is defined in the ontology specification which states: "The lexicon model for ontologies is a model for describing lexical resources in connection to ontologies, it is not a generic vocabulary supporting the publication of any sort of linguistic data"(cf. https://www.w3.org/2016/05/ontolex/). In contrast, MMoOn Core is explicitly considered to serve as a representation basis for the

Module, in short "morph", has been initiated in order to overcome the shortcomings in representing morphological data identified in [P3]. [P4] presents the interim results of the module creation effort. The two main outcomes of this publication are summarised as follows:

The Morphology Module extension: For the first defined module goal to enable the representation of the decomposition of `ontolex:LexicalEntry` and `ontolex:Form` class instances the newly created ontological elements are introduced and explained. These encompass classes and properties that enable the declaration of various sub-types of morph resources, their specification for being inflectional or derivational morphs, and the assignment of word-forms to inflectional paradigms and derived lexemes to a derivational relation, respectively. Furthermore, an explicit interconnection between `ontolex:Form` resources and `ontolex:LexicalEntry` resources as well as the `morph:Morph` resources of which both consist is also realised. As a result, the Morphology Module is embedded within the existing OntoLex-*lemon* modules and finally extends the already provided decomposition of compound words with the necessary vocabulary to represent inflection and derivation. With regard to the formerly existent limitations mentioned in [P3] it has to be noted that morphemes are still not representable. This is due to the general principle underlying OntoLex-*lemon* to use the LexInfo ontology for the representation of grammatical meanings. Moreover, from the new ontology elements presented in [P4] it is obvious that a considerable adoption of elements that already exist within the MMoOn Core vocabulary has been applied. The object property `morph:consistsOf` and all the subclasses of `morph:Morph` for describing stem, root, transfix, simulfix and zero morph resources similarly exist in MMoOn Core and have been regarded as valuable inclusion into the Morphology Module as well. Even though, these elements are only defined within the realm of OntoLex-*lemon* and not aligned with the MMoOn Core ontology, their transfer can be considered as an indirect validation and acknowledgement of the semantic representation of morphological data within MMoOn Core.

The second goal of the Morphology Module to enable an automatic generation of `ontolex:Form` instances for `ontolex:LexicalEntry` resources is still pursued and not implemented to date. Nonetheless, the novelty of this endeavour should be pointed out. It takes up the Item and Process theory of morphology by creating replacement rules for the affected grapheme strings and converting them into regular expressions. These are used to generate all word-forms but also their contained morph resources if they are provided in the source data. These operations will be integrated into the model and, thus, go beyond a mere ontology development. Further, this will compensate for the limitation of MMoOn Core which is designed to represent but not generate morphological data. Still, the applicability

publication of morphological data and, hence, accordingly more complex.

of this approach remains to be seen.

Modelling challenges: A valuable contribution of [P4] lies in the exposure of the modelling choices that eventually lead to the resulting Morphology Module. Due to the ongoing development of OntoLex-*lemon* module extensions but also the growth in community group members and model users, the module development raised new challenges. Those needed to be explicated (also in prospect of future modules) to ensure a consistent model application. Therefore, the difficulties of defining the scope and coverage of the Morphology Module, the consistency barrier that required an adaption of existing vocabulary elements as well as the sensitisation for the chosen terminology of the new model classes and properties have been pointed out. Concomitantly, the choice for the dedicated Morphology Module to be developed in addition to the decomp module demands to ensure a twofold coherence. It has to balance between the integrity within the whole OntoLex-*lemon* model, on the one hand, and all users coming from different disciplinary backgrounds producing and using morphological data in various ways, on the other hand. While the approach to this, as described in Section 5 in [P4], implemented clear solutions, however, these modelling choices might come at the price of usability, e.g. in the confusion of the `ontolex:Affix`, `morph:Morph` and `morph:AffixMorph` classes.

As a result of [P4], the emerging OntoLex-*lemon* Morphology Module extension will be another semantic model for representing morphological language data next to the MMoOn Core ontology. Still a standardised alignment of both is not envisaged so far by the OntoLex community. Nonetheless, the Linked Data framework provides the possibility to extend MMoOn-based or OntoLex-*lemon*-based datasets with a customary interconnection to each other if desired. Based on the outlined scope, functionalities and usage potential of both models, data creators can choose which model most adequately fulfills their representation and application needs. In general it is likely that the Morphology Module will be preferred by computer linguists and users with experience in working with Linked Data. To the contrary, for linguists the formal semantic implications which resulted from the necessary OntoLex-*lemon* adaption are less obvious and might cause an incoherent vocabulary usage. The MMoOn Core ontology, however, is not only the more encompassing domain model for morphological data but also provides a consistent vocabulary linguists, but also computer linguists, are very familiar with.

To summarise, the presented outcomes of the publications [P2], [P3] and [P4] have laid the foundation for an actual cross-disciplinary usage of morphological language data. Due to the semantic data representation that is available with the MMoOn Core ontology semantically and structurally interoperable datasets can be created and reused. The illustrated use cases

in Section 7 of [P2] provide a prospect of how the suggested data reuse cycle introduced in Chapter 3.1.2 could be induced. Therefore, the MMoOn Core ontology will contribute to an extension and exchange of morphological language data that was not realisable so far and is beneficial for every research area that relies on this data. Moreover, it has been shown that the MMoOn Core ontology qualifies as an adequate domain ontology to document morphological data resources that have been formerly described either in the context of grammars or within lexical datasets in a unifying and homogeneous setup.

Simultaneously, all three publications constitute a chronological overview illustrating the development of semantic models for representing morphological data within the past five years. The identified insufficiency of the ontolex and decomp modules ([P3]) caused the development of the MMoOn Core ontology ([P2]) which in turn initiated the creation of the Morphology Module extension ([P4]). As a consequence, next to MMoOn Core a second model will exist. This is regarded as a positive effect because an additional increase in the publication of more morphological language data can be expected due to the automatic data generation feature of the Morphology Module. Ultimately, the MMoOn Core ontology still persists as comprehensive domain model with the complementary focus on more detailed and manually compiled language data which was not representable before.

3.2.3 Publication 5 and 6

Finally, the last two publications [P5] and [P6] present two MMoOn morpheme inventories, i.e. the Hebrew Morpheme Inventory and the Xhosa RDF dataset, respectively. Both are conversions of manually compiled tabular lexical and morphological data which were not applicable for their intended usage in this data structure. Thus, the outcomes of [P5] and [P6], as outlined in the following, fulfill the last requirement for inducing the cross-disciplinary usage of morphological language data in that they demonstrate the added value that is obtained by their transformation to the RDF format.

Publication 5: Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory

The source data of the Hebrew Morpheme Inventory constituted a large and manually not further manageable table based on custom linguistic concepts. It entailed errors, inconsistencies in its editing and repetitions which naturally occur in manual data compilations of this size. At this point, the data creator expressed his need for a more consistent and accurate data structure that would also be able to represent the language-specific peculiarities of Hebrew. Especially the non-concatenative morphological

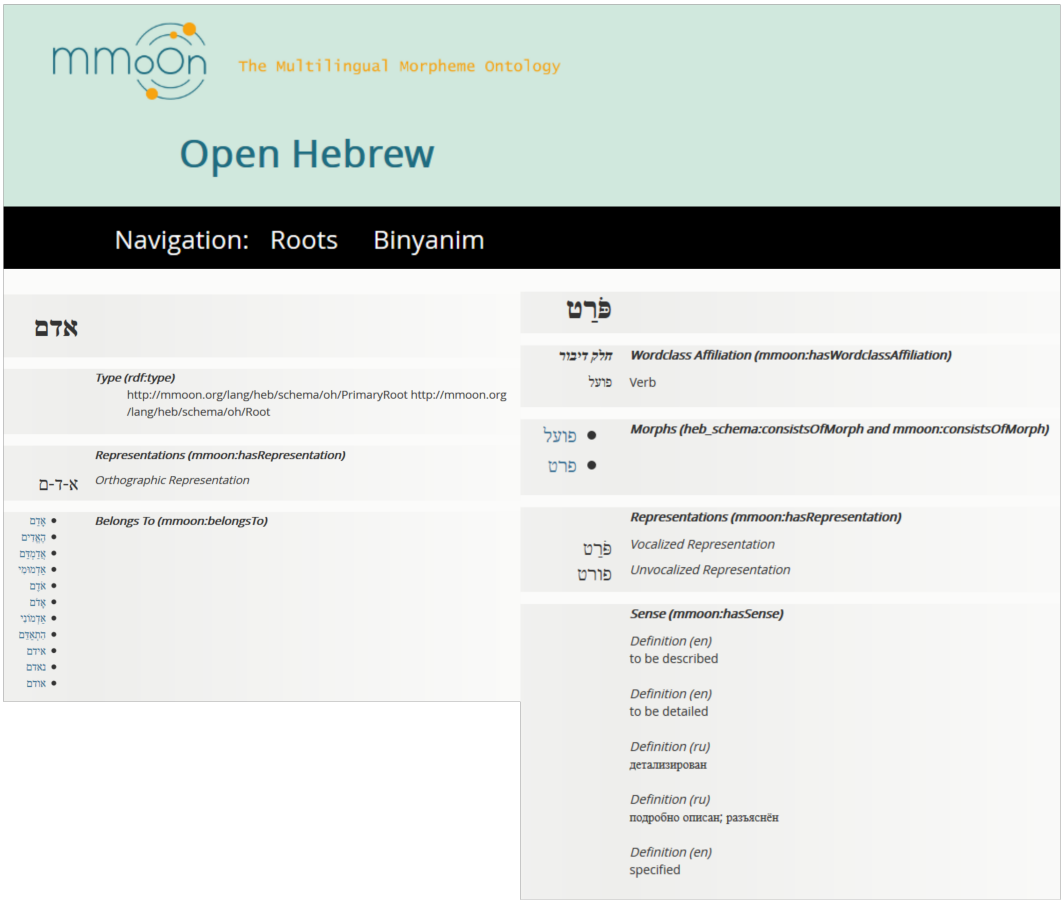
structure of lexemes and word-forms, consisting of consonantal roots to which vowel patterns are applied, should be explicitly stated in a machine-processable manner. For these goals the MMoOn Core ontology was considered as suitable data model and, thus, applied. Simultaneously, this dataset creation could serve as the proof of concept for representing morphological data as MMoOn morpheme inventories in accordance with the proposed architectural setup. As a result, the necessary schema ontology has been created adhering to the language expert's specifications of the Hebrew language.

The main outcome of [P5], therefore, is the created Hebrew Morpheme Inventory with the specific dataset details presented in Section 6 of the publication. It could be proved that the MMoOn Core ontology enabled a full semantic representation of the source data adhering to the special morphological characteristics of the Hebrew language. This includes in particular the reduction of repeated morph entries in the source table to unique resources, their interconnection with all language elements to which they are interrelated, the explication of the underlying linguistic concepts within this particular dataset and the explicit representation of morphologically complex elements. Moreover, the underlying RDF format permitted an external data enrichment by interlinking the Hebrew data with sense and lexical entry resources of the Babelnet[10] dataset.

Given that the dataset is available as Linked Data, a Web interface[11] (Arndt et al., 2019) has been created that allows to browse the data in a human-readable way. Figure 3.4 shows two screenshots displaying a root and a lexeme resource. As can be seen, the MMoOn Core vocabulary has been reused to express the relations between the resources and, since all instances are URIs, they can be used as links to navigate through the dataset. For the root resources all lexemes in which they occur are directly listed and the lexemes are provided with the morphs they contain, i.e. its root and Binyanim (which defines the vowel pattern information). Through this setup a new kind of lexical and morphological language resource for Hebrew has been created that overcomes the shortcomings of traditional print dictionaries.

To conclude, the resulting Hebrew Morpheme Inventory has proven that the MMoOn Core ontology together with its architectural setup is able to represent morphological language data in the way it was intended. Therefore, this dataset is also the first instance of fine-grained morphological data represented with the RDF format, given that the existing ontological model at the time, i.e. *lemon*[12] (the predecessor of OntoLex-*lemon*), did not have the expressivity to transform the same source data without a significant data loss. The Hebrew Morpheme Inventory further enables more accurate data editing, cleaning and verification methods and can be

[10] http://babelnet.org
[11] https://mmoon-project.github.io/JekyllPage/
[12] http://lemon-model.net



https://mmoon-project.github.io/JekyllPage/inventory/oh/Root_פרט https://mmoon-project.github.io/JekyllPage/inventory/oh/Lexeme_אדם

Figure 3.4: Screenshots of a root and lexeme resource in the Hebrew Morpheme Inventory Web interface.

extended with additional Hebrew language data existing in RDF.

Publication 6: Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

The Xhosa RDF dataset presents another conversion from manually compiled tabular language data that is based on the MMoOn Core ontology. This morpheme inventory is placed within the challenging setting of under-resourced languages which rely on collaborative approaches between the language's experts, computer scientists and environments providing the infrastructures for data distribution, exchange and documentation. A central role in this plays the feature of data interoperability which is achieved by the Linked Data framework and especially valuable in the course of providing more resources for under-documented languages like isiXhosa (Eckart et al., 2018). Therefore, the SADiLaR national centre opted for exploring RDF and Linked Data in order to "support research and development in

the domains of language technologies and language-related studies in the humanities and social sciences" (cf. https://www.sadilar.org/).

Within this context a semantic representation model was required to convert the isiXhosa source data described in Section 2 of [P6] into RDF. Again, the main disadvantages of the tabular data structure lied in the repetition of language elements, in this case the noun class affixes, the missing interrelations between the stem, affix and translation instances and the absence of the lexeme resources that are formed based on the provided stems and affixes. The aim of the data conversion was not only to merely provide the isiXhosa source data in the RDF format. Beyond that, it should demonstrate the suitability of the created ontological foundation to account for the language-specific features of the whole Bantu language family as well as both linguistic domains of lexical and morphological data.

Therefore, next to the Xhosa RDF dataset, the developed Bantu Language Model (BLM) constitutes the main outcome of [P6]. Due to its vocabulary, the entire source data could be mapped to the BLM and, moreover, the resulting Xhosa RDF dataset could be extended by three kinds of data which were not contained in the original source data before. With regard to the initial morphological data, first, the given formal and semantic ambiguity of the affixes could be resolved due to the possibility provided in the BLM to explicitly state allomorphy and homonymy relations between any unique prefix or suffix resources that share the same meaning or the same overt form, respectively. Before, every unique morph had to be repeated within the table for every row without any specification of their semantic interrelations. Second, new instance data could be generated by combining the given root and affix resources to form lexemes, thus, producing `blm:LexicalEntry` resources. For these the BLM vocabulary further allowed to establish a meaningful relation to the `blm:Morph` resources, i.e. the roots and affixes of which they consist. Lexical entry resources are the basis for every lexical dataset and enabled the assignment of the translations accordingly. The third kind of additional data constitute external lexical entry resources obtained from WordNet RDF[13]. Because the mere strings within the translation columns of the source data have been turned into `ontolex:LexicalEntry` resources within the Xhosa RDF dataset, links to WordNet RDF resources could be established that provide more detailed lexical information for the translations. Therefore, the original translations are automatically enriched with sense and word-class information by this interlinking. Consequently, showcased for the Xhosa RDF dataset, the BLM proved to be applicable in the intended way. Interoperability between future SADiLaR language resources is provided by the BLM that also guarantees the intrinsic semantic interrelatedness of future Bantu language RDF datasets.

In fact, further language data next to the Xhosa RDF dataset emerged already. The BLM has been reused to convert existing lexical data of the

[13]http://wordnet-rdf.princeton.edu/

Kalanga and Ndebele Bantu languages which have been also interlinked with lexical entries of the English WordNet RDF. Together with Xhosa RDF, lexical data for these three Bantu is now available. The English WordNet RDF has then been used as the pivot language dataset that enabled the identification and creation of new translations between lexical entries among all three languages (Eckart et al., 2019). As a result, the BLM contributed to the creation of new language resources for Bantu in RDF, i.e. multilingual translation data in this case.

The reuse of existing language data to conduct this kind of dictionary alignment as well as the mentioned three types of data that could be created additionally as an enrichment for the Xhosa RDF dataset are especially important for the emergence of language material for the Bantu languages. The usage of an ontology such as the BLM to create semantically modelled language data in RDF proved to be more than adequate for the representation of the commonly very fragmented data available for under-resourced languages. It does not only enable the transformation of the original data without any data loss but, moreover, allows for an internal as well as external dataset enrichment of which the latter is hitherto not achievable with any other data format. Moreover, the ontological modelling of the BLM entails cross-disciplinary potential which goes beyond the intended usage of the Bantu language resources of SADiLaR. Given that the BLM is openly available it can be directly reused for creating Bantu language datasets of any other origin into RDF. It prevents the formation of new data silos in so far that it inherits an alignment to the MMoOn Core ontology as well as to the OntoLex-*l*emon model. As a result, all emerging datasets based on the BLM will be automatically interoperable with existing datasets which are based on one or both of the other two vocabularies. Reversely, already existing lexical datasets of Bantu languages can be now extended with more fine-grained morphological data that was not adequately representable before. Consequently, the BLM is not only a specialised ontology limited to a single dataset but actually bridges the gap between the need for a manageable vocabulary dedicated to the language peculiarities of the Bantu languages and its integration into the existing landscape of ontologies for representing lexical and morphological language data that are already in use.

From the presented outcomes of [P5] and [P6] in can be concluded that the achieved results are fully attributable to the underlying RDF format and in particular to the MMoOn Core ontology. Thus, the content-wise and format-wise homogeneity that is required for creating and reusing morphological language resources across different disciplines has been realised and put into practice. Therefore, with regard to cross-disciplinary usability, both datasets pose additional value to other research focusing on the Hebrew or the Bantu languages. Since the data is now machine-processable, text mining and NLP tools that rely on such high quality and fine-grained language data can use these datasets to improve results. Additionally,

theories in computational linguistics that could not be researched in depth due to the lack of such data can be now explored and applied.

3.3 Impact on Further Research

3.3.1 Implications

The outcomes of the six publications presented in this thesis are embedded into the context of Linked and its impact on linguistic and language data representation and reuse. In particular they are concerned with morphological data as one of many linguistic data domains and semantic data modelling as one among various existing digital means to create and represent data. This specific combination is motivated by the goal to advance language data-driven research by overcoming data barriers and discipline boundaries. Although the achieved results of the publications certainly induce a cross-disciplinary usage of morphological data through the proposed semantic model of the MMoOn Core ontology, it is obvious that the conducted work contributes only the beginning of this investigation. New issues and implications consequently emerged from the outcomes which significantly impact further research.

The presented datasets of [P5] and [P6] constitute a new type of linguistic data artefact within the science of linguistics. The MMoOn Core ontology enabled the creation of a stand-alone domain representation for morphological data as data sources which unify linguistic information that is usually distributed across dictionaries and grammars. It is due to the underlying semantic modelling that a novel conceptualisation of this linguistic domain could be realised which goes beyond the possibilities of their printed counterparts. Therefore, the MMoOn Core ontology as well as the resulting datasets from this model mark a **paradigm shift in the digitisation of language data**. The mere replication of manually compiled source data into a digital medium has turned into the constructive transformation of language data into unique data resources that are interconnected within a semantic network. The crucial change lies within the data representation model which has become an integral and inseparable part of the data it describes. This opens new possibilities to represent a data domain such as morphological data which could not be taken into account before. The MMoOn Core ontology and the OntoLex-*lemon* Morphology Module represent two conceptualisations of the morphology domain that illustrate how the reasoning capabilities of OWL can be utilised. In particular this implies that the dataset is explicitly shaped by the features of the modelling. The MMoOn Core ontology, for instance, makes use of symmetric and inverse properties and, moreover, established interconnections of glosses with meanings. Simultaneously, every dataset based on this ontology can be extended and customised with more specific ontological elements and features. As a consequence, the same source data can result

in different RDF datasets depending on the vocabulary that was chosen to describe it. For the research field of linguistics morphological datasets represented as Linked Data in this constructive rather than reduplicating approach to digitising data is comparably new. In order to understand and create MMoOn morpheme inventories and other Linked Data language resources linguists need to acquire a certain degree of data literacy which is considerably time consuming in the case of Linked Data. However, the technological development is constantly advancing and Linked Data an acknowledged method for digitising research data, not only in the computer sciences but increasingly also in the humanities. According to this, further research and studies in linguistic data management are required to investigate and evaluate the possibilities that arise from applying Linked Data in linguistics.

The semantic modelling approach to creating Linked Data with ontologies has been adopted for the linguistic domain of morphology because it enables cross-disciplinary data interoperability. From the works in this thesis it becomes obvious that an ontology, such as MMoOn Core, never exists in isolation. Even without any explicitly stated interconnections to other vocabularies or specific extensions of ontology elements within the schema files of MMoOn morpheme inventories, it is always embedded within a broader semantic network of other ontologies, at least with the RDF and OWL models. Therefore, every MMoOn-based dataset will be an integrated part of the global unified data graph of the LLOD and LOD clouds (provided that it contains the required number of links to other datasets). Being an integrated part of a larger data cloud is highly conducive for morphological data reuse across disciplines. Moreover, the use of **Linked Data intrinsically promotes the dissolving of discipline boundaries** since it allows the representation of knowledge domains in a discipline-opening way. For the domain of morphology this enabled the joint representation of data and knowledge that was distributed over grammars and dictionaries before. Within the broader semantic network of RDF-based language data it also permits to interconnect it with data from other linguistic domains, such as lexicographic and phonological data described with other vocabularies, e.g. OntoLex-*lemon* or PHOIBLE[14]. Beyond that, a direct enrichment with interlinkings to other knowledge resources provided in RDF is possible. General information about a language, such as details on the country, culture and speakers can be interconnected in addition to the actual language data. Reversely, this fusion of data across different knowledge domains impacts other disciplines as well. Every knowledge area that is subject to the field of content mining is enhanced by the interconnected language data, since information is to a large amount still encoded in natural language sources and can be extracted with the aid of language data. While this technical interconnection of data leads to a shared knowledge base for all research areas it simulta-

[14]http://phoible.org/

neously impacts the way single disciplines gain insights. New issues arise that challenge the established methods of scientific data compilation and management. Publishing a MMoOn morpheme inventory and integrating it into the LLOD cloud has several implications. For example, it means that meta-knowledge that determines the adequacy and usefulness of this dataset is harder to identify. Further research has to investigate on the fact that external data is always more or less interconnected to data created by another person or institution. To this extent, aspects of data provenance and quality in particular need to be reevaluated for language data. For the presented morphological data in this thesis this includes the creation of minimal standards for data reuse which take Linked Data-inherent aspects like the relation between the source data and generated data, expressivity of the underlying ontologies, ontology alignment, resource mappings, interlinking and provenance of all interconnected data sources into account.

Another implication from the works presented in this thesis evolves around the goal of knowledge acquisition gain obtained by using semantically modelled morphological data. For the two morpheme inventories that have been created with MMoOn it became obvious that the generation of new data in addition to the source datasets, as well as the enrichment by interlinking it to external resources, as such are regarded as knowledge acquisition gain. Since an ontology is in itself a meaning creating construct an inherent equivalence between data and knowledge emerges. No datum exists in isolation if it is described with RDF. It is at least a part of the minimal unit of a triple which again is at least a type assertion that interrelates it to an ontological element. The knowledge entailed in a dataset is accessible via the in-built reasoner of OWL that runs over the data. Consequently, resources that are not directly interrelated within a dataset can be still accessed by traversing the whole data graph. For the improvement of NER by using language resources represented as Linked Data, as proposed in [P1], this resource integrating graph structure is very powerful. Named entities in a text corpus can be identified in the data by traversing the graph form *Skialpinisten* to the proper noun *Alpen*. Even though this requires only one query the underlying assumptions that lead to an interconnection of both resources have been explicitly and formally constructed into the semantic modelling based on the domain knowledge of the ontology creator. As a consequence, data and the assumed knowledge about the data are inseparably merged into one dataset. This leads to a **need for the sensitisation of the interrelation of knowledge and data with regard to scientificity**. The knowledge acquisition gain which was originally derived from data by human researchers is largely conducted by machines for RDF-based data. For morphological data in particular this included a separation of data representation, which resulted in dictionaries, as well as the description of its underlying theoretical foundation and the scientific analysis of the data, which resulted in grammars. The merging of both into one morpheme inventory requires now an additional under-

standing of the effects on data representation with RDF on the knowledge it represents as well as the knowledge that is obtained from the automated computational processing of the reasoner. The aspect of interlinking data of separate datasets further entails that data is not static anymore but constantly evolving. Therefore, it is also prone to a conflation of data with formally underlying mismatching or mutually excluding assertions. Further research within single disciplines but also across different research fields is required that elaborates on the status, treatment, integration and evaluation of machine-generated knowledge and its consequences on data analysis for reaching scientific insights.

Finally, the usage of semantic modelling for representing morphological data entails consequences regarding other scientific data principles that are arising from the data format. In particular the basic shape of a single datum in its representation as a URI significantly enhances the aspects of content, interoperability, reusability, discovery and citation because it makes use of the single medium every researcher has access to: the World Wide Web. The publications in this thesis already illustrated the benefits for data content and interoperability in a cross-disciplinary setting. Simultaneously, however, this dependency on the URI structure of the format is also its greatest drawback. Even though data citation and discovery are reduced to opening a link in a browser, URIs are not as persistent as printed books. They have to be maintained and provided by the data creators. This might compromise a whole dataset and diminish its reusability since it consists of URIs only. The Hebrew Morpheme Inventory and the Xhosa RDF datasets do not only depend on the MMoOn Core ontology that is hosted by the author but also on the separately managed persistence of the OntoLex-*lemon*, WordNet RDF, OWL and RDF specification URIs. While the advantages of using URIs for data representation are compelling in theory, the Linked Open Data paradigm simultaneously generates novel obstacles that challenge its potential to **realise the scientific data principles in practice**. The conflation of data and the knowledge derived from this data makes it possible to reference every kind of data or knowledge to every level of granularity via URIs. This design enhances data processing for machines but is not in line with the methodologies and norms researchers have implemented for scientific data reuse. Since a URI can represent literally anything it is impossible for a human to recognise whether it refers to a single morphological element of a language, the grammatical terminology used to categorise it, a whole metadata entry about the dataset, the dataset with or without any external enrichment links or any other customised data fragment that is under investigation. The status of these resources for science is not clearly defined. It remains open at this point if a MMoOn schema ontology file can be regarded equivalent to a printed grammar or if it evolves to an assisting tool for grammar publishing. The created morphemic datasets in this thesis imply a reevaluation of this kind of data representation for human data usage. The semantic

modelling with URIs needs to be investigated with regard to the technical possibilities that realise the scientific data principles for machine processing on the one side and its compliance to existing scientifically established standards of sustainability and transparency on the other side.

These four major implications that result from the conducted work in this thesis emphasise the impact of semantic data representation for the domain of morphological but also language data in general. In contrast to established non-interoperable data representation structures the usage of Linked Data ignites a new scientific discourse about data management in a digitised scientific world in the future. The MMoOn Core ontology and the two resulting datasets illustrate how morphological data that is usually scattered across two linguistic domains can be unified and reused across disciplines, however, with effortful and not yet widely practiced means. Therefore, it will be inevitable to examine the aspects just mentioned in order to integrate and establish Linked Data resources within the scientific data practice based on shared cross-disciplinary conventions.

3.3.2 Limitations

The semantic modelling approach towards a cross-disciplinary reuse of morphological language data by creating ontology-based datasets in the RDF format contains accompanying limitations that need to be pointed out. These arise from the complexity of the application of the Linked Data framework and ontologies in general. In their adoption of language data they particularly entail compromises in exchange for the gained cross-disciplinary usage they enable. Therefore, the most significant consequences that emerge from creating and using morphological data as proposed within this thesis are outlined in what follows.

Consistent usage: The MMoOn Core ontology presents one possible modelling of the morphology domain that can be adapted for the descriptive needs of a certain language documentation. The interoperability effect, which is the main driver for a shared data reuse, largely depends on the consistent as well as greatest possible reuse of the ontological elements. In this respect, the model constrains the applicability of the proposed ontology elements in a prescriptive manner. The classes, properties, their definitions, instances and established restrictions of the MMoOn Core ontology cannot be changed without affecting the semantics of datasets already using this vocabulary. Therefore, the desired cross-disciplinary usability relies on the acceptance and appropriate application of the linguistic terminology and modelling of MMoOn Core by the data creators from various research backgrounds. A misuse of the vocabulary to suit non-intended representations, the recreation of vocabulary elements with different labels or an unconnected creation of new classes or properties in the schema ontologies increase the danger of arriving at isolated data silos again.

Domain delimitation: The intradisciplinary overlap of linguistic domains implies that a domain ontology like MMoOn Core cannot be exhaustive. Even though it covers morphological data to a great extent and interconnections to OntoLex-*lemon* account for an overlap with the domain of lexical language data, certain linguistic aspects are not representable with this ontology. This includes an explicit modelling of morpho-phonological operations such as the deletion, addition, alternation or reduplication of segments. To this extent the expressivity of MMoOn Core is limited to the description of the initial and resulting morphemic elements but does not model processes or operations as such. Further, the ontology is not specified for describing diachronic morphological data like etymology or evolutionary reconstructions. However, both areas of linguistic processes and historical language data affect not only the morphology domain but also the lexical, syntactic and phonological data domains. As a result, they should be separately semantically modelled with the emerging ontologies being adequately aligned to MMoOn Core and even reusing existing elements when appropriate. Such a generic alignment of linguistic domain ontologies is hardly realisable for one ontology creator alone but requires all ontology creators and the potential data users to establish economic and meaningful domain ontology interconnections.

Ontology expertise: The creation, understanding and usage of MMoOn morpheme inventories requires comprehensive multilingualism with regard to the so called ontology vocabularies. Next to the MMoOn Core-specific vocabulary knowledge about OWL, RDF and OntoLex-*lemon* is required. Due to the linguistic domain overlap it is possible that the MMoOn Core ontology alone will not suffice to represent morphological data for some data creators. This might include necessary meta data that needs to be expressed with the dataset or indispensable additional ontological elements that have to be newly created. Therefore, data creators have to search for already existing relevant ontologies, understand the concepts and relations they describe and evaluate and compare them with regard to the effects of combining only parts of them. Consequently, the semantic modelling approach proposed with MMoOn Core is embedded within a wider ontological network that concomitantly needs to be studied in advance. Since the Linked Data principles generally pose no restrictions on the way ontologies are reused to represent data the cross-disciplinary applicability of evolving morphological datasets relies on the ability of the researchers to literally translate their data with the most suitable selection of ontological elements into RDF in a way that also accounts for a shared data reuse.

Expressivity: In adherence to the Linked Data principles the MMoOn Core ontology is based on the standards of OWL and RDF. With regard to its expressiveness it is, therefore, ultimately limited to the realm of formal semantics defined by them. In the design of Linked Data the so called resources are mainly intended to be references to entities in the real world like persons and objects, i.e. which ontologically (in the philosophical sense)

correspond to first-order entities that are relatively uncontroversial. A large part of linguistic data, however, is considered to consist mostly of abstract entities (excluding attestations) which are often very controversially described across the scientific discourse. Two direct consequences follow from this with regard to the linguistic adequacy of morpheme inventories. First, linguistic data is the product of a theoretical interpretation and can consequently be described in various ways leading to multiple unique resources as references to the same entity. The differences of inter-annotator agreements illustrate this issue. The same word in a corpus can result in different categorisations and would require separate resources to represent the diverging interpretations. As a result, more explicit semantics has to be established to express these differences in order to prevent a distortion of the OWL reasoner outputs. This also demands a very cautious handling of URIs as representing different entities accordingly. Second, the required type assertion that assigns an instance to be a member of a class is restricted to a binary choice. This creates categorisation issues with the concept of OWL classes for linguistic categories. For theoretical reasons it can happen that a linguistic expression or unit cannot be clearly categorised. The part of speech classes, for example, are determined by semantic features that are organised along a continuum. E.g. the feature of time stability constitutes such a continuum which results in assumptions leading to debates whether a word is classified as a verb or a noun resulting in conflating categories like verbal nouns. With OWL, however, it is not possible to express that some language element is a member of a class to a certain extent and the data creator is forced to make an inaccurate choice. These two differences in the nature of linguistic data directly affect the reasoning power of OWL and, hence, also of the MMoOn Core ontology in terms of the validity of the obtained inferences. Workarounds to account for the lack of expressivity can be implemented but might lead to diminished data interoperability. Moreover, they often require to add more formalised semantics to a dataset. Such an increase in ontological expressivity then results in a greater reasoning power but may also involve a higher error rate.

Human data processing: With the approach of semantic modelling with ontologies the task to derive insights from data that was formerly reserved to humans is handed over to machines which are equipped with formal semantics to handle data volumes the human mind cannot process. However, the URI-based structure of Linked Data datasets is very machine-centred. It enables the computational access and processing of the data but the resulting graph data is not straightaway understandable for the human user. Even though the link structure allows to navigate through the graph the entirety and interconnectedness is still difficult to comprehend. In order to actually serve research and support cross-disciplinary data reuse a transfer back from RDF into a human-readable data representation is required. Open source tools are available to generate an ontology view, table and

list outputs or transformations into other formats like XML. Yet, more specific software and tools that reconstruct interlinear glossed text or facilitate the creation, management and analysis of Linked Data language resources are still missing. Moreover, all these technical extensions around Linked Data need to be acquired at first by the data creators and users in order to exploit its full potential. Such extensions additionally involve means to query, evaluate and edit RDF and pose a fundamental prerequisite for linguists if they want to integrate semantic data representation into their everyday research work. Finally, the MMoOn Core ontology and the resulting datasets are limited to the machine-readable RDF view on morphological data. The other aspects that have been mentioned and influence the human capabilities of data processing and reuse across different disciplines depend on the ability of the individual researchers to work with Linked Data.

Overall, the considerations just described indicate potential pitfalls that will eventually limit or reduce the reuse of morphological data if not carefully addressed by the data creators. Beyond that, they illustrated that the MMoOn Core ontology as a necessary condition for inducing cross-disciplinary morphological data use simultaneously entails further technological challenges that impact its adequate application in language data-driven research.

Chapter 4

Conclusion

Within this thesis the scientific foundation has been established to induce the cross-disciplinary usage of morphological language data. It has been illustrated that the three research fields of linguistics, computational linguistics and content mining can reach a knowledge acquisition gain in their individual fields by reusing morphological language data that originated in the other fields. The prevalent format barriers that prevented a shared data reuse so far could be overcome by the application of the semantic modelling approach that implemented Linked Data and ontologies for creating interoperable morphological data in the RDF format.

The six publications contributed to this thesis present the conducted research that realised the hitherto missing requirements in order to achieve this result. Evidence for an improvement of the results of cross-disciplinary tasks has been provided in [P1] by a linguistic investigation of the morphological complexity of German proper nouns for the computational linguistic task of NER. Incorporating fine-grained morphological data into NER systems can lead to better system performances and enables direct entity linking in the context of content mining. The necessary domain ontology to create such morphological data has been developed in the form of the MMoOn Core ontology. Publication [2] and [3] described the ontology in detail and elaborated on its adequacy to build the foundation for the creation of semantically represented and interoperable morphological data in RDF. Initiated by the work on the MMoOn Core ontology, the OntoLex-*lemon* Morphology Module emerged and has been introduced in [P4]. This ontology is another semantic modelling possibility that broadens the ontological application scope by targeting automated morphological data generation and, thus, potentially increases the amount of future morphological datasets. Finally, The Hebrew Morpheme Inventory and the Xhosa RDF dataset outlined in [P5] and [P6] proved the applicability of the MMoOn Core ontology. They demonstrated the added value obtained by the datasets' internal and external enrichment that could be achieved due to the underlying ontological data structure and RDF data format.

Beyond the two datasets created within [P5] and [P6], the MMoOn Core

ontology has been taken up by other researchers as well. One application is envisaged in the context of the linguistic documentation and description of the highly inflectional Cherokee language based on very detailed syllabary text sources (Bourns, 2019). This case aims at data interoperability as well as the best possible granularity of the linguistic representation within a multilingual project setting focusing on preserving and studying endangered languages among Native American speech communities. Another incidence of the reuse of the MMoOn Core ontology is in the area of NLP, where it has been used for creating a corpus analyser tool (Mukhamedshin et al., 2020). These two use cases confirm the intended cross-disciplinary reusability of MMoOn Core for the representation of morphological language data.

As a result, the outcomes of the six publications can be regarded as the verification of the hypothesis that semantically modelled and represented morphological data enhances the cross-disciplinary usage of language data in general. Given that morphological data based on MMoOn Core overcomes the former restrictions posed on morphological data that was distributed across lexicons and grammars, it can be now integrated into the wider network of linguistic domain ontologies. This enables a direct incorporation of morphological data into language datasets of other linguistic domains. Further, it can be expected that the comprehensive collection of meanings and morphemic glosses available in MMoOn Core introduce the investigation of new use cases which are of interest to multiple disciplines, e.g. the assessment of the meanings of unknown words by analysing their sub-word units or the development of a morphological tagger by using the glosses.

To conclude, the goal of this thesis, to contribute to the creation of more openly available morphological language data in the RDF format in order to enhance language data-driven research in general, has been achieved. Nonetheless, the conducted research represents only the beginning of the adoption and usage of RDF for morphological data. After all, the research area of semantically modelled morphological language data constitutes a cross-disciplinary field in itself. Therefore, the researchers of traditional linguistics, computational linguistics and language data-based content mining will be hopefully encouraged by this work to jointly create morphological data, share it with the wider research communities as well as to explore the possibilities of Linked Data-based language data reuse.

Chapter 5

Future Work

Both, the area of morphological data and semantic modelling are extensive research fields. The MMoOn Core ontology, as the result of their application, laid the foundation for a cross-disciplinary usage of morphological language data. However, several issues could not be pursued in the realm of this thesis and remain open for future work.

Some possibilities of the MMoOn Core ontology were not fully explored in practice. The classes `mmoon:Morpheme`, `mmoon:MorphologicalRelationship` and the derivational meanings contained in the `mmoon:Meaning` class advance the descriptive range for morphological data but could not be extensively realised in the scope of the Hebrew Morpheme Inventory and Xhosa RDF dataset. The usefulness of their modelling within morphological datasets, therefore, still needs to be demonstrated with more data.

Although the MMoOn Core ontology comes with a large set of meanings, a continuous vocabulary extension is envisaged in order to expand the range of available and ready to use language-independent meanings. An ongoing observation of future MMoOn morpheme inventories is required to include emerging linguistic categories, word-classes and grammatical as well as derivational meanings into the MMoOn Core ontology.

The capability of morphological data to enhance language data in general clearly increases with the interconnection of MMoOn Core with ontologies of other linguistic domains. Therefore, an alignment with the OntoLex-*lemon*, Ligt and PHOIBLE vocabularies needs to be conducted to integrate morphological data representation into the overlapping domains of lexical, interlinear glossed and phonological language data.

Despite the created possibility to express specific grammatical information in the description of the schema ontologies, this functionality has not been actually used by the data owners of the Hebrew and Xhosa RDF datasets. This circumstance leaves room for investigating the practical usability and potential of language-specific ontologies as a formal and machine-processable equivalent of printed grammars.

Finally, the concomitant implications for the knowledge acquisition gain by a shared reuse of morphological and language data through the means

of ontologies and Linked Data point out the future directions for research. Therefore, the researchers of all disciplines that use and create language data are invited to elaborate on the following open questions in more detail:

(i) Will the possibilities of Linked Data and ontologies lead to new methodologies for morphological language data representation that overcome the traditional division into lexicon and grammar?

(ii) To what extent should machine-generated data serve as an additional basis of knowledge acquisition in linguistics as an empirical science?

(iii) What can be done in order to prevent data quality loss caused by the open data policy enabling unsupervised data reuse and interlinking by third parties?

(iv) How can a consistent and more user-oriented application of RDF datasets be ensured which does not exclude less Linked Data-proficient researchers in terms of the whole data management process?

The outcomes of these examinations will ultimately determine how much more RDF datasets containing morphological data will be created and if semantic modelling can be established as a common practice for cross-disciplinary language data-driven research.

Chapter 6

Declaration of Contributions

In the following the author declares her individual contributions to all publications presented in this thesis. This excludes [P3] for which she holds the single authorship.

[P1] Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge

For this publication the investigation of the morphological complexity of German named entities and lexemes that are built from proper nouns according to the identified morphological parameters of inflectional, derivational and compounding degrees has been conducted by the author. This includes the full manual annotation of the three data sub-sets, reconstructing the target named entities from the source named entities given in the corpus data, as well as the error annotations that can be consulted here: https://raw.githubusercontent.com/AKSW/germeval-morph-analysis/master/data/annotation_imports/compl-issues-ann-ranks.tsv. Further, the author states her sole contribution of the Sections 3.3, 4.1 and 5.1 of the publication in addition to co-writing the Sections 1, 5.2 and 6.

[P2] MMoOn Core – The Multilingual Morpheme Ontology

Apart from proof-reading and minor additions to single paragraphs by the co-authors, the author takes credit for the textual contribution of the entire publication. Further, the author worked out the conceptualisation of the MMoOn-Core ontology in its design, purpose, scope, functional details and architectural setup for creating morpheme inventories. The preceding domain analysis as well as the manual description and implementation of all classes, properties and instances resulting in the published ontology file http://mmoon.org/core.rdf have also been conducted by the author. The co-authors shared their expertise on Semantic Web technologies with the author who, at the time of the ontology creation, was still developing her own expertise and proficiency in the field of ontology engineering. To this extent, the technical advice and assistance in implementing the realisation and documentation of the ontology for its publication on the Web is attributed to the co-authors.

[P4] Challenges for the Representation of Morphology in Ontology Lexicons

The author declares to have written the Sections 5, 6, 7 and 8 fully on her own. In addition, she has contributed the first and last two paragraphs of Section 4 to the publication. Moreover, the author has led the development of the Morphology Module within the Ontology-Lexicon community group which resulted in this intermediate module report. In this function she has been highly active in consolidating the individual modelling proposals and arising issues discussed among all group members. The outcomes of this effort are included by the author in this publication in the form of the Morphology Module diagram, the suggestion of new definitions for the module elements and the example graph in Figure 4, which is based on the ontology code that has been created by the author as well (and will result in the publication of the final module specification).

[P5] Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory

Except for Section 5 the author declares to have written the publication on her own. The three co-authors approached the author with the aim to convert their existing tabular Hebrew language data into RDF. Consequently, this publication is a joint effort by all authors who contributed their language expertise on Hebrew, the original Hebrew dataset and their technical knowledge in implementing the data generation and linking. The author created the model in conjunction with the development of the architectural setup that enabled the data transformation into MMoOn-RDF based on the necessary language studies and technical details she acquired from working with the co-authors.

[P6] Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

Sections 3 and 4 of this publication have been written by the author. The initiator of this publication was Dr. Sonja Bosch who was looking for a solution to transform the original Xhosa source data into a reusable and extendable format and asked the authors from the Natural Language Processing Group of the Leipzig University for their assistance. The author was approached by them, because they became aware of the author's work on the MMoOn Core ontology. As a result, the author contributed her expertise in creating language datasets containing morphological data. She developed the BLM based on the architectural setup of MMoOn morpheme inventories including the necessary interconnections to the OntoLex-*lemon* vocabulary. Furthermore, the author provided the mapping of the source data to BLM-RDF and supervised the transformation into the Xhosa RDF dataset by taking the linguistic specifications of the Xhosa language as well as the technical realisation possibilities into account.

Bibliography

Arndt, N., Zänker, S., Sejdiu, G. & Tramp, S. (2019). Jekyll RDF: Template-Based Linked Data Publication with Minimized Effort and Maximum Scalability. In *International Conference on Web Engineering.* Springer, 331–346.

Aronoff, M. (1976). Word formation in generative grammar. In *Linguistic Inquiry Monographs Cambridge, 1.* MIT Press, 1–134.

Aronoff, M. & Fudeman, K. (2011). *What is Morphology?* Fundamentals of Linguistics (Vol. 8). John Wiley & Sons.

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

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In Aberer, K. et al. (Eds.): *The Semantic Web. Lecture Notes in Computer Science.* Springer, 722–735.

Baierer, K., Dröge, E., Eckert, K., Goldfarb, D., Iwanowa, J., Morbidoni, C. & Ritze, D. (2017). DM2E: A linked data source of digitised manuscripts for the digital humanities. In *Semantic Web Journal, 8*(5). IOS Press, 733–745.

Bauer, L. (2004). *A glossary of morphology.* Edinburgh: Edinburgh University Press.

Beesley, K. R. (2003). Computational morphology and finite-state methods. In *NATO Science Series, III: Computer and Systems Sciences, 188.* IOS PRESS, 61–100.

Bender, E. M. (2013). Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. In *Synthesis Lectures on Human Language Technologies, 6*(3). Morgan & Claypool Publishers, 1–184.

Berners-Lee, T. (2006). *Linked Data - Design Issues.* http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. In *Scientific American, 284*(5). JSTOR, 34–43.

Berry, D. M. & Fagerjord, A. (2017). *Digital humanities: Knowledge and critique in a digital age*. John Wiley & Sons.

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Bird, S. & Simons, G. (2003). Seven dimensions of portability for language documentation and description. In *Language, 79*(3). JSTOR, 557–582.

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - The Story So Far. In *International Journal on Semantic Web and Information Systems, 5*(3). IGI Global, 1–22.

Blanke, T., Bodard, G., Bryant, M., Dunn, S., Hedges, M., Jackson, M. & Scott, D. (2012). Linked data for humanities research - The SPQR experiment. In *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. IEEE, 1–6.

Boisson, C., Kirtchuk, P. & Béjoint, H. (1991). Aux origines de la lexicographie: Les premiers dictionnaires monolingues et bilingues. In *International Journal of Lexicography, 4*(4). Citeseer, 261–315.

Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

Booij, G., Lehmann, C., Mugdan, J. & Skopeteas, S. (2000). *Morphologie: Ein internationales Handbuch zur Flexion und Wortbildung / An international handbook on inflection and word-formation* (Vol. 1). Walter de Gruyter.

Booij, G., Lehmann, C., Mugdan, J. & Skopeteas, S. (2004). *Morphologie: Ein internationales Handbuch zur Flexion und Wortbildung / An international handbook on inflection and word-formation* (Vol. 2). Walter de Gruyter.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (n.d.). Towards a module for lexicography in OntoLex. In McCrae, J. P. et al. (Eds.): *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*. CEUR Workshop Proceedings 1899, 74–84.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Gómez-Pérez, A. (2018). Models to represent linguistic linked data. In *Natural Language Engineering, 24*(6). Cambridge University Press, 811–859.

Bourns, J. (2019). Cherokee syllabary texts: Digital documentation and linguistic description. In Eskevich, M. et al. (Eds.): *2nd Conference*

on Language, Data and Knowledge (LDK 2019). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 18:1–18:6.

Chiarcos, C., Hellmann, S. & Nordhoff, S. (2012). The Open Linguistics Working Group of the Open Knowledge Foundation. In Chiarcos, C., Hellmann, S. & Nordhoff, S. (Eds.): *Linked Data in Linguistics*. Springer, 153–160.

Chiarcos, C., McCrae, J. P., Cimiano, P. & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In Oltramari, A. et al. (Eds.): *New Trends of Research in Ontologies and Lexical Resources*. Springer, 7–25.

Chiarcos, C., Moran, S., Mendes, P. N., Nordhoff, S. & Littauer, R. (2013). Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments. In *The People's Web Meets NLP*. Springer, 315–348.

Chiarcos, C., Nordhoff, S. & Hellmann, S. (2012). *Linked Data in Linguistics. Representing Language Data and Metadata.* Heidelberg: Springer.

Ciotti, F., Lana, M. & Tomasi, F. (2014). TEI, Ontologies, Linked Open Data: Geolat and Beyond. In *Journal of the Text Encoding Initiative, 8*. Text Encoding Initiative Consortium.

De Saussure, F. (1989). *Cours de linguistique générale* (Vol. 1). Otto Harrassowitz Verlag.

Eckart, T., Bosch, S., Goldhahn, D., Quasthoff, U. & Klimek, B. (2019). Translation-based dictionary alignment for under-resourced Bantu languages. In Eskevich, M. et al. (Eds.): *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 17:1–17:11.

Eckart, T., Klimek, B., Goldhahn, D. & Bosch, S. (2018). Using linked data techniques for creating an IsiXhosa lexical resource – A collaborative approach. In Skadina, I. and Eskevich, M. (Eds.): *CLARIN Annual Conference 2018*. 26–29.

Ekinci, E. & İlhan Omurca, S. (2020). Concept-LDA: Incorporating Babelfy into LDA for aspect extraction. In *Journal of Information Science*. SAGE Publications Sage UK: London, England, 406–418.

Ekinci, E. & Omurca, S. İ. (2018). Babelfy-Based Extraction of Collocations from Turkish Hotel Reviews. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. IEEE, 1–5.

Farber, B., Freitag, D., Habash, N. & Rambow, O. (2008). Improving NER in Arabic Using a Morphological Tagger. In Calzolari, N. et al. (Eds.): *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA.

Färber, M., Noullet, K. & El Asmar, B. (2018). Annotating Domain-Specific Texts with Babelfy: A Case Study. In *Proceedings of the 1st International Workshop on EntitY REtrieval (EYRE 2018)*.

Farrar, S. & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. In *GLOT International*. 97–100.

Forkel, R. (2014). The cross-linguistic linked data project. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A. & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. In *Scientific Data*. Nature Publishing Group, 1–10.

Frodeman, R. (2013). *Sustainable knowledge: A theory of interdisciplinarity*. Springer.

Granger, S. & Paquot, M. (2012). *Electronic lexicography*. OUP Oxford.

Gruber, T. (2016). Ontology. In Liu, Ling and Özsu, M. Tamer (Eds.): *Encyclopedia of Database Systems*. New York: Springer, 1–3.

Gudivada, V. N. (2018). Natural language core tasks and applications. In *Handbook of Statistics*. Elsevier, 403–428.

Halle, M. & Marantz, A. (1992). Distributed Morphology and the Pieces of Inflection. In *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. MIT Press, 111-176.

Haspelmath, M. (2014). *The Leipzig Style Rules for Linguistics*. http://www.uni-regensburg.de/sprache-literatur-kultur/sprache-literatur-kultur/allgemeine-vergleichende-sprachwissenschaft/medien/pdfs/haspelmath_2014_style_rules_linguistics.pdf

Haspelmath, M. & Sims, A. D. (2013). *Understanding morphology*. Routledge.

Heyer, G., Quasthoff, U. & Wittig, T. (2006). *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. Bochum: W3L.

Hockett, C. F. (1954). Two models of grammatical description. In *Word*. Taylor & Francis, 210–234.

Hudson, R. (1984). *Word grammar*. Oxford: Blackwell.

Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J. & Keravuori, K. (2019). BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research. In *European Semantic Web Conference*. Springer, 574–589.

Janicki, M. (2019). *Statistical and Computational Models for Whole Word Morphology* (Doctoral dissertation). Universität Leipzig.

Karttunen, L. & Beesley, K. R. (2001). A short history of two-level morphology. In *ESSLLI-2001 Special Event titled "Twenty Years of Finite-State Morphology"*.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K. & Chiarcos, C. (2019). Challenges for the representation of morphology in ontology lexicons. In Kosem, I. et al. (Eds.): *Electronic Lexicography in the 21st Century (elex 2019): Smart Lexicography*. Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 570–591.

Koskenniemi, K. (1983). Two-level model for morphological analysis. In *International Joint Conferences on Artificial Intelligence*. 683–685.

Marcińczuk, M., Kocoń, J. & Janicki, M. (2013). Liner2–a customizable framework for proper names recognition for Polish. *Intelligent tools for building a scientific information platform*. Berlin, Heidelberg: Springer, 231–253.

Matthews, P. H. (1972). *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*. Cambridge: Cambridge University Press.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The Ontolex-Lemon model: Development and applications. In Kosem, I. et al. (Eds.): *Proceedings of eLex 2017 conference: Lexicography from Scratch*. Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 19-21.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S. et al. (2016). The open linguistics working group: Developing the Linguistic Linked Open Data cloud. In Calzolari, N. et al. (Eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. ELRA, 2435–2441.

Mendes, P. N., Jakob, M., Garcıúa-Silva, A. & Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. 1–8.

Moro, A., Raganato, A. & Navigli, R. (2014). Entity linking meets word sense disambiguation: A unified approach. In *Transactions of the Association for Computational Linguistics*. MIT Press, 231–244.

Mukhamedshin, D., Nevzorova, O. & Kirillovich, A. (2020). Using FLOSS for Storing, Processing and Linking Corpus Data. In *IFIP International Conference on Open Source Systems*. Springer, 177–182.

Nielsen, S. (2017). Lexicography and interdisciplinarity. In *Routledge Handbook of Lexicography*. Routledge, 93–104.

Nurmikko-Fuller, T., Jett, J., Cole, T. W., Maden, C., Page, K. R. & Downie, J. S. (2016). A Comparative Analysis of Bibliographic Ontologies: Implications for Digital Humanities. In *DH*. 639–642.

Osselton, N. E. (1990). English lexicography from the beginning up to and including Johnson. In *Wörterbücher: Ein internationales Handbuch zur Lexikographie*. Mouton de Gruyter, 1943–53.

Payne, T. E. (1997). *Describing morphosyntax: A guide for field linguists*. Cambridge University Press.

Perlmutter, D. (1988). The split morphology hypothesis: Evidence from Yiddish. In *Theoretical morphology: Approaches in modern linguistics*. San Diego: Academic Press, 79–100.

Scalise, S. (1988). Inflection and derivation. In *Linguistics*. Berlin/New York: Walter de Gruyter, 561–582.

Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In *Electronic Lexicography*. Oxford University Press, 107–118.

Van Leeuwen, T. (2005). Three models of interdisciplinarity. In *A new agenda in (critical) discourse analysis: Theory, methodology and interdisciplinarity*. John Benjamins Publishing, 3–18.

Vanhoutte, E. (2013). *Defining Digital Humanities: A Reader*. Ashgate Publishing, Ltd.

Weichselbraun, A., Gindl, S. & Scharl, A. (2014). Enriching semantic knowledge bases for opinion mining in big data applications. In *Knowledge-based systems, 69*. Elsevier, 78–85.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data, 3*. Nature Publishing Group.