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.

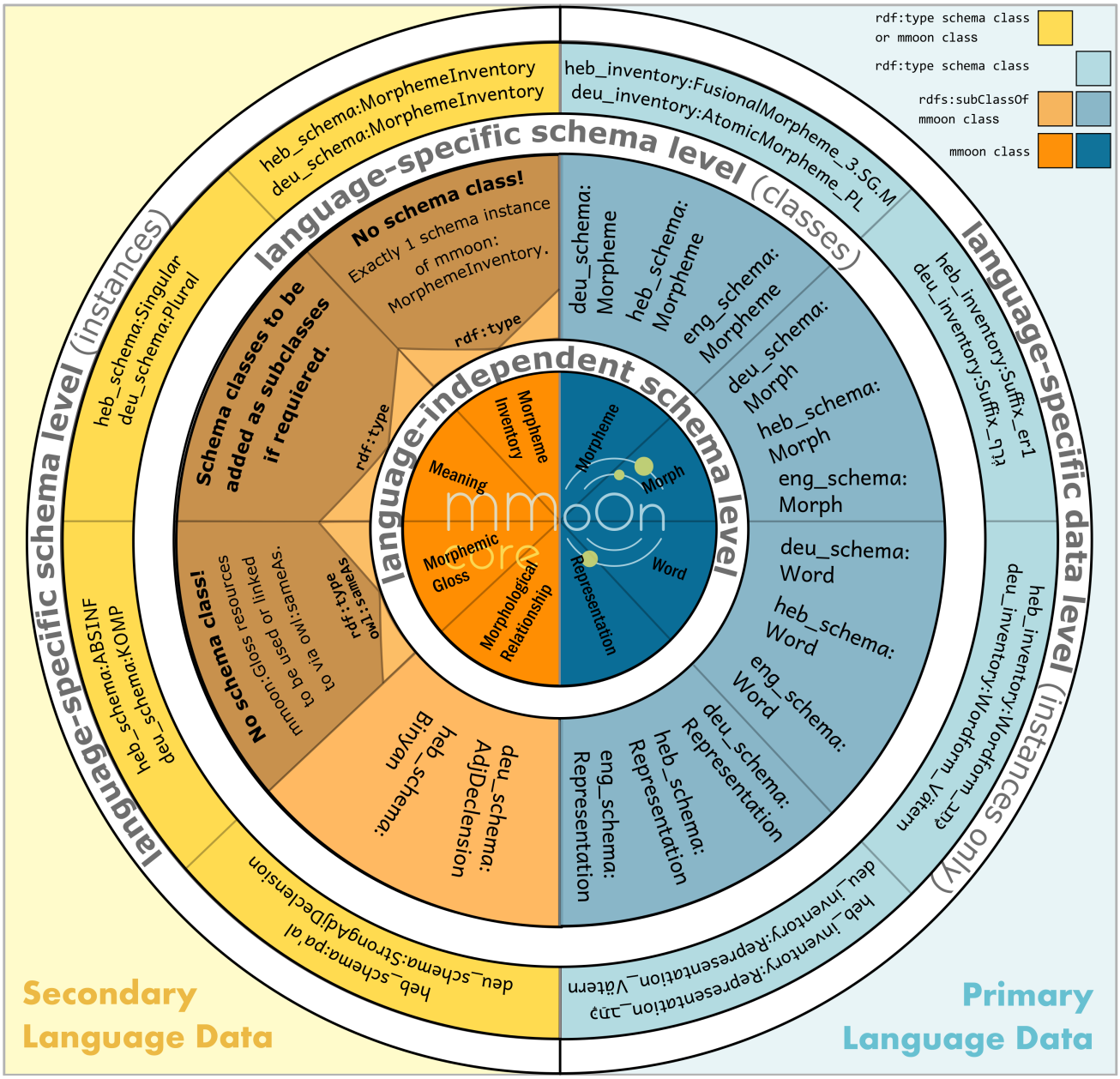

Figure 1: Architectural setup of MMoOn morpheme inventories.

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.

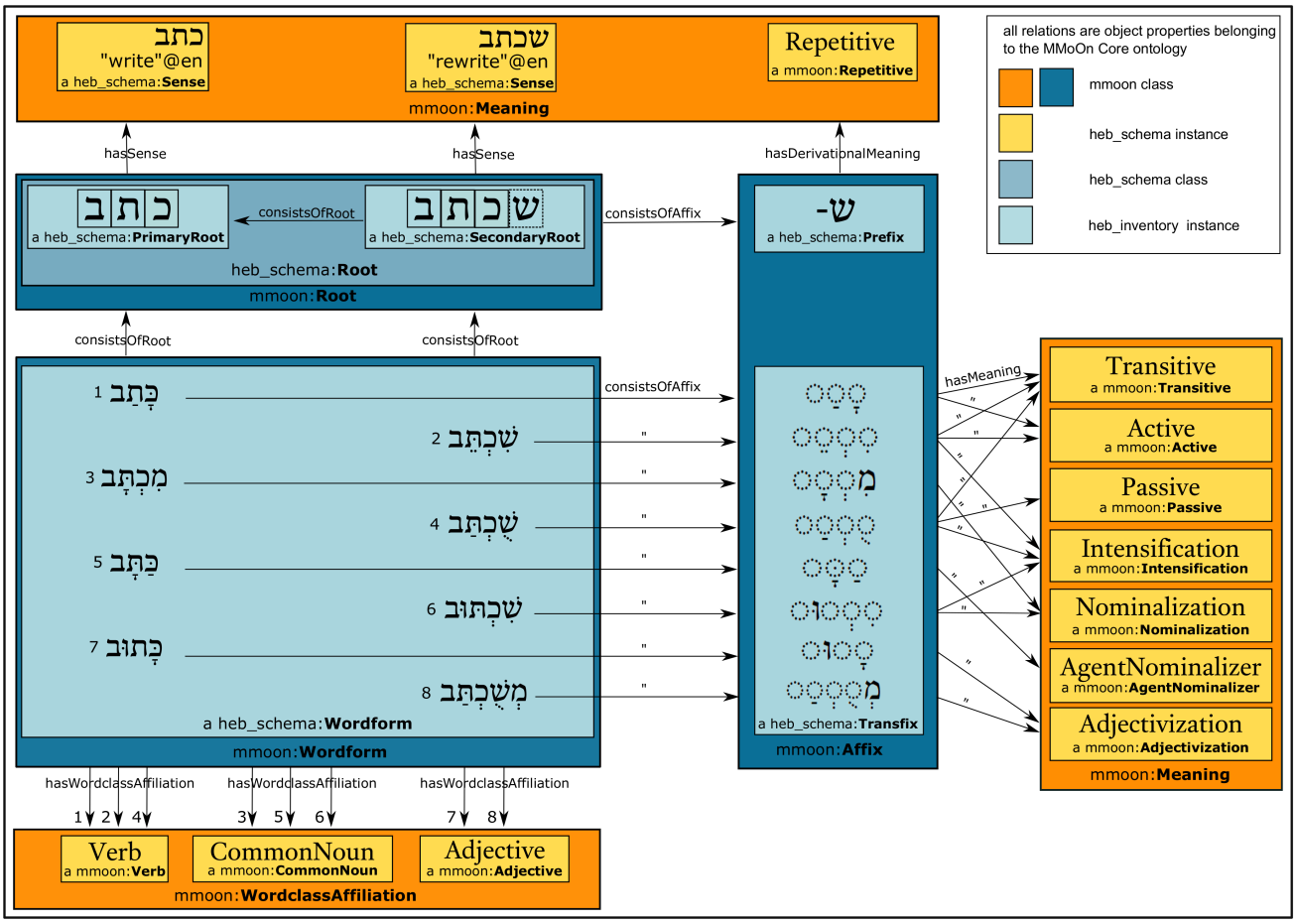

Figure 2: An example for Hebrew root derivation modelled with MMoOn.

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

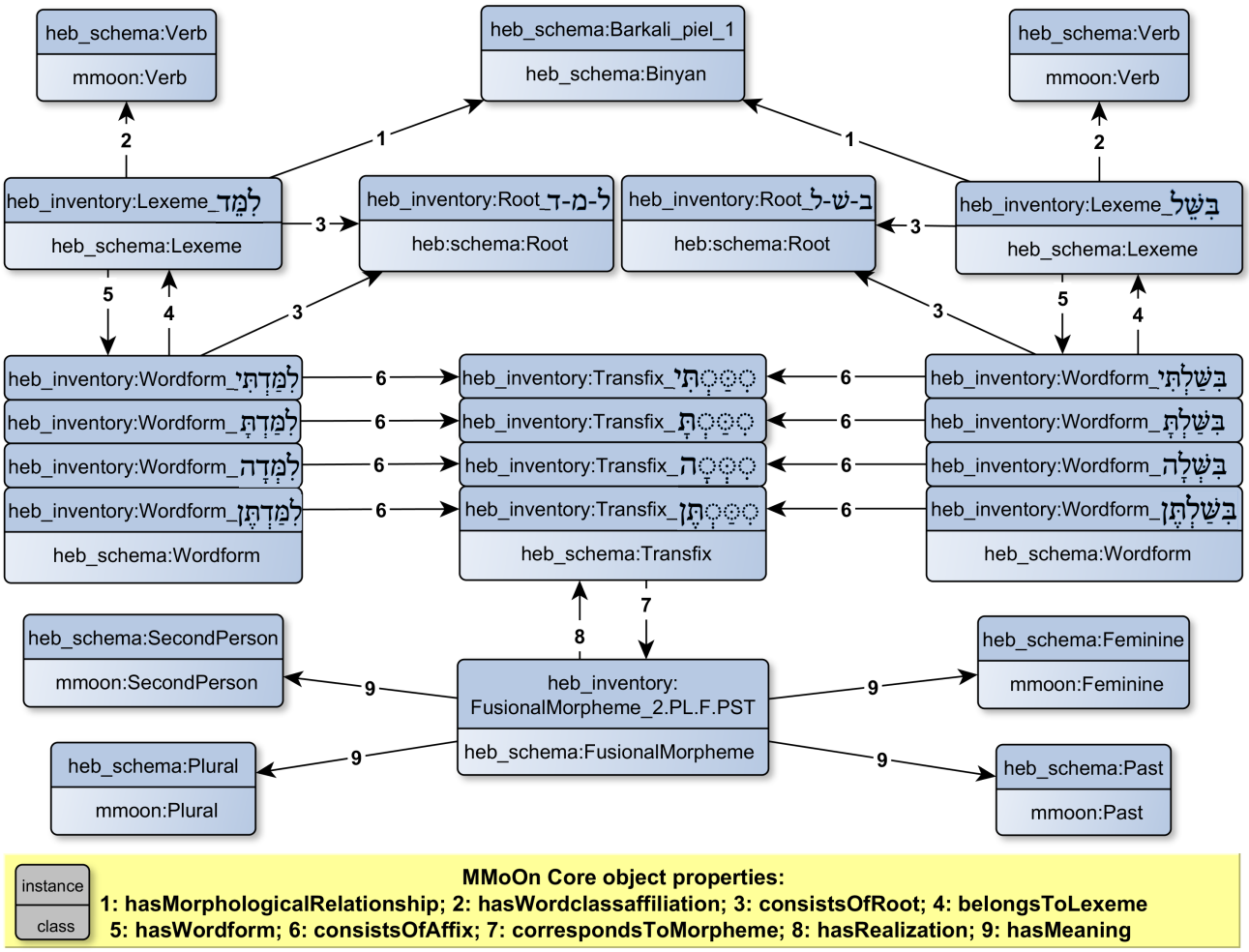

Figure 3: Morphological data of verbs in the Hebrew Morpheme Inventory.

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*.