Towards the Representation of Hashtags in Linguistic Linked Open Data Format

Thierry Declerck
Dept. of Computational Linguistics,
Saarland University, Saarbrücken,
Germany
declerck@dfki.de

Piroska Lendvai
Dept. of Computational Linguistics,
Saarland University, Saarbrücken,
Germany
piroska.r@gmail.com

Abstract

A pilot study is reported on developing the basic Linguistic Linked Open Data (LLOD) infrastructure for hashtags from social media posts. Our goal is the encoding of linguistically and semantically enriched hashtags in a formally compact way using the machine-readable *OntoLex* model. Initial hashtag processing consists of data-driven decomposition of multi-element hashtags, the linking of spelling variants, and part-of-speech analysis of the elements. Then we explain how the *OntoLex* model is used both to encode and to enrich the hashtags and their elements by linking them to existing semantic and lexical LOD resources: *DBpedia* and *Wiktionary*.

1 Introduction

Applying term clustering methods to hashtags in social media posts is an emerging research thread in language and semantic web technologies. Hashtags often denote named entities and events, as exemplified by an entry from our reference corpus that includes Twitter [1] posts ('tweets') about the Ferguson unrest[2]: *"#foxnews #FergusonShooting is in a long line of questionable acts by the police. Because some acted out does not excuse the police."*

In recent work (Declerck and Lendvai, 2015) we have applied string and pattern matching to address lexical variation in hashtags with the goal of normalizing, and subsequently contextualizing hashtagged strings. Types of contexts for a hashtag can be derived from e.g. hashtag co-occurrence and semantic relations between hahstags; representing such contexts necessitates the understanding of the linguistic and extra-linguistic environment of the social media posting that features the hashtag.

In the light of recent developments in the Linked Open Data (LOD) framework, it seems relevant to investigate the representation of language data in social media so that it can be published in the LOD cloud. Already the classical Linked Data framework included a growing set of linguistic resources: language data – i.e. human-readable information connected to data objects by e.g. RDFs annotation properties such as 'label' and 'comment' –, have been suggested to be encoded in machine-readable representation[3]. This triggered the development of the *lemon* model (McCrae et al., 2012) that allowed to optimally relate, in a machine-readable way, the content of these annotation properties with the objects they describe.

While LOD enables connecting and querying databases from different sources[4], the recently emerging Linguistic Linked Open Data (LLOD) facilitates connecting and querying also in terms of linguistic constructs. Based on the activities of the Working Group on Open Data in Linguistics[5] and of projects such as the European FP7 Support Action "LIDER"[6], the linked data cloud of linguistic resources is expanding.

Our goal in the current study is to develop and promote the modeling of linguistic and semantic phenomena related to hashtags, adopting the On-

[1] twitter.com
[2] https://en.wikipedia.org/wiki/Ferguson_unrest
[3] (Declerck and Lendvai, 2010) discussed already the possible benefits of the linguistic annotation of this type of language data.
[4] A more technical definition of Linked Data is given at http://www.w3.org/standards/semanticweb/data
[5] http://linguistics.okfn.org/
[6] http://www.lider-project.eu/.

Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, pages 16–22,
Hissar, Bulgaria, 11 September 2015.

toLex model[7]. This model, a result of the W3C Ontology-Lexicon community group[8], lies at the core of the publication of language data and linguistic information in the LLOD cloud[9]. In the next sections we briefly present the current state of OntoLex, then summarize our approach to hashtag processing, after which our LOD and LLOD linking efforts are explained in detail, finally leading us to future plans.

2 The OntoLex model

The OntoLex model has been designed using the Semantic Web formal representation languages OWL, RDFS and RDF[10]. It also makes use of the SKOS and SKOS-XL vocabularies[11]. OntoLex is based on the ISO Lexical Markup Framework (LMF)[12] and is an extension of the *lemon* model. OntoLex describes a modular approach to lexicon specification.

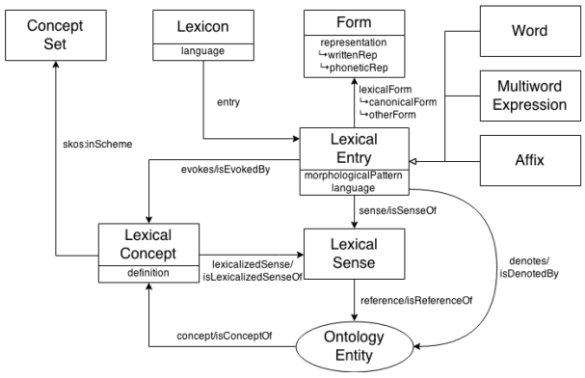

Figure 1: The core model of OntoLex. Figure created by John P. McCrae for the W3C Ontolex Community Group.

With OntoLex, all elements of a lexicon can be described independently, while they are connected by typed relation markers. The components of each lexicon entry are linked by RDF encoded relations and properties. Figure 1 depicts the overall design of the core OntoLex model.

An important relation for us will be 'reference' that represents a property that supports the linking of senses of lexicon entries to knowledge objects available in the LOD cloud so that the meaning of a lexicon entry can be referred to appropriate resources on the Semantic Web.

Additionally to the core model of OntoLex, we make use of its decomposition module[13], which is important for the representation of segmented hashtags. The relation of this module to a lexical entry in OntoLex is displayed in Figure 2.

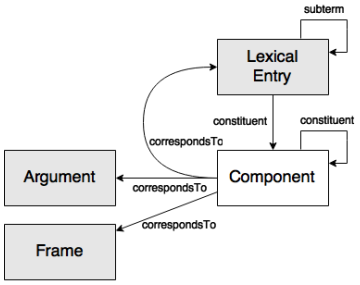

Figure 2: The relation between the decomposition module and the lexical entry of the core module. Figure created by John P. McCrae for the W3C Ontolex Community Group.

3 Hashtag analysis and decomposition

The hashtag set we work with originates from tweets collected about both the Ferguson and the Ottawa shootings[14], as part of a journalistic use case defined in the PHEME project[15]. Below we give examples of the hashtags that we encoded in a lexicon using the OntoLex guidelines:

#FergusonShooting, #fergusonshooting, #FERGUSON, #FERGUSONSHOOTING, #FergusonShootings, #OttawaShooting, #ottawashooting, #Ottawashooting, #Ottawashootings, #ottawashootings, #OttawaShootings, #Ottawa #SHOOTING, #ottwashooting, #OttwaShooting, #Ottwashooting

In Declerck and Lendvai (2015) we reported on the relation between a hashtag processing approach that we apply in our present study as well, and previous work from the literature. Our goal was to examine if hashtags can be segmented and normalized in a data-driven way. In that study, we processed a different, much larger corpus of

[7] http://www.w3.org/community/ontolex/wiki/Main_Page, and more specifically: http://www.w3.org/community/ontolex/wiki/Final_Model_Specification
[8] https://www.w3.org/community/ontolex/ and https://github.com/cimiano/ontolex
[9] http://linguistic-lod.org/llod-cloud
[10] http://www.w3.org/TR/owl-semantics/ http://www.w3.org/TR/rdf-schema/ http://www.w3.org/RDF/
[11] http://www.w3.org/2004/02/skos/
[12] Francopoulo et al. (2006) and http://www.lexicalmarkupframework.org/
[13] For details see http://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Decomposition_.28decomp.29
[14] See https://en.wikipedia.org/wiki/Ferguson_unrest and https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottawa
[15] http://www.pheme.eu/

tweets than the data set we take as an example in the current paper. We analyzed the distribution of hashtags and devised a simple offline procedure that generates a gazetteer of hashtag elements via collecting orthographical information: element boundaries in hashtags were assumed based on e.g. camel-cased string evidence and collocation heuristics. Using this approach on our current corpus, the hashtag *#JusticeForMikeBrown* will be segmented into four elements, while *#michaelbrown* into two elements. Subsequently, we can establish a link between *'Mike'* and *'michael'*, and type it as *lexical variant*, which we later might want to further categorize into specific types relating to normalization such as paraphrase, orthographic variant, and so on, depending on the goal.

We also proposed morpho-syntactic analysis in terms of part-of-speech and dependency analysis; the latter would detect the semantic head in a hashtag, allowing to establish lexical semantic taxonomy relations between hashtag elements such as hyper-, hypo-, syno- and antonymy. In our current study, part-of-speech information is obtained from the NLTK platform[16], while dependency information is not used.

4 Linking and exploiting LOD resources

We connected hashtags and their elements in the OntoLex model to existing linguistic and semantic LOD resources: *wiktionary.dbpedia.org* and *DBpedia*[17]. The use of other resources in the Linked Data framework, such as BabelNet[18], DBnary[19] and Freebase[20] is also relevant and will be explored in further experiments. The *lemon* model, which is the immediate predecessor of OntoLex, is utilized by wiktionary.dbpedia.org, BabelNet and DBnary.

DBpedia provides access to a rich encyclopedic resource, mainly extracted from Wikipedia infoboxes. It also provides links to popular knowledge bases such as Freebase, wikidata[21], yago[22], but does not provide linguistic information. We access DBpedia via the Python package *SPARQLWrapper*[23]. To link hashtags and hashtag elements to LOD data, we query the following properties in DBpedia[24]:

- 'rdfs:label'
- 'rdfs:comment'
- 'dct: subject'
- 'dbo:abstract'
- 'owl:sameAs'
- 'dbo:wikiPageRedirects'.

The added value of information linked via the **'dbo:wikiPageRedirects'** property is that we are able to link hashtags, or their elements, to alternative spellings and variants that were unseen in our Twitter corpus; e.g. for both hashtag variants seen in our corpus *'foxnews' and 'FoxNews'*, the query returns *FOXNEWS, FOXNews, FOXNews.com, FOX NEWS, FOX News*, etc.

It is also possible to designate a preferred form of a hashtag named entity via this property, e.g. querying DBpedia for *'foxnews'* yields *http://dbpedia.org/resource/Fox_News_Channel*. Since this query returns a URL, we get an indication that it is the full span of this hashtag that designates an existing knowledge object. We use this as a heuristic for preventing our system from proposing a compositional analysis of '#FoxNews', but allow its segmentation into "Fox News". In case no such a result is returned when querying a multi-item hashtag, its segmented elements are subject to individual LOD querying and linking (e.g. *#myCanada*, *#besafeottawa*).

The **'owl:sameAs'** property is used to retrieve multilingual equivalents of hashtags or hashtag elements. For example, querying DBpedia for the values of the owl:sameAs property associated to *'shooting'*, returns among others the following results:

http://fr.dbpedia.org/resource/Tir
http://de.dbpedia.org/resource/Schusswaffengebrauch
http://ja.dbpedia.org/resource/射撃
http://es.dbpedia.org/resource/Tiro_(proyectil)
http://id.dbpedia.org/resource/Penembakan
http://it.dbpedia.org/resource/Tiro_(balistica)
http://ko.dbpedia.org/resource/사격
http://nl.dbpedia.org/resource/Schieten
http://pt.dbpedia.org/resource/Tiro_(balística)

[16] http://www.nltk.org/
[17] http://datahub.io/dataset/wiktionary-dbpedia-org and http://wiki.dbpedia.org/
[18] http://babelnet.org/ and (Navigli and Ponzetto, 2012).
[19] http://kaiko.getalp.org/about-dbnary/ and (Sérraset, 2014).
[20] https://www.freebase.com/
[21] https://www.wikidata.org/wiki/Wikidata:Main_Page
[22] www.mpi-inf.mpg.de/yago/
[23] https://rdflib.github.io/sparqlwrapper/
[24] The prefixes 'dbo' and 'dct' stand for http://dbpedia.org/ontology/ and http://purl.org/dc/terms/subject, respectively.

wiktionary.dbpedia.org provides an "open-source framework to extract semantic lexical resources from Wiktionary, including information about language, part of speech, senses, definitions, lexical taxonomies, and translations"[25]. For this LOD dataset there is also a SPARQL endpoint[26] that we query. A query on *'shooting'* returns a number of results, out of which we select the relevant one for our hashtag lexicon: i.e., the senses for the English noun 'shooting', given that our tweets are in English and from NLTK we know that shooting is a noun[27]:

- *http://wiktionary.dbpedia.org/resource/shooting-English-Noun-2en*
- *http://wiktionary.dbpedia.org/resource/shooting-English-Noun-1en*

Verbs and adjectives, as well as sense disambiguation is currently unaddressed in our system.

5 OntoLex Encoding of Hashtags

5.1 Lexicon

The first step in creating the OntoLex representation of hashtags is to define a lexicon that is the container for the hashtag entries.

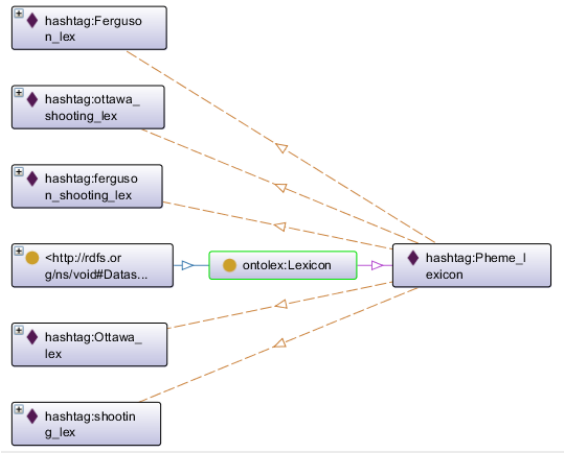

Figure 3: Graphical view of the hashtag lexicon with a entries

The graphical representation of this lexicon and its entries (here in limited numbers) is given in Figure 3[28]. Figure 4 provides the legend for arc colors displayed in all the representation graphics.

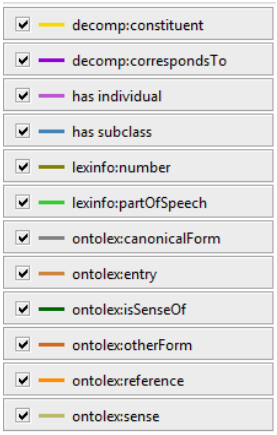

Figure 4: Legend for arc colors in graphical representations of our OntoLex model.

The RDF code underlying the representation in Figure 3 is:

```
hashtag:Pheme_lexicon
  rdf:type ontolex:Lexicon ;
  ontolex:entry hashtag:Ferguson_lex ;
  ontolex:entry hashtag:Ottawa_lex ;
  ontolex:entry
hashtag:ferguson_shooting_lex ;
  ontolex:entry hashtag:ottawa_shooting_lex
;
  ontolex:entry hashtag:shooting_lex ;
.
```

5.2 Lexical Entries

Lexical entries are instances of the class ontolex:LexicalEntry. As shown in Figure 5, the class LexicalEntry introduces three sub-classes: Word, MultiWordExpression and Affix, for now we populate the model with instances for the classes ontolex:Word and ontolex:MultiWordExpression. The corresponding coding for the entries "shooting_lex" and "ferguson_shooting_lex" is given below. We discuss the use of the property ontolex:denotes in Section 5.4.

```
  hashtag:shooting_lex
    rdf:type ontolex:Word ;
    ontolex:canonicalForm
  hashtag:shooting_form ;
    ontolex:denotes
  <http://dbpedia.org/page/Shooting> ;
  .
  hashtag:ferguson_shooting_lex
    rdf:type ontolex:MultiWordExpression ;
    rdf:_1 hashtag:ferguson_component ;
    rdf:_2 hashtag:shooting_component ;
    rdfs:label "fergusonshooting"@en ;
    decomp:constituent
  hashtag:ferguson_component ;
```

[25] Quotation from http://datahub.io/dataset/wiktionary-dbpedia-org
[26] http://wiktionary.dbpedia.org/sparql
[27] Details follow in Section 5.
[28] The ontology graphs presented in this paper are generated by the OntoGraf – Protégé Desktop plug-in. For more details, see http://protegewiki.stanford.edu/wiki/OntoGraf.

```
  decomp:constituent
hashtag:shooting_component ;
  ontolex:canonicalForm
hashtag:ferguson_shooting_form ;
  ontolex:language "en"^^xsd:string ;
  ontolex:otherForm
hashtag:shooting_in_ferguson_form ;
.
```

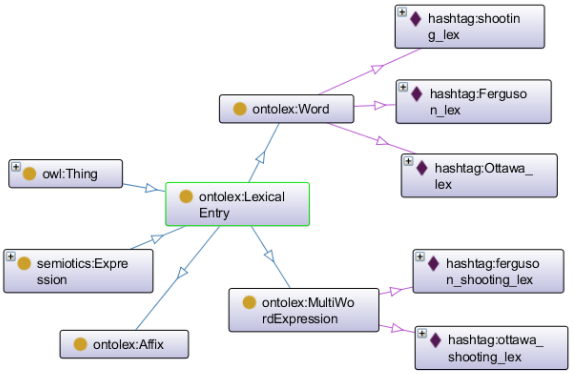

Figure 5: Subclasses of LexicalEntry, with instances for Word and MultiWordExpression.

5.3 Decomposition Module

We focus here on the "ferguson_shooting_lex" entry, an instance of the class ontolex:MultiWordExpression, to see how OntoLex supports the encoding of components of complex hashtags that have been segmented by the algorithms described in (Declerck & Lendvai, 2015). The decomposition of the hashtag is marked by the property: decomp:constituent. The value of this property is an instance of the class ontolex:Component. Since the hashtag has been decomposed in two components, the entry will introduce two decomp:constituent properties, with the current values hashtag:ferguson_component and hashtag:shooting_component

We use rdf_1 and rdf_2 as instances of the property rdfs:ContainerMembershipProperty [29] for marking the order of the two components in the compound hashtag. Keeping this information will be relevant for further contextual interpretation. The form "ferguson_shooting" is marked as preferred written representation for the entry, while an alternative form is "shooting_in_ferguson". These two forms are considered paraphrases. Other types of variants are not introduced as instances of a class, but will be added to the values of the relational data type property "writtenRep", with domain "ontolex:Form" and range string values.

[29] See http://www.w3.org/TR/rdf-schema/ for more details.

The interplay between the ontolex:Component instances and the ontolex:MultiWordExpression instances is graphically shown in Figure 6. 'Ferguson' is marked as a component, and as such it will be put to use in decomposing expressions in our corpus such as "Fergusonvigil", "FergusonPD", etc. The property `decomp:corresponds` links the components to the lexical entries in which they occur.

Part-of-speech and Named Entity information is gained from the combined use of the NLTK tagger (delivering 'NN') and the information from DBpedia that 'Ferguson' is a locality. These pieces of information are mapped to the tagset for linguistic information from the lexinfo ontology[30], which is imported into the OntoLex model.

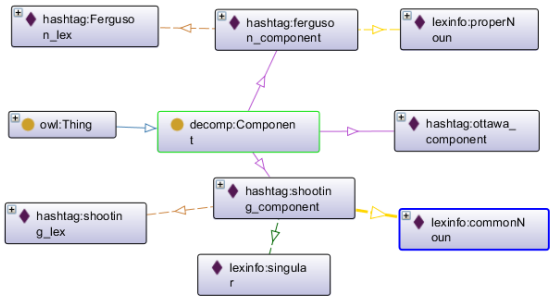

Figure 6: Interplay between components and MultWordExpression entries

Figure 7 supplies more details of the relation between instances of ontolex:Component and ontolox:MultiWordExpression, showing a component ('shooting') shared by various entries.

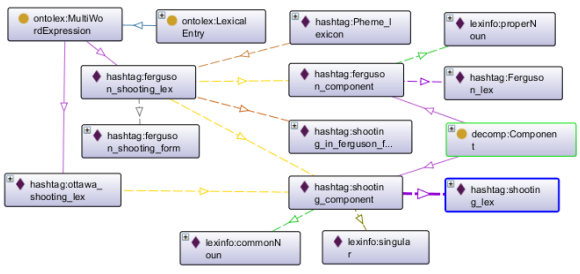

Figure 7: More details of the interplay between Components and MultiWordExpressions, showing how a component ('shooting') is shared by various lexical entries (see the yellow lines).

[30] See http://lexinfo.net/.
Figure 9 shows the lexinfo hierarchy for morpho-syntactic information.

5.4 Linking to LOD resources

In OntoLex there are two ways for linking entries to external semantic resources available in the LOD: ontolex:denotes and ontolex:reference. An example for ontolex:denotes is:

```
hashtag:Ferguson_lex
  rdf:type ontolex:Word ;
  ontolex:denotes
<http://www.dbpedia.org/page/Ferguson,_
Missouri> ;
.
```

Here we see that the lexical entry is linked directly to a DBpedia resource that contains encyclopedic knowledge, via the ontolex:denotes property. Since 'Ferguson' is a Named Entity it is important to know the type of this entity so the disambiguation task related to this string would focus on selecting the correct type. Likewise, to disambiguate common nouns, a selection of correct sense needs to be made. OntoLex offers a property to encode senses of entries, e.g. for the 'shooting' entry in the following way:

```
hashtag:shooting_lex
  rdf:type ontolex:Word ;
  ontolex:canonicalForm
hashtag:shooting_form ;
  ontolex:denotes
<http://dbpedia.org/page/Shooting> ;
  ontolex:otherForm
hashtag:shootings_form ;
  ontolex:sense
hashtag:shooting_noun_sense1 ;
  ontolex:sense
hashtag:shooting_noun_sense2 ;
.
```

The piece of code additionally exemplifies that for this lexical entry we can employ two ways to link to an external LOD resource. Either directly to DBpedia (or another source) via the ontolex:denotes property, or indirectly via the explicit listing of senses and the corresponding property ontolex:sense that has the class ontolex:LexicalEntry as domain and ontolex:LexicalSense as range. The corresponding instances of ontolex:LexicalSense for 'shooting' are:

```
hashtag:shooting_noun_sense1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "An instance of shooting
(a person) with a gun."@en ;
  ontolex:isSenseOf
hashtag:shooting_lex ;
  ontolex:reference
<http://wiktionary.dbpedia.org/page/sho
oting-English-Noun-1en> ;
```

and

```
hashtag:shooting_noun_sense2
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "The sport or activity
of firing a gun."@en ;
  ontolex:isSenseOf
hashtag:shooting_lex ;
  ontolex:reference
<http://wiktionary.dbpedia.org/page/sho
oting-English-Noun-2en> ;
```

The different senses are made explicit to the human reader by the use of the rdfs:comment property. The reader can observe that via the property ontolex:reference we can also link to LOD resources, as we did earlier with the property ontolex:denotes. The main difference between the two properties is the specification of the corresponding domains and ranges, as observable in Figure 1.

Another difference lies in the fact that with ontolex:reference we link to resources encoding lexical senses[31]. This provides more precise and specific semantic information and also creates a more accurate ground for possible translations of the entries. The relation between an entry ('shooting') and its senses is graphically represented in Figure 8:

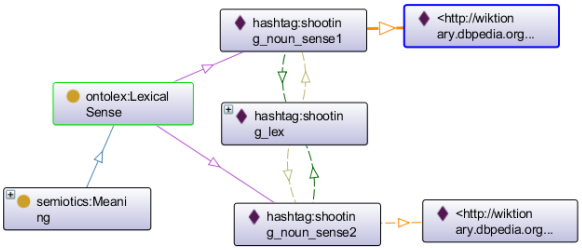

Figure 8: Relation between an entry and its senses

5.5 Part-of-Speech

Concerning the morpho-syntactic information, we map all the information obtained from the NLTK tagger onto the information structure offered by the lexinfo ontology.[32] We display in

[31] But there is no way to enforce this guideline.
[32] As a reminder: http://lexinfo.net/

Figure 9 the relevant part of the lexinfo class hierarchy. There, lexinfo:PartOfSpeech introduces 228 different categories. 'Noun' is defined in lexinfo by reference to the ISOcat http://www.isocat.org/rest/dc/1256 and http://www.isocat.org/datcat/DC-385 entries. Using OntoLex and lexinfo caters for re-using standards from the field of lexical markup.

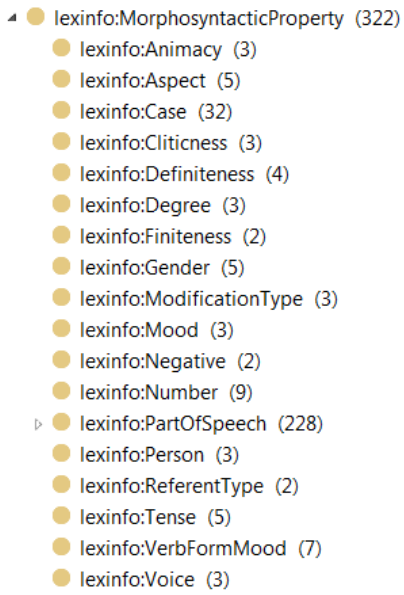

Figure 9: The lexinfo hierarchy of morpho-syntactic information.

Since we are focusing on English data which are morphologically poor, and since OntoLex does not yet provide a final model for the description of morphological information, we postpone the issue of morphological markup till an updated version of our lexical-ontology work on hashtags.

6 Conclusion and future work

We described the current status of our work on porting results of our approach to hashtags normalization onto a standardized representation format suitable for publishing hashtag data in the Linguistic Linked Open Data cloud. The OntoLex model has proven to be an adequate platform for this endeavor.

Next steps of our work will consist in applying the porting algorithm to a larger dataset. The goal is to publish the resulting data in the LLOD cloud, and so to make it semantically interoperable and machine-readable for a variety of language technology applications. To achieve this, we will also integrate our OntoLex representation of hashtags into broader semantic representations of social media data, and transfer the approach to hashtag processing and representation in languages other than English.

Acknowledgments

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233).

References

Cimiano, P. and Unger, C. (2014). Multilingualität und Linked Data. In: T. Pellegrini, H. Sack & S. Auer (eds.) *Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien.* Springer, pp. 153-175.

Declerck, T, and Lendvai, P. (2010). Towards a Standardized Linguistic Annotation of the Textual Content of Labels in Knowledge Representation Systems. *Proceedings of LREC 2010.*

Declerck, T. and Lendvai, P. (2015). Processing and Normalizing Hashtags. *Proceedings of RANLP 2015.*

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J.-P., Cimiano, P. & Navigli, R. (2014). A Multilingual Semantic Network as Linked Data: *lemon*-BabelNet. In C. Chiarcos, J.-P. McCrae, P. Osenova & C. Vertan (eds.) *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, pp. 71-76.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria. (2006). Lexical Markup Framework (LMF). *Proceedings of the fifth international conference on Language Resources and Evaluation.*

McCrae, J.-P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, P., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), pp. 701-719.

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

Sérasset, G. (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web* 05/2015; 6(4):355-361. DOI: 10.3233/SW-140147