Representation of Polarity Information of Elements of German Compound Words

Thierry Declerck[1,2]
[1] Saarland University, Department of Computational Linguistics
[2] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
E-mail: declerck@dfki.de

Abstract

We present on-going work on using formal representation frameworks for encoding polarity information that can be attached to elements of German compound words. As a departure point we have a polarity lexicon for German words that was compiled and ranked on the basis of the integration of four pre-existing polarity lexicons that were available in different formats. As for the formal representation frameworks we are considering for the encoding of the lexical data the *lexicon model for ontologies* (lemon), more specifically its modules *ontolex* (Ontology-lexicon interface) and *decomp* (Decomposition), which have been developed in the context of the W3C Ontology-Lexica Community Group. For the encoding of the polarity information we adopt a slightly modified version of the Marl ontological modelling, developed at the Universidad Politécnica de Madrid.

Keywords: Lemon, Ontology-lexicon interface, Decomposition, Polarity

1. Introduction

Emerson and Declerck (2014) describe algorithms developed in order to generate SentiMerge, a resource that encodes polarity information for German words on the basis of integration processes performed on four pre-existing polarity lexicons for German (Clematide and Klenner, 2010; Remus et al. 2010; Waltinger, 2010 and Klenner et al., 2012). The resulting merged lexicon[1] consists of 15.287 lemmas marked with either positive or negative polarity, indicated by real numbers (from -1.0 to 1.0), to which also a confidence measure is associated. There are 5 levels of confidence, from low (3.536) to high (14.527), with the intermediate levels (5.823, 7.966 and 12.389).

Entry	POS	Polarity Value	Confidence
arbeitslos	AJ	-0.968	14.527
freihalten	V	0.777	7.966
goldhochzeit	N	0.628	5.823
rotsperre	N	-0.628	5.823

Table 1: Examples from SentiMerge

The four examples displayed in Table 1 (*jobless, to keep free, golden wedding anniversary, red card suspension*) show a negative polarity adjective and a negative polarity noun (both marked by the minus sign), a positive polarity verb and a positive polarity noun[2]. In the last column of Table 1, the reader can see the confidence measure computed by the algorithm described in (Emerson and Declerck, 2014).

The examples are compound words and our interest lies in the possibility of marking elements of such compound words with polarity information and, in the longer term, to be able to propose an algorithm for computing the polarity of unknown compound words (i.e. words not included in the SentiMerge lexicon) on the basis of the polarity of their elements, if those are included in the lexicon. Furthermore, our intuition is that the position of an element within a compound is playing a role when it comes to compute the polarity of the compound word.

For our investigation, there is thus the need to be able to represent elements of compound words, including their position within such words. Our choice therefor is the *lexicon model for ontologies* (lemon), which has been first developed within the European project "Monnet" (McCrae et al., 2012) and further refined in the larger context of the W3C Ontology-Lexica Community Group[3]. Of particular relevance for our work are 1) the core module of *lemon,* which describes the so-called Ontology-lexicon interface (*ontolex*) and 2) the Decomposition module (*decomp*) of *lemon*, which marks those elements of the lexicon that are compound or multi-word lexical entries.

This choice is also supported by a study we provided on the use of those *lemon* modules for representing the result of the decomposition of complex English hashtags used in Twitter posts, examples of which are "#StopTheRiots" and the like (Declerck and Lendvai, 2015).

For the representation of polarity information we opted for the Marl ontology (Westerski and Sánchez-Rada, 2013), which has already been adopted for use in the context of sentiment lexicons published in the Linguistic Linked Open Data[4] framework (Buitelaar et al., 2013). We use in this study a slightly modified version of Marl, which has been developed in the context of the European project "TrendMiner" (Krieger and Declerck, 2014), where we called this version of Marl the OP ontology.[5]

[1] Downloadable at https://github.com/guyemerson/SentiMerge

[2] Neutral polarity is indicated by the value „0.0", so for „Abdeckblech" (*cover plate*): abdeckblech N 0.0 7.966.

[3] See https://www.w3.org/community/ontolex/

[4] See http://www.linguistic-lod.org/ for more details.

[5] See http://www.dfki.de/lt/onto/trendminer/OP/opinion.owl

2. The core Module (ontolex) of *lemon*

The *ontolex* model has been designed using the Semantic Web formal representation languages OWL, RDF(S) and RDF[6]. It also makes use of the SKOS vocabulary[7]. *ontolex* has been inspired by the ISO Lexical Markup Framework (Francopoulo et al., 2006)[8], which is based on XML[9].

Ontolex describes a modular approach to lexicon specification. All elements of a lexicon can be described independently, while they are connected by typed relation markers. The components of each lexicon entry in the core module are linked by RDF, SKOS and *ontolex* properties, as this can be seen in Figure 1. A main motivation for the development of *ontolex* is to support the specification of the meaning of lexical entries by pointing to objects described in ontological frameworks, using for this the properties ontolex:denotes or ontolex:reference, offering thus a bridge – or interface – between *knowledge of words* and *knowledge of the world.*

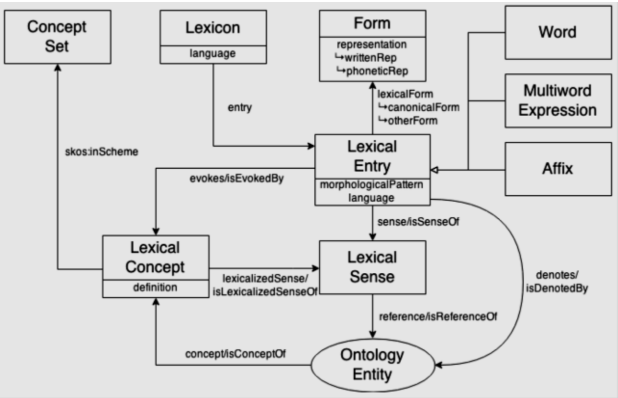

Figure 1: The core model (*ontolex*)
Figure created by John P. McCrae for the W3C Ontology-Lexica Community Group.

3. The decomp Module of *lemon*

Additionally to the core module of *lemon*, we make use of its decomposition module (*decomp*)[10], which has been designed for the representation of multi-word or compound lexical entries. The relation of *decomp* to the core module, and more particularly to the class ontolex:LexicalEntry, is displayed in Figure 2. There, the reader can observe that the components of a compound (or a multi-word) entry are pointed to by the property: decomp:constituent. The range of this property is an instance of the class decomp:Component.

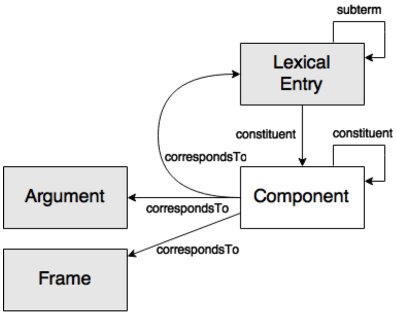

Figure 2: The relation between the decomposition module and the LexicalEntry class of *ontolex*.
Figure created by John P. McCrae for the W3C Ontology-Lexica Community Group

.

As an example (see (1) below), let us consider the German word “Rotsperre” (*red card suspension*). This word is built out of two components, introducing two decomp:constituent properties, with the associated values :Rot_comp and :sperre_comp, which are instances of the class decomp:Component . Those instances reflect the particular form of the components of the compound word. The property decomp:subterm instead “segments” the compound (or multi-word) entry to the corresponding lexical entries. We use rdf_1 and rdf_2 as instances of the property rdfs:ContainerMembershipProperty for marking the order of the two components in the compound word. Keeping this information on the position of the elements can be relevant for further contextual interpretation.

```
(1) :Rotsperre_lex
      rdf:type ontolex:LexicalEntry ;
      lexinfo:partOfSpeech lexinfo:noun ;
      rdf:_1 :Rot_comp ;
      rdf:_2 :sperre_comp ;
      decomp:constituent :Rot_comp ;
      decomp:constituent :sperre_comp ;
      decomp:subterm :Sperre_lex ;
      decomp:subterm :rot_lex ;
      ontolex:denotes   <http://www.oeaw.ac.at/acdh/compound#
            https://www.wikidata.org/wiki/Q1827> .
```

Examples (2) and (3) below show the encoding of the instances of the class decomp:Component:

```
(2) :Rot_comp
      rdf:type decomp:Component ;
      decomp:correspondsTo :rot_lex .
```

```
(3) :sperre_comp
      rdf:type decomp:Component ;
      decomp:correspondsTo :Sperre_lex .
```

[6] See respectively http://www.w3.org/TR/owl-semantics/, https://www.w3.org/TR/rdf-schema/, https://www.w3.org/RDF/
[7] https://www.w3.org/2004/02/skos/
[8] See also http://www.lexicalmarkupframework.org/
[9] Differences between LMF and lemon-ontolex are described at http://lemon-model.net/lemon-cookbook/node46.html
[10] http://www.w3.org/community/ontolex/wiki/Final_Model_Specification

Those instances of decomp:Component are linked to their corresponding lexical entries by the use of decomp:correspondsTo property.

We stress here that instances of decomp:Component can be pointed to by an arbitrary number of compound (or multi-word) lexical entries, like "Löschsperre" (*deletion block*) or the semantically more closely related "Gelbsperre" (*temporary suspension*) for :sperre_comp, or "Rotwein" (*red wine*) for :Rot_comp. This capability leads to the possibility of listing all German strings that play a role as a component in compound words. We consider this approach to the representation of elements of compounds very intuitive and potentially very economical, since one component can be linked to by a large number of entries, or could be used in the context of the generation of compound words.

We note though that we are still investigating if we should keep the capitalization properties of the compound word for marking the components: "Rotsperre" vs "blutrot" (*crimson*). It is yet unclear if we should have the two instances :Rot_comp and :rot_comp.

4. The Marl Ontology

As mentioned above, we opted for the Marl model, described in (Westerski and Sánchez-Rada, 2013), for the encoding of polarity information. Our inspiration for using this model for SentiMerge is the approach proposed in the past Eurosentiment project[11] and in (Buitelaar et al., 2013). The (simplified and slightly modified) encoding of the Spanish word "abandonar" (*to abandon*) in the Eurosentiment project is displayed below (examples 4 and 5):

```
(4)
<http://www.eurosentiment.eu/dataset/general/es/opener/0044/lex
icalentry/abandonar>
    ontolex:sense
     http://www.eurosentiment.eu/dataset/general/es/opener/00
     44/lexicalentry/sense/abandonar_0
    lexinfo:partOfSpeech  lexinfo:verb .
(5)
<http://www.eurosentiment.eu/dataset/general/es/opener/0044/lex
icalentry/sense/abandonar_0>
    a                ontolex:LexicalSense ;
    ontolex:reference
     <http://wordnet-rdf.princeton.edu/wn31/200551194-v> ;
    marl:hasPolarity     marl:negative ;
    marl:polarityValue    -1.0 .
```

Example (4) introduces a lexical entry "abandonar" that has the object ".../abandonar_0" as the value of the property ontolex:sense. Example (5) shows how the polarity information is encoded within this instance of the class ontolex:LexicalSense. As the reader can see, the name of the instance ".../abandonar_0" is underscored with a number. This reflects the possibility that a lexical entry can have various senses, here encoded by referential links to elements of the WordNet resource. By its decision to encode the polarity information within instances of the class ontolex:LexicalSense, the Eurosentiment project relates thus the various polarities an entry can have with its different senses. Since this seems to be a reasonable assumption, we adopt this approach as well. Example (6) displays the lexical sense we associate with the lexical entry "Rotsperre" (see example (1) above).

```
(6) :rotsperre_sense
      rdf:type ontolex:LexicalSense ;
       op:assessedBy :SentiMerge ;
       op:hasPolarity op:Negative ;
       op:maxPolarityValue "1.0"^^xsd:double ;
       op:minPolarityValue "-1.0"^^xsd:double ;
       op:polarityValue "-0.628"^^xsd:double ;
       rdfs:label "Sense for the German word \"Rotsperre\""@en ;
       ontolex:isSenseOf :Rotsperre_lex ;
      ontolex:reference
     <http://de.dbpedia.org/resource/Wettkampfsperre> .
```

The ontological reference that is associated to this sense is the DBpedia entry for "competition ban". Polarity information can be recognized by the use of the prefix "op". We have only one sense for the entry "Rotsperre", but there are more senses for the word "Sperre". Examples (7) and (8) show the encoding for 2 different senses, including also polarity information.

```
(7) :sperre_sense1
       rdf:type ontolex:LexicalSense ;
      op:assessedBy :TD ;
       op:hasPolarity op:Neutral ;
       op:maxPolarityValue "1.0"^^xsd:double ;
       op:minPolarityValue "-1.0"^^xsd:double ;
       op:polarityValue "0.0"^^xsd:double ;
       rdfs:label "A sense for the German word \"Sperre\""@en ;
       ontolex:isSenseOf :Sperre_lex ;
       ontolex:reference <http://de.dbpedia.org/resource/Lock> .
(8) :sperre_sense2
       rdf:type ontolex:LexicalSense ;
       op:assessedBy :SentiMerge ;
       op:hasPolarity op:Negative ;
       op:maxPolarityValue "1.0"^^xsd:double ;
       op:minPolarityValue "-1.0"^^xsd:double ;
```

[11] See http://eurosentiment.eu/. Adopting the approach suggested by Eurosentiment is also instrumental for publishing our lexicon in the Linguistic Linked Open Data (see http://linguistic-lod.org/llod-cloud).

```
 op:maxPolarityValue "1.0"^^xsd:double ;
rdfs:label "A sense for the German word \"Sperre\""@en ;
ontolex:isSenseOf :Sperre_lex ;
ontolex:reference
<http://de.dbpedia.org/resource/Wettkampfsperre> .
```

In (8) we can see that the ontological reference is identical to the one of the sense of “Rotsperre” displayed in (6). Since we are primarily interested in encoding elements of compounds with polarity information, we need to adapt the encoding of the instances of the class decomp:Component (examples (2) and (3)). So for example decomp:sperre_comp needs to be reduplicated in various instances that are linking to the distinct senses of the lexical entry ontolex:Sperre_lex.

```
(9) :sperre1_comp  a            decomp:Component ;
     decomp:correspondsTo  :Sperre_lex ;
     ontolex:sense         :sperre_sense1 .
(10) :sperre2_comp  a           decomp:Component ;
     decomp:correspondsTo  :Sperre_lex ;
     ontolex:sense         :sperre_sense2 .
```

A possible issue with our approach consisting in adding the property ontolex:sense lies in the fact that the domain of this property is in fact the class ontolex:LexicalEntry.

5. Conclusion

We presented in this short paper on-going work dealing with an extension of a German polarity lexicon with polarity information being attached not only to full entries, but also to elements of compound words. We tested and integrated for this purpose two formal representation frameworks: *lemon* and Marl. Future work will consist in applying the suggested modelling to other lexicons as SentiMerge and in trying to derive rules for the segmentation of compounds not included in lexicon, due to the very productive nature of compounding.

6. Acknowledgements

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233) and by the FREME H2020 project (grant No. 644771). The author would like to thank the anonymous reviewers for their very helpful and constructive comments.

7. References

Buitelaar, P., Arcan, M., Iglesias, C.A., Sánchez, J.F. and Strapparava, C. (2013). Linguistic Linked Data for Sentiment Analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL 2013): Representing and linking lexicons, terminologies and other language data*. Collocated with the Conference on Generative Approaches to the Lexicon, Pisa, Italy.

Clematide, S, Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In *Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*. Held in conjunction to ECAI 2010, Lisbon, Portugal.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U. and Wiegand, M. (2012). MLSA - A Multi-layered Reference Corpus for German Sentiment Analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Declerck, T. and Lendvai, P. (2015). Towards the Representation of Hashtags in Linguistic Linked Open Data Format. In *Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data*. Hissar, Bulgaria.

Emerson, G and Declerck, T. (2014). SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework. In *Proceedings of the 2014 Workshop on Lexical and Grammatical Resources for Language Processing*. Dublin, Irland.

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

Klenner, M., Clematide, S., Petrakis, S. and Luder, M. (2012). “Compositional syntax-based phrase-level polarity annotation for German”. In *Proceedings of the 10th International Workshop on Treebanks and Linguistic Theories (TLT 2012)*, Heidelberg, Germany.

Krieger, H.-U. and Declerck, T. (2014). TMO - The Federated Ontology of the TrendMiner Project. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*

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.and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation,* 46(4), pp. 701-719.

Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10).*

Waltinger, U. (2010). Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features”. In *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10).*

Westerski, A. and Sánchez-Rada, J.F. (2013). Marl Ontology Specification, V1.0 May 2013. Available at http://www.gsi.dit.upm.es/ontologies/marl