Adding Polarity Information to Entries of the Database of Bavarian Dialects in Austria

Thierry Declerck[1], Amelie Dorn[2], Eveline Wandl-Vogt[2]
[1]Saarland University, [2]Austrian Center for Digital Humanities
e-mail: declerck@dfki.de, amelie.dorn@oeaw.ac.at, eveline.wandl-vogt@oeaw.ac.at

Abstract

In this short paper we describe an on-going work, which consists in adding polarity information to entries that are part of the Database of Bavarian Dialects in Austria (DBÖ). By "polarity information" we mean the positive or negative interpretation a word can carry. The starting point of our study is given by SentiMerge, a lexical resource that encodes polarity information for standard German words on the basis of integration processes performed on four pre-existing polarity lexicons. The lexical information of the entries of SentiMerge is encoded using the Ontolex model, which has been developed in the context of the W3C Ontology-Lexica Community Group. The polarity information, on the other hand, is encoded using the MARL ontological model, which has been developed at the Universidad Politécnica de Madrid. The use of such formal representation frameworks is aimed at supporting the publication of the resulting data in the Linked Open Data cloud. In the first phase of our work we focus on headwords describing colour terms, taking also compound words into consideration.

Keywords: polarity; DBÖ; colour terms; Ontolex; MARL

1 Introduction

We investigate the possibility of adding polarity information to entries of the Database of Bavarian Dialects in Austria (DBÖ). With "polarity information" we refer to the positive or negative interpretation a word may have. While this work can be partly achieved by analysing the textual content of examples or definitions associated to the entries in the DBÖ, it can also be achieved by establishing correspondences to existing polarity lexicons for standard German. One such lexicon is SentiMerge, described in Emerson & Declerck (2014).[1] SentiMerge is a lexical resource that encodes polarity information for German words on the basis of integration processes performed on four pre-existing polarity lexicons (Clematide & Klenner 2010; Remus et al. 2010; Waltinger 2010 and Klenner et al. 2012). The resulting merged and cleaned lexicon consists of 15,287 lemmas marked with either positive or negative polarity, indicated by real numbers (from -1.0 to 1.0, neutral polarity being marked by the value "0.0"), to which a confidence measure is also associated. There are 5 levels of confidence, low (3.536) and high (14.527), as well as three intermediate levels (5.823, 7.966 and 12.389). 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. In the last column of Table 1, the reader can see the confidence measure computed by the algorithms described in Emerson & Declerck (2014).

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

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 examples in Table 1 are compound words. Our interest lies in the possibility of marking elements of such compound words with polarity information and, in the longer term, the possibility of proposing an algorithm for computing the polarity of unknown compound words (i.e. words not listed in the SentiMerge lexicon) on the basis of the polarity of their elements, if those are included in the lexicon. For this study, there is thus the need to be able to encode elements of compound words, including their position in different compound words. Our choice here is the Ontolex model, which has been developed in the context of the W3C Ontology-Lexica Community Group.[2] For the representation of polarity information we opted for the MARL model[3], which has already been adopted for use in the context of sentiment lexicons published in the Linguistic Linked Open Data framework, as this has been described in details in Buitelaar et al. (2013).[4]

2 Context of our Study: The exploreAT! Project

The study we present in this paper is part of the Digital Humanities project "*exploreAT!* - exploring Austrian culture through the language glass"[5] carried out at the Austrian Centre for Digital Humanities.[6] It is based on the *Database of Bavarian dialects in Austria (DBÖ)* and the *Dictionary of Bavarian Dialects in Austria* (WBÖ). This large collection of 20th century dialect data is highly heterogeneous, and contains much information on dialectal word formations, as well as interesting cultural information relevant to the cultural heritage of the present-day Austria and the Austrian-Hungarian monarchy.

The data was originally collected by means of questionnaires, containing around 20,000 questions. The collection is estimated to contain 200,000 headwords in a set of about 4 million records. In this context, colour terminology is receiving particular attention as colours are an essential component of the vocabulary of almost all languages in the world (Berlin & Kay 1969). We concentrate thus on the extended semantic field of colour terms in Bavarian dialects, dealing also with various stages of the complex digital composition of non-standard linguistic data. At the same time, this rather local exploration of non-standard language material, combined with novel encoding methods in the field of (Linguistic) Linked Open Data enables the investigation and sustainability of lexicographic and digital humanities resources, supporting their linking to external dictionary data sets.

[2] Examples of encoding of German compound words in Ontolex are given in (Declerck 2016). See also http://www.w3.org/community/ontolex/wiki/Final_Model_Specification for the Ontolex model.

[3] See http://www.gsi.dit.upm.es/ontologies/marl

[4] In the Appendix of this paper we display graphical views of the two models, Ontolex and MARL.

[5] See (Wandl-Vogt et al. 2015).

[6] http://www.oeaw.ac.at/acdh/en/node/187 [last access: 2016.05.01] or http://exploreat.usal.es/ [last access: 2016.05.01].

3 Encoding of Polarity Lexicon in Ontolex and MARL

We present first the encoding of the SentiMerge entry “Rotsperre” (*red card suspension*, see Table 1) in Ontolex and MARL and show how this can easily be ported to the lexical data contained in DBÖ.

```
(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> .
```

Example (1) displays the Ontolex encoding for the compound word “Rotsperre”. We represent the fact that the entry consists of 2 components (:Rot_comp and :sperre_comp), which correspond to two entries in the generic German lexicon. Examples (2) and (3) show the encoding of these components and their linking to their corresponding lexical entry by the use of decomp:correspondsTo property:[7]

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

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

Now we can integrate the MARL vocabulary for marking the polarity of the compound word “Rotsperre” (see Table 1) and its components. As example (4) shows, we do this in the context of the Ontolex sense class.[8] Inclusion of MARL vocabulary is indicated by the use of the “op” prefix:

```
(4) :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
```

In a generic lexicon we see that the word “Sperre” has different meanings, one of them being in line with the sense of “Rotsperre”, sharing thus the same ontological reference: http://de.dbpedia.org/resource/Wettkampfsperre (*suspension from a competition*). The corresponding sense of the word “Sperre” is displayed in example (5):

```
(5) :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 ;
    op:polarityValue "-0.777"^^xsd:double ;
```

[7] See the Decomp module graphical representation in the Appendix.
[8] See the Ontolex graphical representation in the Appendix.

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

The adjective "rot" (*red*) in SentiMerge is marked as neutral (has polarity value "0.0"), and so we can see that the polarity value of the word "Rotsperre" is in the range of the combination of the polarity values of "red" and "Sperre2".

Looking now for example at the entry "(Stall)rôt:1" in DBÖ, which is not listed in SentiMerge. We first get the indication that the entry is a compound word, a fact indicated by the use of the parentheses. The first component is "Stall" (*stable*, a place for keeping animals, like cattle), and the second component "rôt" is a colour term (*red*). The components of the DBÖ entry are included in SentiMerge, and both are marked as being neutral. But the DBÖ definition of "(Stall)rôt:1" is stating that the word is about a *cattle disease* ("eine Krankheit des Hornviehs"). And diseases of animals (the entry: "Tierkrankheit") is marked with the value "-0.897" in SentiMerge. Therefore, similar to example (4), the DBÖ entry "(Stall)rôt:1" can be enriched with the information displayed in example (6), where we specify that the op:polarityValue is the one we get from the corresponding "Tierkrankeit" (*animal disease*) entry in SentiMerge, being "-0.897":

```
op:assessedBy :SentiMerge ;
op:hasPolarity op:Negative ;
op:maxPolarityValue "1.0"^^xsd:double ;
op:minPolarityValue "-1.0"^^xsd:double ;
op:polarityValue "“-0.897"^^xsd:double ;
```

An interesting fact is that the value of the polarity of the compound word does not correspond to the combination of the polarity values of its components, giving us a hint that the usage of the word is only indirectly or metaphorically related to a colour term. As a matter of fact, all terms pointing to a colour as such are marked as being neutral.

4 Conclusion

In this paper we presented an on-going work dealing with marking dialect data with polarity information which can be gained from both a specialised standard German polarity lexicon, and from the interpretation of definition of the entries of the dialect collection. We described how we encode all this integrated information using formal models for lexical data and for polarity information, supporting thus the future publication of the extended dialect dictionary in the Linked Data cloud.

5 References

Berlin, B. & Kay, P. (1969). *Basic color terms: their universality and evolution*. University of California Press.

Buitelaar, P., Arcan, M., Iglesias, C.A., Sánchez, J.F. & 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. &

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. (2016). Representation of Polarity Information of Elements of German Compound Words. In *Proceeding of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources,* Portorož, Slovenia.

Declerck, T. & 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 & 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, Ireland.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & 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. & 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. & 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, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012*). Interchanging lexical resources on the Semantic Web*. Language Resources and Evaluation, 46(4), pp. 701-719.

Remus, R., Quasthoff, U. & 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).*

Wandl-Vogt, E., Kieslinger, B., O´Connor, A. & Theron, R. (2015): „exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts“. In *Proceedings of the DHd-Tagung 2015*. Graz. Austria.

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).*

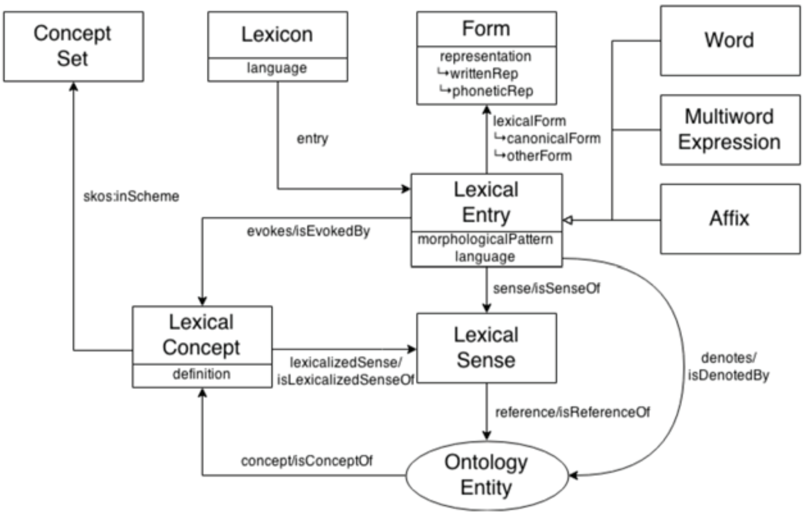
Westerski, A. & Sánchez-Rada, J.F. (2013). *Marl Ontology Specification, V1.0 May 2013.* Accessed at http://www.gsi.dit.upm.es/ontologies/marl [30/03/2016].

Acknowledgements

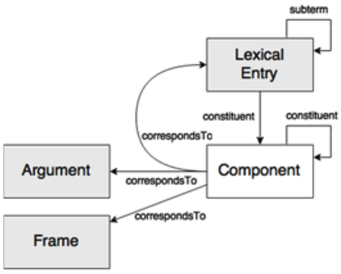
The work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233) and by the FREME H2020 project (grant No. 644771) and the “explore.AT!” project, funded by the Nationalstiftung für Forschung, Technologie und Entwicklung (Nationalstiftung FTE) at the Austrian Academy of Sciences.

Appendix

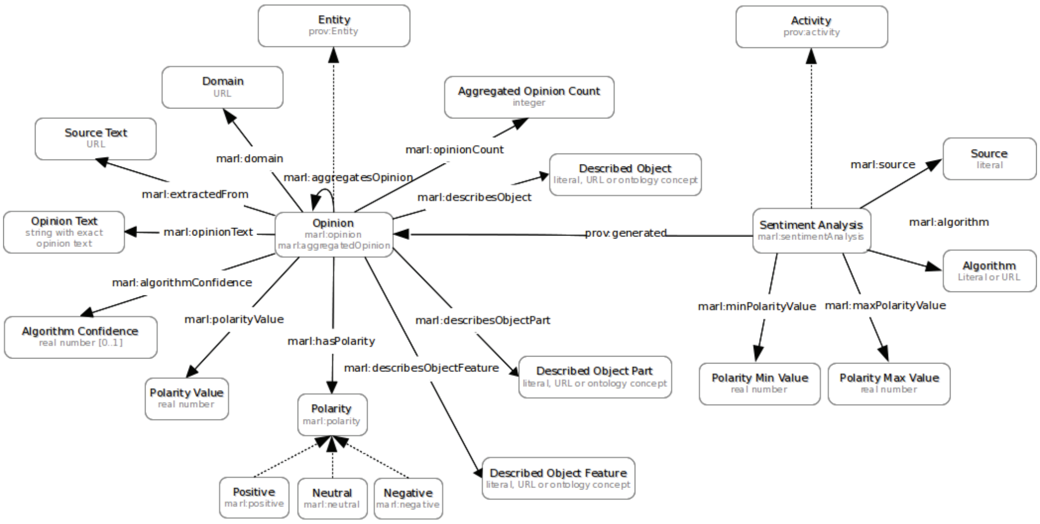
In this appendix we display the graphical representations of Ontolex, of the Decomp module attached to Ontolex and of MARL



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



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.



The MARL Model.