OntoLex 2017
1st workshop on the
OntoLex model

The W3C OntoLex Community Group was launched in 2011, with Paul Buitelaar (INSIGHT. National University of Ireland, Galway) and Philipp Cimiano (CITEC, Bielefeld University, Germany) as chairs, with the goal to define a model for allowing to represent lexical knowledge in connection to ontologies [1]. The rationale behind the model is that semantics are captured by ontologies, and the role of the lexicon-interface is to link lexical entries to ontological entities expressing their denotational meaning, following a principle called *Semantics by Reference*. Based on five years of intensive discussions and work, the new OntoLex-Lemon model was launched in May 2016 [2] and is becoming the primary method for representing linked lexical resources on the Web of Data, not only for capturing the lexicon-ontology interface but for the representation of lexicographic resources as well. The model facilitates bridging the gap between the NLP and data science communities by making available and linking large amounts of quality lexical information to the knowledge represented on the semantic web, for example in graphs such as DBpedia, applications

Issue 5 (T2). Semantic selection

Some dictionaries indicate the semantic features of the lexical items that an entry (in one of its senses) selects or even the exact lexical items with which it collocates. This is usually indicated either with a specific tag (e.g. KD's Range Of Application), or in-between parentheses at the beginning of a definition. Examples are, for instance, the dictionary entry for the German verb *dämmen*, which in its sense 'to insulate, absorb, mute' selects arguments that denote warmth or sound (German *Wärme*, *Schall*, etc.) (KD), the adjective *cozy*, meaning beneficial to all those involved and possibly somewhat corrupt if predicated from a transaction or an arrangement (*Google Dictionary*); or the collocational measure words of *luck*: stroke, piece of (*Oxford Collocations Dictionary*). The OntoLex Syntax-Semantics Module (*synsem*) class `synsem:OntoMap` allows to map a syntactic frame to an ontology entity, so that the frame and its arguments are linked to the ontology elements that they lexicalize. Even though dictionaries commonly include information on subcategorization (transitive/intransitive/reflexive etc. annotations for verbs, for instance), details on the syntactic frame are not always provided beyond those annotations. Since in dictionary conversion we often lack a given ontology and detailed syntactic information is not provided, the mapping between syntactic arguments and ontology entities seems difficult to establish automatically via `synsem:OntoMap`: how do we automatically represent that the adjective *cozy* has a meaning only applied to transaction or agreement or that the measure words that collocate with *luck* are *stroke* or *piece* if the morphosyntactic information provided in the dictionary is just that *cozy* is an adjective and *luck* a noun? Furthermore, `synsem:condition` (in its turn subsuming `synsem:propertyRange` and `synsem:propertyDomain`) enables us to state constraints on the arguments of a predicate in a given ontology.[14] The possibility of reusing it to state the constraints on syntactic arguments even in cases in which we lack a given ontology and therefore are not mapping to given ontology properties has to be further analyzed. In addition, the potential links between the modeled entries (e.g. *piece* and *luck*), i.e. the links at the lexical level, are also to be considered, for instance, by taking into account recent proposals on the representation of lexical functions as LLD (Fonseca et al. 2016).

4. A module for lexicography

The previous section dealt with some of the issues we encountered in our work with dictionaries and the potential ones that may rise with other lexicographic works that have not been migrated to LLD yet. In the following we draft a potential solution which can serve as a basis for a new module in OntoLex specifically developed for the representation of dictionaries after thorough revision and improvement according to the community's feedback.

In order to keep track of the dictionary representation and prevent any loss of information mentioned in Issue 1, related to the splitting of dictionary entries in several lexical entries, we propose a new class `DictionaryEntry`. This new class would both enable to group together lexical entries as well as to associate any information shared by all of them. In our view, we distinguish lexical entries and lexicons (as containers of lexical entries), from the original dictionary entry (the new class `DictionaryEntry`) and the original dictionary resource (`Dictionary`), which would serve in turn to record the provenance of each dictionary entry. Mirroring the `lime:Lexicon-ontolex:LexicalEntry` relation we suggest a `Dictionary-DictionaryEntry` one. Any lexical entry created during the conversion to LLD but not originally provided in the resource would then belong to a `lime:Lexicon`, but not to the instance of `Dictionary` representing that resource. A `lime:Lexicon` in English, for example, could aggregate lexical entries created on the fly by the LLD expert or original ones coming from as many English dictionaries as desired. These dictionaries can in turn differ in their modeling and their views on the data, their criteria of sense ordering or their structure.

Regarding Issue 2, the `DictionaryEntry` class would allow to divide a single lexical entry into several ones if desired, each with a different preferred form, while maintaining the original dictionary representation. If the dictionary entry is not split, the option of linking a sense to a grammatical restriction on gender or number from an external catalog would solve the issue, although the implications of this solution (its benefits and drawbacks) will need further analysis.

In order to represent usage examples and their translations (Issue 3) we propose to go back to `lemon:UsageExample` and link it to a `LexicalSense`. A new class `ExampleCluster` would link to `UsageExamples` that are translations from each other. The use of the vartrans

[14] http://w3.org/2016/05/ontolex/#conditions

module to model translations among senses would imply the creation of lexical senses for each example, and therefore treating the example as a lexical entry, which we deem is beyond the definition of lexical entries.

Issue 4 was concerned with the order of senses in a dictionary entry and the order of homographs in the macrostructure of the dictionary. There are different possible approaches to resolve this: reusing already available RDF mechanisms, reifying the sense order in a new class `SenseOrder`, or defining a new property `senseOrder` attached to the lexical sense. The first option involves the reuse of `rdfs:Container`[s] to declare with e.g. `rdf:_1, rdf:_2` that a particular sense is the first or the second one. However, cases in which a set of senses allows for various orderings, depending on the ordering criterion, or in which some senses come from different dictionaries (each with its order), should also be accounted for. The second option suggests that the sense order is reified in a class `SenseOrder` linked to the lexical sense. This class would enable us to record the position of that sense, its provenance (presumably an instance of the class `Dictionary`), and, if desired, the ordering criterion. If repeated senses were identified (e.g. senses that share a definition in both dictionaries), `SenseOrder` would allow us to have one single lexical sense with two different positions according to the two different orderings and dictionaries, in a similar fashion as two containers with two different sequences of senses. Alternatively, if we assume that a lexical sense always comes from just one dictionary source, a property `senseOrder` would suffice.

Issue 5, dealing with semantic selection, has been brought up for further discussion in this paper to see whether it could be covered by synsem module mechanisms or whether it would require new entities in the context of the lexicography module. As part of the conversion of the KD's *Global Spanish Multilingual Dictionary* (Bosque-Gil et al. 2016), the semantic selection information provided by KD's tag RangeOfApplication was captured by the use of `synsem:condition`. In that approach, `synsem:condition` would link a lexical sense to a blank node[15] with an `rdf:value` recording the strings given as arguments in the original data. This modeling allowed us to deal with the lack of a given ontology and detailed information on the syntactic frames of lexical entries for each of their senses. Thus, the focus was set on representing the data just as it was in its original format while being compliant with the OntoLex formal specification and reusing its elements as much as possible. We argue that the lexicography module should aim to set the basis to exploit at the dictionary's macro-structure level the potential benefits of establishing semantic relations among lexical senses based on lexical selection or among syntactic frames and arguments and the ontology entities that they denote. To this aim, overcoming the lack of detailed syntactic information in the dictionary as well as the lack of a given ontology to lexicalize becomes essential.

5. Conclusion

OntoLex is increasingly being used to convert linguistic resources to LLD outside the scope of ontology lexicalization. In this position statement we have drawn attention to a series of issues raised in the literature on LLD related to the conversion of dictionaries to LD and to five of the ones we came across in the same line of work and after a later analysis of several additional dictionaries. We argue that the OntoLex model should enable the preservation of the content and the structure of the original resource, even if the LLD expert opts for a different representation that is better suited to the data exploitation by external applications or is more in line with his or her view on the lexicon-ontology interface. We have outlined some of our insights on how to address these issues in a new module for lexicography. It would be compatible with the mechanisms suggested in the state-of-the-art on dictionaries represented as LLD, as of the moment of writing, and also with other potential modules for the encoding of specific lexical aspects (e.g. etymology). The final module is intended to be dictionary-agnostic in the sense that it should be applicable (and combined with other modules if necessary) to different kinds of dictionaries (e.g., general, collocations, learner's, etymological, historical, etc.). This would bring linked data (LD) closer to lexicography not only with the aim of leveraging already available dictionaries in LD for NLP tasks, but also for introducing LD in the work carried out in that discipline.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R), the Excellence Network ReTeLe (TIN2015-68955-REDT), and the Juan de la Cierva program and by the Spanish

as part of the Global WordNet Association Collaborative Interlingual Index, or existing resources such as BabelNet, DBnary and commercial dictionaries.

OntoLex 2017 was co-located with the Language, Data and Knowledge conference [3, and see p.13] and presented the first opportunity for practitioners to meet to discuss the model, its applications and future development [4]. The participants have shown interest in continuing the development of OntoLex-Lemon, particularly with regard to lexicographic resources. In consequence, the group is starting to work on a new best practice document that will provide modeling examples and guidelines for how to use OntoLex specifically to represent lexicographic resources such as dictionaries. Then it will be decided whether to stipulate the proposed vocabulary elements or modeling solutions into the status of a new module or to keep the best practices proposal as an informal document.

Philipp Cimiano
Universität Bielefeld

[1] https://w3.org/community/ontolex/
[2[https://w3.org/2016/05/ontolex/
[3] http://ldk2017.org/
[4] http://ontolex2017.linguistic-lod.org/

[15] `synsem:condition` has `rdfs:Resource` defined as its range.