LREC 2016 Workshop

LDL 2016

5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources

PROCEEDINGS

Edited by

John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, Sebastian Hellmann

24 May 2016

Proceedings of the LREC 2016 Workshop
"LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources"

24 May 2016 – Portorož, Slovenia

Edited by John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, Sebastian Hellmann

Organising Committee

- John P. McCrae, National University of Ireland, Galway, Ireland*
- Christian Chiarcos, Goethe University Frankfurt, Germany*
- Elena Montiel Ponsoda, Universidad Politécnica de Madrid, Spain*
- Thierry Declerck, Saarland University, Germany*
- Petya Osenova, IICT-BAS, Bulgaria*
- Sebastian Hellmann, AKSW/KILT, Universität Leipzig, Germany*
- Julia Bosque-Gil, Universidad Politécnica de Madrid, Spain
- Bettina Klimek, AKSW/KILT, Universität Leipzig, Germany

*: Main editors and chairs of the Organising Committee

Programme Committee

- Guadalupe Aguado, Universidad Politécnica de Madrid, Spain
- Núria Bel, Universitat Pompeu Fabra, Spain
- Claire Bonial, University of Colorado at Boulder, USA
- Paul Buitelaar, National University of Ireland, Galway, Ireland
- Steve Cassidy, Macquarie University, Australia
- Nicoletta Calzolari, ILC-CNR, Italy
- Damir Cavar, Indiana University, USA
- Philipp Cimiano, Bielefeld University, Germany
- Gerard de Melo, Tsinghua University, China
- Alexis Dimitriadis, Universiteit Utrecht, The Netherlands
- Judith Eckle-Kohler, Technische Universität Darmstadt, Germany
- Francesca Frontini, ILC-CNR, Italy
- Jeff Good, University at Buffalo, USA
- Asunción Gómez Pérez, Universidad Politécnica de Madrid, Spain
- Jorge Gracia, Universidad Politécnica de Madrid, Spain
- Yoshihiko Hayashi, Waseda University, Japan
- Nancy Ide, Vassar College, USA
- Fahad Khan, ILC-CNR, Italy
- Vanessa Lopez, IBM Europe, Ireland
- Steven Moran, Universität Zürich, Switzerland and Ludwig Maximilian University, Germany
- Roberto Navigli, University of Rome "La Sapienza", Italy
- Sebastian Nordhof, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
- Antonio Pareja-Lora, Universidad Complutense Madrid, Spain
- Maciej Piasecki, Wroclaw University of Technology, Poland

- Francesca Quattri, Hong Kong Polytechnic University, Hong Kong
- Mariano Rico, Universidad Politécnica de Madrid, Spain
- Laurent Romary, INRIA, France
- Felix Sasaki, Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
- Andrea Schalley, Griffith University, Australia
- Gilles Sérraset, Joseph Fourier University, France
- Kiril Simov, Bulgarian Academy of Sciences, Sofia, Bulgaria
- Milena Slavcheva, JRC-Brussels, Belgium
- Aitor Soroa, University of the Basque Country, Spain
- Armando Stellato, University of Rome "Tor Vergata", Italy
- Cristina Vertan, University of Hamburg, Germany
- Piek Vossen, Vrije Universiteit Amsterdam, The Netherlands

Preface

Since its establishment in 2012, the Linked Data in Linguistics (LDL) workshop series has become the major forum for presenting, discussing and disseminating technologies, vocabularies, resources and experiences regarding the application of **Semantic Web standards and the Linked Open Data paradigm to language resources** in order to facilitate their visibility, accessibility, interoperability, reusability, enrichment, combined evaluation and integration. The Linked Data in Linguistics workshop series is organized by the Open Linguistics Working Group of the Open Knowledge Foundation, and has contributed greatly to the development of the Linguistic Linked Open Data (LLOD) cloud. This workshop builds on the existing success of previous instances of this workshop over the last four years, firstly at the 34th Annual Conference of the German Linguistics Society (DGfS) in 2012, followed by a second appearance at the 6th Annual Conference on Generative Approaches to the Lexicon (GLCON). In 2014, the workshop was held at the previous edition of LREC in Reykjavik, where we attracted a very large number of interested participants. Last year, the workshop was co-located with ACL-IJCNLP 2015 in Beijing, China.

Publishing language resources under open licenses and linking them together has been an area of increasing interest in academic circles, including applied linguistics, lexicography, natural language processing and information technology, and to facilitate exchange of knowledge and information across boundaries between disciplines as well as between academia and the IT business. By collocating the 5th edition of the workshop series with LREC, we encourage this interdisciplinary community **to present and to discuss use cases, experiences, best practices, recommendations and technologies** among each other and in interaction with the language resource community. We particularly invite contributions discussing the application of the Linked Open Data paradigm to linguistic data as it might provide an important step towards making linguistic data: i) easily and uniformly **queryable**, ii) **interoperable** and iii) **sharable** over the Web using open standards such as the HTTP protocol and the RDF data model. While it has been shown that linked data has significant value for the management of language resources in the Web, the practice is still far from being an accepted standard in the community. Thus it is important that we continue to push the development and adoption of linked data technologies among creators of language resources. In particular linked datas ability to increase the **quality**, **interoperability** and **availability** of data on the Web has led to us focus on **managing, improving and using language resources** on the Web as a key focus for this years workshop.

John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, Sebastian Hellmann. May 2016.

Programme

Opening Session

09.00 – 09.30 Welcome and introduction

09.30 – 10.30 Invited talk by Damir Cavar (Indiana University)
On the role of Linked Open Language Data for Language Documentation, Linguistic Research, and Speech and Language Technologies

10.30 – 11.00 `Coffee break`

First Oral Session

11.00 – 11.30 Vladimir Alexiev and Gerard Casamayor
FN goes NIF: Integrating FrameNet in the NLP Interchange Format

11.30 – 12.00 Frank Abromeit, Christian Chiarcos, Christian Fäth and Maxim Ionov
Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF

12.00 – 12.20 Jim O Regan, Kevin Scannell and Elaine Uí Dhonnchadha
lemonGAWN: WordNet Gaeilge as Linked Data

12.20 – 12.40 Vít Baisa, Sara Može and Irene Renau
Linking Verb Pattern Dictionaries of English and Spanish

12.40 – 14.00 `Lunch break`

Poster Session

14.00 – 14.40 Fahad Khan, Javier Díaz-Vera and Monica Monachini
The Representation of an Old English Emotion Lexicon as Linked Open Data
Elena Gonzalez-Blanco, Gimena Del Rio Riande and Clara Martínez Cantón
Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires
Sotiris Karampatakis, Sofia Karampataki, Charalampos Bratsas and Ioannis Antoniou
Linked Open Lexical Resources for the Greek Language
Invited posters from related projects and initiatives

Second Oral Session

14.40 – 15.10 Petya Osenova and Kiril Simov
Linked Open Data Dataset from Related Documents

15.10 – 15.40 Andrea Salfinger, Caroline Salfinger, Birgit Pröll, Werner Retschitzegger and Wieland Schwinger, *Pinpointing the Eye of the Hurricane - Creating a Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data*

15.40 – 16.00 Thierry Declerck
Representation of Polarity Information of Elements of German Compound Words

16.00 – 16.30 `Tea break`

Third Oral Session

16.30 – 16.50 Vanya Dimitrova, Christian Fäth, Christian Chiarcos, Heike Renner-Westermann and Frank Abromeit, *Building an ontological model of the BLL Thesaurus: First step towards an interface with the LLOD cloud*

16.50 – 17.10 Claus Zinn and Thorsten Trippel
Enhancing the Quality of Metadata by using Authority Control

17.10 – 17.30 Christian Chiarcos, Christian Fäth and Maria Sukhareva
Developing and Using Ontologies of Linguistic Annotation

Closing Session

17.30 – 18.00 Wrap up

Table of Contents

Posters

FN goes NIF: Integrating FrameNet in the NLP Interchange Format

Vladimir Alexiev, Gerard Casamayor

Ontotext Corp, Universitat Pompeu Fabra
vladimir.alexiev@ontotext.com, gerard.casamayor@upf.edu

Abstract

FrameNet (FN) is a large-scale lexical database for English developed at ICSI Berkeley that describes word senses in terms of frame semantics. FN has been converted to RDF LOD by ISTC-CNR, together with a large corpus of text annotated with FN. NIF is an RDF/OWL format and protocol for exchanging text annotations between NLP tools as Linguistic Linked Data. This paper reviews the FN-LOD representation, compares it to NIF, and describes a simple way to integrate FN in NIF, which does not use any custom classes or properties.

Keywords: FrameNet, NLP Interchange Format, NIF, Linguistic LOD

1. Introduction

FrameNet (FN) [9] is a large-scale linguistic resource for English developed at ICSI UC Berkeley. It documents the syntactic and semantic combinations (valences) of predicative word senses in terms of frames, lexical units, frame elements, and relations between them. More precisely:

- **Frames** are conceptual situations along with their participants (e.g. `frame:Statement` corresponds to an event in which a statement is made)
- **Lexical Units** (LU) are phrases or words that evoke frames (e.g. `lu:announce.v` and `lu:declare.v` both evoke `frame:Statement`)
- **Frame Elements** (FE) are roles taken by participants in a frame: things, entities, times, places, etc (e.g. `fe:Speaker.statement`, `fe:Message.statement`)

The FN lexical database also comprises a corpus of annotated sentences that exemplify all the above. FN has been converted to Linked Open Data (LOD) by ISTC-CNR [7], a conversion henceforth refered to as FN-LOD. This conversion covers not only FrameNet's lexical database but also FrameNet's corpus of sentences annotated with frames, FEs and other linguistic information.

The NLP Interchange Format (NIF) [3] is a set of ontologies, specifications and software to enable the exchange of linguistic annotations as RDF/OWL between Natural Language Processing (NLP) tools. The NIF model includes a core ontology to represent textual annotations and binding to text, and reuses NLP vocabularies, such as: ITS and NERD for Named Entity Recognition (NER) (individuals and classes respectively), OLIA for modeling model tagsets produced by various types of NLP tools, MARL for sentiment/opinion, etc.

In the last years NIF has gained wide-spread adoption in the Linguistic LD community, with a variety of linguistic corpora being published as NIF. For example, the Manually Annotated Sub-Corpus (MASC) [8] has been published as LOD NIF with additional links to linguistic resources in two recent efforts [5][10].

See [1] for a brief overview of Linguistic LD and related ontologies. An extensive bibliography is available on Zotero.

We are not aware of any alignment or example of using FN-LOD and NIF together. While [4] describes plans to interlink FN-LOD and MASC as LOD, neither it nor the two MASC LOD datasets cited above include FN-LOD.

This paper reviews the FN-LOD representation, compares it to NIF, and describes a simple way to integrate FN in NIF so that FrameNet-based annotations can be produced and consumed by NIF-compliant services. Crucially, this integration is achieved without resorting to any custom vocabulary (no new classes or properties). Instead, we align the core items of NIF and FN-LOD.

This FN-NIF integration is an important step towards bulding NIF-compliant pipelines of text analysis and Information Extraction (IE) components capable of producing LOD corpora with rich linguistic and semantic annotations. Such corpora is important for a wide range of tasks ranging from corpora analysis in linguistic research, to downstream applications like semantic indexing and summarization. The FN-NIF model presented here is used in the Multisensor project (MS) [6][11], which applies semantic technologies to the analysis of multimedia (including news articles and social media) and where NIF has been adopted as the data model for data exchange between text processing components. More precisely, the FN-NIF integration is used to encode the output of a relation extraction implementation that produces annotations of FrameNet-based n-ary relations. By using NIF to store the extracted relations as annotations, it is possible to integrate them with annotations produced by other text analysis services. Thus, for instance, relations in Multisensor can have as arguments entities annotated by NER and concept extraction modules.

The rest of this paper is structured as follows. First, we introduce an example sentence which we will use through the paper to illustrate discussions. Then we describe FN-LOD in detail and compare it to NIF. The FN-NIF model is presented, followed by sample queries to get information out of it. **Accompanying materials** are available for download, including Orgmode source (`org`) and local files referenced in the paper as relative links `./*`: Turtle RDF (`ttl`), ontologies in Manchester Notation (`omn`), bigger figures in

PlantUML (`puml`) and `png`.
Through this paper we'll use the following sentence to illustrate discussions:

> Electrolux announced today the theme for its design competition.

1.1. SEMAFOR

Some softwares are available for automatic FN annotation. We used SEMAFOR [2] to annotate the sample sentence.
SEMAFOR uses a dependency parse (shown on top of Fig 1) to generate candidate frames for the sentence (shown at the bottom). Here we have highlighted the `Statement` frame, invoked by `lu:announce.v` and having FEs `Speaker`, `Time` and `Message`. The other candidate frames are dimmed out.
It may be easier to see the candidate frames in SEMAFOR's vertical layout (Fig 2). Here each column represents a frame.

1.2. SEMAFOR Candidate Frame Filtering

SEMAFOR offers a JSON format (./SEMAFOR.json) where one can see the candidate frames and their targets (LUs) and FEs. It includes a `score` for each frame, which can help us pick the best frames:

Frame	Score
Statement	113.2
Competition	54.6
Coming_up_with	50.7
Calendric_unit	30.4
Topic	25.4

In this case the two top-scoring candidates (`Statement` and `Competition`) are the best frames. `Calendric_unit` is too small (equal to `lu:Time.statement`), `Coming_up_with` is wrong, and `Topic` is part of `Statement`.
We propose a simple approach to filter candidate frames based on score and a dependency tree structure (see Fig 6):

- Order candidate frames by decreasing score
- Repeat:
  - Add the highest scoring frame *f*
  - Discard any frames that are governed by *f* in the dependency tree

2. FN-LOD Ontologies

Major impediments to real world uses of FN-LOD include the complexity of the involved ontologies, the fact that there are two to choose from (see sec 2.2. and sec 2.3.), the lack of an overall picture of how classes and properties fit together, and the lack of adequate documentation for some ontology elements.
The OWL ontology representation of FN-LOD is described in [7], but it is necessary to be familiar with the documentation of the FrameNet project [9] in order to understand the ontologies. While there is a partial ontology diagram in [7], it doesn't show all classes and relations. Some elements are commented extensively using texts from the FN Book [9], but we found these texts more understandable when reading them in the book, since the comments do not capture the context. Many elements are not documented, e.g. class `fn:Header`, data property `fn:frame_cBy` (xsd:string), etc. One can only surmise that `fn:frame_cBy` is the ID of the person who created the frame.
In order to understand the FN-LOD ontologies, we diagrammed classes and properties. Sample data (see sec 2.5.) played a crucial role in building this understanding. Since the data is very large, we had to extract smaller connected fragments to be able to understand them. In this section we describe the available FN-LOD ontologies and RDF data files, provide diagrams to facilitate understanding, and derived files that are easier to consume.

2.1. Prefixes

FN-LOD uses the following prefixes, which we registered in prefix.cc, an online prefix registry:

prefix	description
fn:	FN metamodel (tbox)
frame:	frame
fe:	frame element
lu:	lexical unit
st:	semantic type

2.2. fntbox ontology

The *FN terminology box* fntbox is the FN-LOD metamodel. It's an OWL ontology that uses Restrictions extensively, and is easiest to understand in Manchester notation (OMN): ./fntbox.omn.It has 16 Classes, 67 ObjectProperties, 49 DataProperties. Online documentation (OWLDoc) is available.
Most relations have inverses, but the PROV ontology designers have concluded that inverses actually harm interoperability by exerting a higher reasoning or querying cost:

> *When all inverses are defined for all properties, modelers may choose from two logically equivalent properties when making each assertion. Although the two options may be logically equivalent, developers consuming the assertions may need to exert extra effort to handle both (e.g., by either adding an OWL reasoner or writing code and queries to handle both cases). This extra effort can be reduced by preferring one inverse over another.*

We agree with them and recommend to use exactly the FN-LOD properties shown in Fig 5, and **not** their inverses.
Inverses also hinder understanding the *data hierarchy* implied by the ontology. To aid understanding, we made a diagram (Fig 3) (./fntbox.png, source ./fntbox.puml) showing all classes, their relations (object properties) and fields (data properties). For some properties we had to figure out the range from Restrictions; properties having a Union as domain are shown several times on the diagram.
To understand **fntbox** consider the classes in two groups and navigate top-down.
First are classes that represent texts and their annotation with frame instances and other linguistic info:

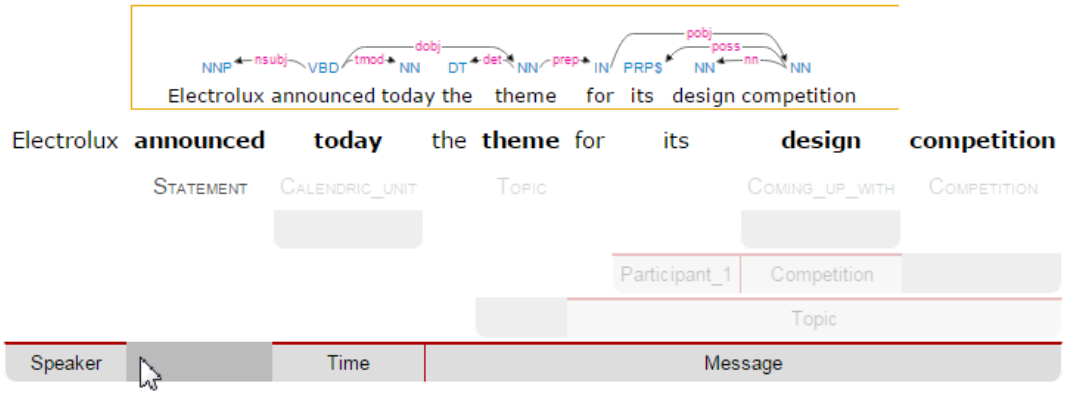


Figure 1: SEMAFOR Output, Horizontal Layout

	Statement	Calendric_unit	Topic	Coming_up_with	Competition
Electrolux	Speaker				
announced	Statement				
today	Time	Calendric_unit			
the	Message				
theme			Topic		
for			Topic		
its					Participant_1
design				Coming_up_with	Competition
competition					Competition

Figure 2: SEMAFOR Output, Vertical Layout

- `Header` holds together all `FullTextAnnotation` and `CorpDoc` about the same frame.
- `FullTextAnnotation` represents a mode of annotation where sentences are "preselected" by a given text.
- `CorpDoc` is a corpus comprising of documents and sentences that are carefully chosen by lexicographers to illustrate the possible valences of LUs, i.e. make various frames for each sense of each LU.
- `Sentence` holds the `text` being annotated and some identifying information.
- `AnnotationSet` is a set of annotations about one frame. One sentence may have several frames and they may even overlap.
- `Layer` is a subset of annotations with a single purpose, indicated in `fn:layer_name`. Often used ones:
  - **Target**: LU that is target of the frame. Such layer has a single label.
  - **FE**: frame elements
  - **PENN**: part of speech (e.g. VBD, VVN, dt, nn)
  - **PT**: phrase type (e.g. NP, AJP, PP, PPing)
  - **GF**: grammatical function (e.g. Ext, Obj, Dep, Comp)
  - **NER**: named entity recognition (e.g. person, location)
- `Label` is a word or phrase in an annotated `Sentence` (indicated by index `label_start, label_end`) that:
  - Plays the role of LU instance. This is indicated by `fn:label_name` being "Target", and it's the single `Label` in a layer having the same `fn:layer_name`
  - Or plays the role of FE instance. In this case `fn:label_FE` points to the FE definition (e.g. `fe:Speaker.statement`) and `fn:label_name` corresponds (e.g. "Speaker"),
  - Or carries a grammatical or POS tag in `label_name`,

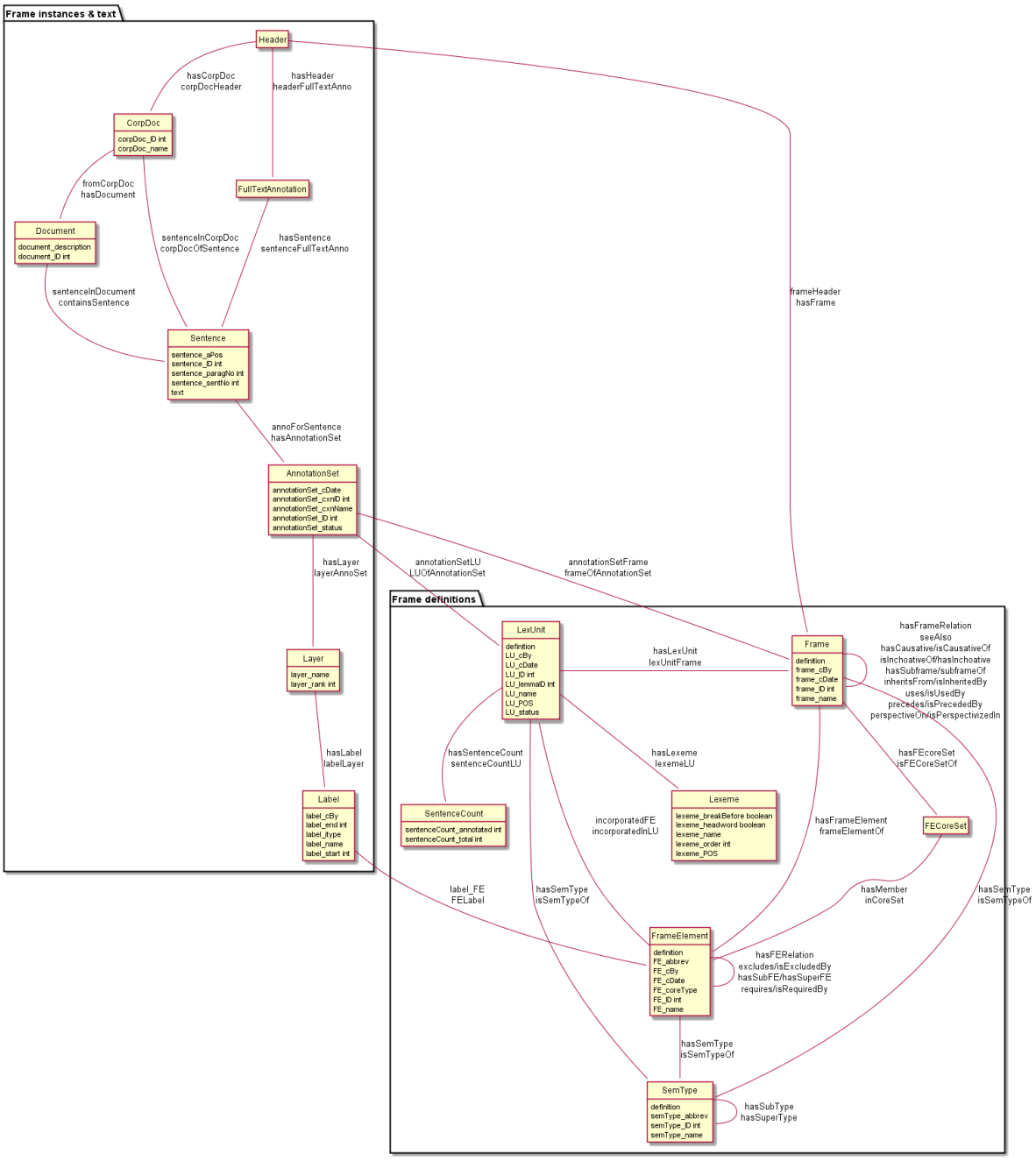


Figure 3: fntbox Ontology

– Or indicates a lexically omitted FE (see [9] sec 3.2.3 Null instantiation) using `fn:label_itype` (e.g. "CNI", "DNI", etc), in which case `label_start, label_end` are omitted.

Then are frame definition classes:

- `Frame` is a structure that abstracts over real-world situations, obtained through linguistic attestation
- `LexUnit` is the head-word of a sentence or subsentence that invokes the frame. An important goal of the FN project is to capture the meaning of words through annotated examples, that's why the LU can point to an `AnnotationSet` that supports it. It can also carry simple statistics (`SentenceCount`) used for managing the work of annotators.
- `Lexeme` is the linguistic representation of a LU. One LU can have several lexemes.
- `FrameElement` are entities (things, actors, times, messages, etc) that participate in a frame. They

are classified with `FE_coreType` into Core-Unexpressed, Extra-Thematic, Peripheral.

- `FECoreSet` describes a set of alternative FEs, one of which must be present in the frame. A frame can have several core sets.
- `SemType` classifies frames, FEs and LUs by type. E.g. some sem types are:
  - for Frame: `Non-perspectivalized_frame`, `Non-Lexical_Frame`
  - for FE: `Sentient` (an agent), `Artifact`, `Message`, `State_of_affairs`

2.3. framenet ontology

The framenet ontology is an alternative version of **fntbox**. It is significantly more complex: 33 Classes, 71 ObjectProperties, 23 DataProperties, and 18 Individuals. We converted it to Manchester notation (./framenet.omn) and made two diagrams:

- ./img/framenet.png (source ./framenet.puml). This is nearly unreadable, showing the complexity of this ontology
- ./img/framenet-nolabel.png (source ./framenet-nolabel.puml), which elides edge labels to avoid clutter, but is still too complex to show here.

This ontology perhaps corresponds better to what is described in the FN Book [9], but since it is not used in the RDF files described below, we do not give it further consideration.

2.4. fnabox ontology

The FN-LOD *assertion box* ontology fnabox is an RDF representation of all frame definitions. It includes only individuals, not classes nor property definitions. It used some illegal URI chars (spaces and parentheses) that we converted to underscores (e.g. transformed `lu:swing_(into).v` to `lu:swing__into_.v`). Then we converted it to ./fnabox.ttl, which is more readable: all individuals are sorted by name and all statements about an individual are together. For instance, the triples for `frame:Statement` include:

```
frame:Statement
  fn:hasFrameElement fe:Time.statement,
    fe:Iteration.statement... ;
  fn:hasLexUnit lu:gloat.v, lu:explain.v,
    lu:declaration.n, lu:talk.v... ;
  fn:isInheritedBy frame:Telling,
    frame:Reveal_secret, frame:Recording... ;
  fn:isUsedBy frame:Unattributed_information,
    frame:Adducing... ;
  fn:uses frame:Communication .
```

And these are the triples for a couple of the core FEs in that frame:

```
fe:Speaker.statement a fn:FrameElement ;
  fn:hasSemType st:Sentient ;
  fn:hasSuperFE fe:Speaker.speak_on_topic...
fe:Message.statement a fn:FrameElement ;
  fn:hasSemType st:Message ;
  fn:hasSuperFE fe:Message.encoding,
    fe:Message.communication...
```

2.5. fndata

fndata_v5 is a corpus of FrameNet annotations provided in RDF by ISTC-CNR, consisting of 540Mb of RDF/XML (292Mb Turtle, 1.03Gb NTriples) and comprising 3.8M triples. It includes 5946 sentences and 20361 frame instances (`annotationSetFrame`), i.e. 3.4 frames per sentence. The info about each sentence takes 640 triples on average; about a quarter of these are pure frame instance info (45 triples per frame).

We extracted all triples about *iran_missile_fullTextAnnotation_sentence_52* into ./iran_missile_sentence_52.ttl. This, for instance, is sentence 3 of paragraph 10 of a fullTextAnnotation corpus named "iran_missile":

This project was focused on the development of a longer ranged (150-200 km) and more heavily armed version of the Israeli Gabriel anti-ship missile (not as sometimes reported with the development of a ballistic missile based upon Israeli Jericho surface-to-surface missile technology)

.

Extracting the triples was fairly trivial since the URLs of nodes in these triples share the same base. The resulting set of triples for the above sentence played a crucial role in allowing us to understand the structure of FN-LOD data and the meaning of most fields (see Fig 3 and field descriptions above). It includes 6 manually annotated frames: *Gizmo*, Bearing_arms, Cause_to_make_progress (twice), Project and Type. SEMAFOR reports these frames and a number of smaller frames (often consisting of a single word): Artifact, Cardinal_numbers, Degree, Duration_attribute, Frequency, Increment, Part_inner_outer, Place_weight_on, Range, Statement, Vehicle and Weapon. While *Gizmo* is invoked by this phrase: "*surface-to-surface missile **technology***", it is not recognized by SEMAFOR, as it may have an older set of frame definitions.

3. Comparing FN-LOD to NIF

Since our goal is to integrate FN-LOD to NIF, we'll start with a comparison between the two. Compare **fntbox** (Fig 3) to the NIF class and property diagram (Fig 4).

3.1. Text Framing

The document is the basic level at which there is correspondence between FN-LOD and NIF: `fn:Document` and `nif:Context`. The text is stored in `fn:text`, respectively `nif:isString`.

At the level above document, FN-LOD has `fn:CorpDoc` or `fn:FullTextAnnotation` (two kinds of corpora). NIF uses `nif:Context` for this, using `nif:broaderContext` to point to higher-level contexts (but we are not aware of NIF data actually using this pattern).

Below document, `fn:Sentence` is the basic FN-LOD level to which frames are attached. Then follow `fn:AnnotationSet, fn:Layer, fn:Label`.

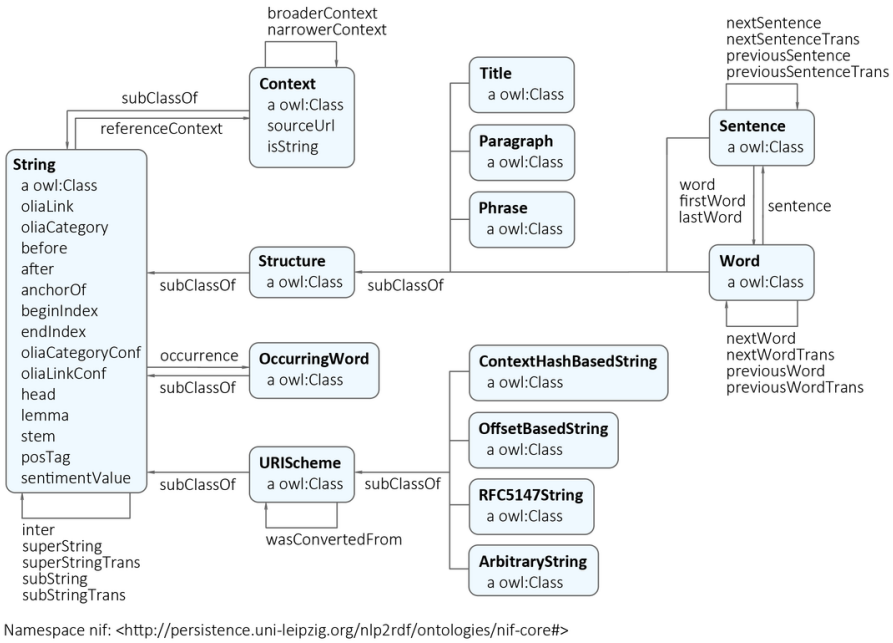


Figure 4: NIF Core Ontology

Char offsets are attached to `fn:Label`: `fn:label_start`, `fn:label_end`. NIF uses a generic class `nif:Structure` with subclasses `Paragraph`, `Sentence`, `Phrase`, `Word`, etc. Char offsets are specified at each level (`nif:beginIndex`, `nif:endIndex`). One can also provide the text at this level (`nif:anchorOf`), though this is redundant because `referenceContext/isString` is mandatory and contains the full text.

3.2. Text Links

Every NIF string (`Paragraph`, `Sentence`, `Phrase`, `Word` etc) must point to the enclosing context (`nif:referenceContext`). NIF has property `nif:subString` (and inverse `nif:superString`) that can be used to point uniformly from higher level texts to lower level texts (e.g. from Paragraph to Sentence to Phrase to Word). However it is not often used. There is also a specialized property `nif:word` (inverse `nif:sentence`) that points from a sentence down to its words; but it is not declared as specialization of `nif:subString`. One can also make chains of sentences (`nif:previousSentence, nif:nextSentence`) and words (`nif:previousWord, nif:nextWord`), and point to the first/last word of a sentence.

In contrast, FN-LOD has non-uniform treatment of links: to navigate from `Sentence` to its strings (`Label`), one has to follow the property path `sentenceInDocument/annoForSentence/ hasLayer/hasLabel`.

3.3. Text Nodes

FN-LOD doesn't recommend any convention for the URLs of text nodes, but you can see a pattern in sec 2.5.. E.g. `iran_missile_fullTextAnnotation_sentence_52_annotationSet_6_layer_2_label_0` is the URL of label 0 in layer 2 in set 6 of sentence_52 (which is actually sentence 3 of paragraph 10 of the fullTextAnnotation corpus. Note: labels, layers and sets use only even numbers in this representation). This label represents the phrase *surface-to-surface missile* (from offset 282 to 253) representing `fe:Use.gizmo` of `frame:Gizmo`. This convention makes labels **relative** to annotation sets (frame instances), and indeed this is borne out by the fntbox class diagram (sec 2.2.).

In contrast, NIF strongly recommends adopting a URL scheme that is based on character offsets and is thus **global** within the document (`nif:Context`). The class `nif:RFC5147String` provides such a scheme. The above phrase would be addressed like this (`<#char=0,2353>` represents the complete text).

```
<#char=282,253> a nif:Phrase;
  nif:referenceContext <#char=0,2353>.
```

The reason is to ensure interoperability between different NLP tools that all output NIF format over the same text. Using a uniform node addressing scheme ensures that the triples produced by the different tools will "mesh" together. This is perhaps the most significant difference between FN-LOD and NIF:

- FN-LOD defines Labels "as needed" by linguistic annotation, and locally. Several Label nodes can point to the same piece of text (offsets in the document). Labels are not shared between different annotations (NLP features).
- NIF typically defines Strings for every word and sentence of the document, globally. Each piece of text is

represented by one node (but of course, Words overlap their containing Phrases and Phrases overlap their containing Sentences).

Several NLP features can be attached to this node:

- `nif:oliaLink` for syntactic individual
- `nif:oliaCategory` for syntactic class
- `its:taIdentRef` for Named Entity individual
- `its:taClassRef` for Named Entity class; etc

4. Integrating FN-LOD in NIF

As we have seen in the previous section, the FN-LOD and NIF models for representing annotated text are totally different. Therefore we propose to represent the minimum possible FN nodes, and point to them from `nif:String` using `nif:oliaLink`.

We propose a representation that integrates FN-LOD in NIF (Fig 5), relying on a dependency parse of the sentence. Let *head* be a head-word that governs *word1..N* (and by extension, the phrases governed by these words). Assume *head* corresponds to *lexUnit* that invokes *frame*, and the frame has elements *frameElement1..N*, corresponding to *word1..N*. Just for illustration, assume the frame also has a lexically omitted FE *frameElementN+1* of type CNI (constructional null instantiation).

The easiest way to understand the representation is to think of `fn:AnnotationSet` as **frame instance** and think of `fn:Label` as **FE instance**. The representation consists of 3 parts:

1. **NIF** includes word offset info, as well as the dependency tree from *head* to *word1..N* (not shown). `nif:dependency` or specific dependency parsing properties are used for that tree. E.g. MS uses `upf-deep:deepDependency`
2. **Frame instance** connects `nif:Words` to frames.
3. **Frame definition** is defined in the fnabox ontology (sec 2.4.)

We don't use `fe:label_start` and `fe:label_end` because those would duplicate `nif:beginIndex` and `nif:endIndex` unnecessarily. The same word could participate in several frames (as LU or FE), in which case it will have multiple `nif:oliaLink`. The lexically omitted FE *labelN+1* (of type CNI) has no corresponding NIF node. Nevertheless, it is a full participant in the frame.

The nodes *labelLU* and *layerLU* are redundant and carry no information (except the fixed string "Target"). There's a direct link *nif:oliaLink* from *head* to *annoSet*, which itself points to *frame* and *lexUnit*, so there's little reason to use the indirect path *fn:hasLayer/fn:hasLabel*. In fact the indirect path can be considered harmful, since it causes *head* to have two *nif:oliaLink*, which could cause confusion if *head* participates in several frames. We have included these redundant nodes in Fig 5 to be faithful to the fntbox ontology 2.2.. But they can safely be omitted, which we have done in sec 4.2..

The links of *label1..N+1* (`fn:hasLabel` and `fn:label_FE`) are not redundant. The former ties the frame **instance** together, while the latter points the specific FE in the frame **definition**.

4.1. Querying FN-NIF

FN-LOD in NIF involves a fairly complex graph structure. In this section we show a few queries to extract data from that graph. We use SPARQL property paths liberally (including inverses `^`) and indicate the input parameter of a query with `$`. We don't bother to check the types of intermediate nodes, relying that the specific FN properties will occur only on appropriate nodes.

Find the Frame and LU corresponding to a head-word (if indeed it is the head-word of a frame-annotated phrase):

```
select * {
  $head nif:oliaLink ?annoSet.
  ?annoSet fn:annotationSetLU ?lu;
    fn:annotationSetFrame ?frame}
```

We could also use the round-about path

```
select * {
  $head nif:oliaLink [
    fn:label_name "Target";
   ^fn:hasLabel/^fn:hasLayer ?annoSet.
  ?annoSet fn:annotationSetLU ?lu;
    fn:annotationSetFrame ?frame]}
```

After getting the Frame and LU, we'd want to get all FE and the corresponding *word1..N*:

```
select ?fe ?word ?itype {
  # Find the ?annoSet and ?frame
  $head nif:oliaLink ?annoSet.
  ?annoSet fn:annotationSetFrame ?frame.
  # Get all ?fe, ?label, (optionally) ?word
  ?frame fn:hasFrameElement ?fe.
  ?annoSet fn:hasLayer/fn:hasLabel ?label.
  ?label fn:label_FE ?fe.
  optional {?word nif:oliaLink ?label}
  optional {?label fn:label_itype ?itype}}
```

Each row of the result-set will have a `?fe` of the frame, and either `?itype` (for lexically omitted FEs) or the corresponding NIF `?word`. We don't return `?label` because it's used only for connectivity but doesn't carry useful info.

Find all frames of a sentence together with the corresponding `fn:AnnotationSet`. Usually `nif:word` is used to point out the words of a sentence (that is also the practice in MS):

```
select * {
  $sentence nif:word/nif:oliaLink ?annoSet.
  ?annoSet fn:annotationSetFrame ?frame}
```

Find all frames of the complete text (`nif:Context`) together with the corresponding `fn:AnnotationSet`. NIF mandates that `nif:referenceContext` is used to connect each word to the complete text:

```
select * {
  $context ^nif:referenceContext/
           nif:oliaLink ?annoSet.
  ?annoSet fn:annotationSetFrame ?frame}
```

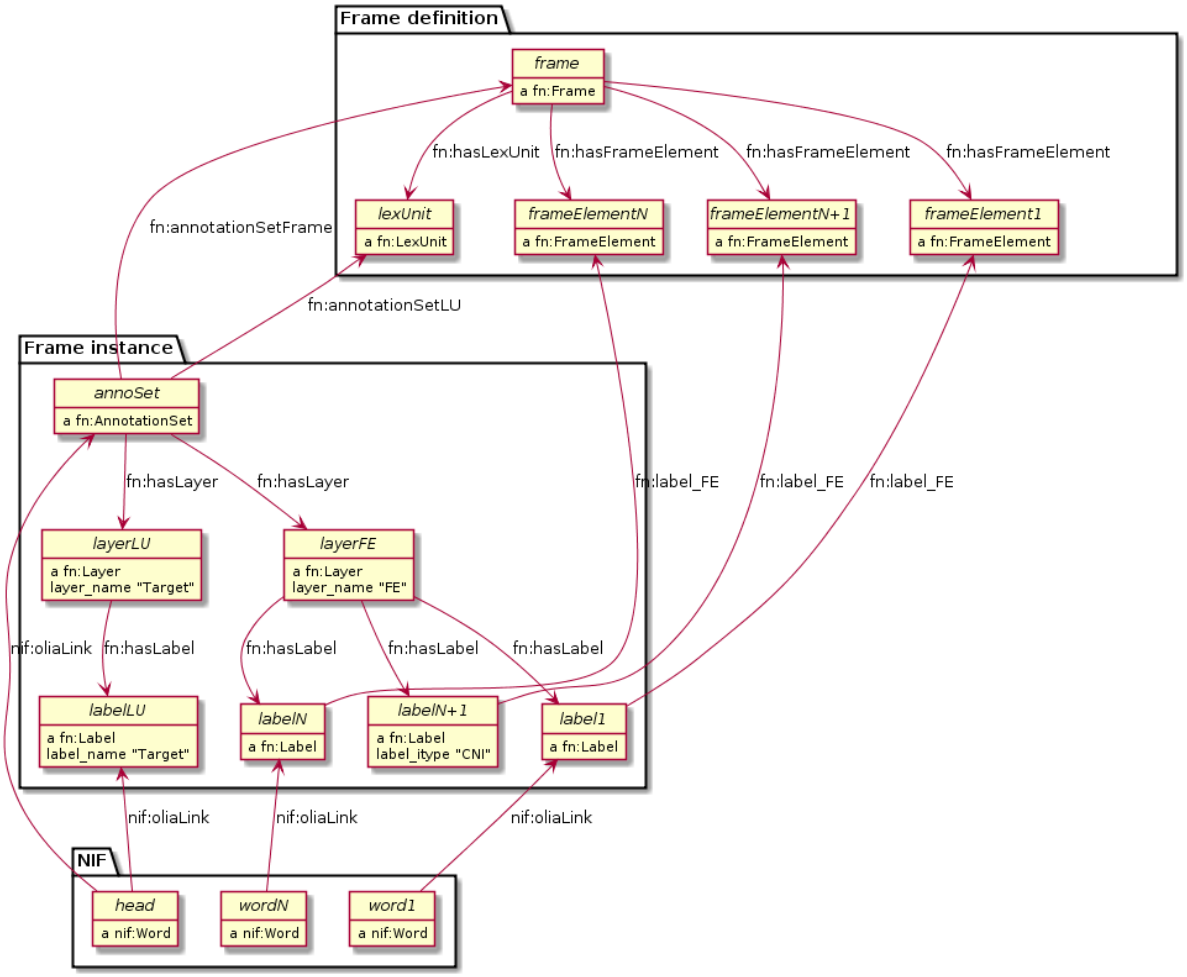


Figure 5: FrameNet Integration in NIF

4.2. Representing the Sample Sentence in FN-NIF

Fig 6 represents the sample sentence as NIF, adding FN-LOD annotations. We represent 3 of the 5 candidate frames (Statement, Topic, Competition); the filtering described in sec 1.2. would leave only the top frame `Statement`

- The top layer shows Frame definitions (fntbox)
- The bottom layer shows NIF words and dependency links between them
- The dotted arrows represent frame instances, connecting words to frames. For simplicity, we don't show the `Label`, `Layer`, `AnnotationSet` nodes (see sec 4.)

./fn-nif-example.ttl represents all SEMAFOR candidate frames. Compared to sec `*Integrating FN in NIF`, we elide the redundant nodes *labelLU* and *layerLU*.

5. Conclusions

We presented an integration of FN-LOD into NIF that allows us to emit various linguistic info about text corpora in NIF in an integrated way: frames (FN), POS tagging (e.g. Penn), morphological, syntactic and dependency parsing (OLIA), named entities (ITS), etc. This integrated representation is used by the MS project.

5.1. Future Work

5.1.1. Represent Confidence

Sec 1.2. remarked that SEMAFOR emits a confidence `score` for each candidate frame. It would be useful to emit this score, allowing clients to select the most probable frames.

- NIF has a property `nif:oliaConf` (confidence of `nif:oliaLink` and `nif:oliaCategory`). But we cannot use it, since the same word may participate in several frames and thus have several `nif:oliaLink`.
- We could use the NIF Stanbol profile to associate several annotations with the same String and emit confidence for each one. But compared to NIF Simple, it uses completely different properties, e.g. `fise:entity-reference` vs

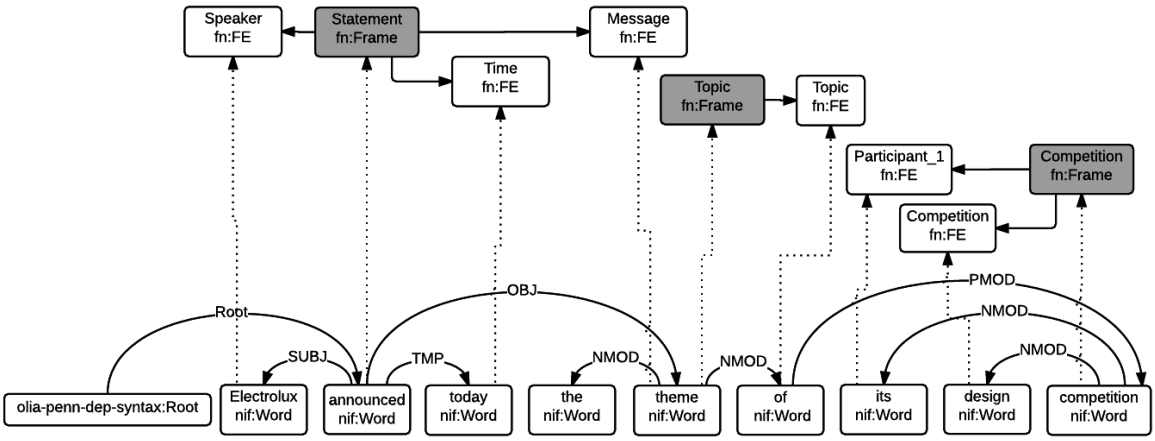


Figure 6: FN-NIF Example

`its:taIdentRef` and `fise:entity-type` vs `its:taClassRef`. And there are stability problems: NIF Stanbol shows different classes and properties compared to [3] fig.3 and Stanbol EntityAnnotation Structure, e.g.

NIF Stanbol	NIF and Stanbol
nif:EntityAnnotation	fise:EntityAnnotation
nifs:extractedFrom	fise:extracted-from
nif:oliaConf	fise:confidence

- Recently a new proposal Provenance and Confidence for NIF annotations was made, motivated by the FREME project. It is part of a developing NIF 2.1 specification currently at Release Candidate stage (NIF 2.1 RC), see source. It offers two options: Using only Generic Provenance and Confidence Properties, or Using Companion Properties (see last 2 columns below). But it is still in flux, e.g. on 14 Mar 2016 a number of properties were split to a separate namespace `nif-ann:`

5.1.2. Create an RDF Shape Description

Our representation doesn't define any new properties: it only combines FN-LOD and NIF properties in an appropriate way. From this point of view, it is not an ontology but an *application profile*, *data pattern* or *RDF Shape*. Recently the W3C RDF Shapes working group has made great advances in analyzing requirements for defining data shapes and formalizing languages to describe them.
It would be useful to define the FN-NIF integration (Fig 6) as an RDF Shape. We could use the brief ShEx language or the more formal SHACL language. However, they are still under development.

5.2. Acknowledgements

This work is part of the MultiSensor project that has received funding from the European Union under grant agreement FP7 610411. The 4 anonymous referees made useful suggestions for improving the article. Object diagrams are made with PlantUML.

6. References

1. Alexiev V. Linguistic Linked Data presentation, Multisensor Project Meeting, Bonn, Germany, October 2014.

2. ARK Syntactic & Semantic Parsing. Noah's ARK research group, Carnegie Mellon University.

3. Hellmann S., Lehmann J., Auer S., and Brümmer M. Integrating NLP using Linked Data. In *International Semantic Web Conference* (ISWC) 2013.

4. Ide N., FrameNet and Linked Data. In *Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, pages 18–21. Baltimore, Maryland USA, 27 June 2014.

5. Moro A., Navigli, R., Tucci, F.M., and Passonneau. R.J. Annotating the MASC Corpus with BabelNet. In *Language Resources and Evaluation* (LREC 2014), Reykjavik, Iceland, May 26-31, 2014. Download page, checked 17 Mar 2016.

6. MultiSensor project. CERTH, DW, Eurecat, Everis, Linguatec, Ontotext, Pressrelations, UPF. Nov 2013 - Oct 2016.

7. Nuzzolese A.G., Gangemi A., and Presutti V. Gathering lexical linked data and knowledge patterns from FrameNet. In *Knowledge Capture* (K-CAP'11), pages 41–48. June 26-29, 2011, Banff, Alberta, Canada

8. Passonneau R., Baker C., Fellbaum C., and Ide N. The MASC Word Sense Sentence Corpus. In *Language Resources and Evaluation Conference* (LREC-12), Istanbul, Turkey. Download page, checked Jan 2016 (offline on 17 Mar 2016).

9. Ruppenhofer J., Ellsworth M., Petruck M.R.L, Johnson C.R., Scheffczyk J. FrameNet II: Extended Theory and Practice, Sep 2010

10. Siemoneit, B., McCrae, J. P., and Cimiano, P. Linking four heterogeneous language resources as linked data. *Workshop on Linked Data in Linguistics: Resources and Applications* (LDL-2015). (2015). Beijing, China, 31 July, 2015. Download page, checked 17 Mar 2016.

11. Vrochidis, S. et al. MULTISENSOR: Development of multimedia content integration technologies for journalism, media monitoring and international exporting decision support. *IEEE International Conference on Multimedia & Expo Workshops* (ICME15). Turin, Italy, 2015. 10.1109/ICMEW.2015.7169818

Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF

Frank Abromeit, Christian Chiarcos, Christian Fäth, Maxim Ionov
Goethe Universität Frankfurt am Main, Germany
{abromeit|chiarcos|faeth|ionov}@informatik.uni-frankfurt.de

Abstract

This paper describes the development of a Linked Data representation of the *Tower of Babel* (*Starling*), a major resource for short- and long-range etymological relations. Etymological dictionaries are highly multilingual by design, they usually involve cross-references to additional monolingual and etymological dictionaries, and thus represent an ideal application of Linked Data principles. So far, however, the Linguistic Linked Open Data (LLOD) community rarely addressed etymological relations. In line with state-of-the-art LLOD practice, we represent Starling data in accordance with the lemon vocabulary developed by the W3C Ontolex Community Group, we discuss the state of the art, experiences and suggest extensions for etymological dictionaries. The conversion we describe is conducted automatically and applicable to *any* Starling dictionary. This paper focuses on modelling issues, using the Turkic Etymological Dictionary for illustration.

Keywords: Linguistic Linked Open Data (LLOD), lemon, ontolex, lemon extensions, etymological dictionaries, Starling

1. Background and Motivation

Etymological dictionaries are highly multilingual by design, they often involve cross-references to multiple monolingual dictionaries as well as to etymological resources for other languages. For this kind of data, we thus expect particularly clear benefits of a Linked Data representation. Because etymological dictionaries are massive collections of cross-references, a Linked Data conversion will provide integrated access to different resources originally linked with each other only at the bibliographical level.

In parts, such cross-references are resolved in existing etymological databases (e.g., in the data discussed here, but also in the commercial Leiden Indo-European Etymological Dictionaries[1]) already, however, this is limited to content residing in *the same database*. Providing etymological data as Linked Data facilitates integrating such resources with etymological, dialectal or historical content, or with corpora maintained by different providers. Linked data allows for expansion beyond the limitations of the existing stand-alone resource, and it thereby contributes to liberating and aggregating content scattered in the web of documents as well as in the web of data.[2]

The *Tower of Babel*, also known as *Starling*,[3] is a web portal on historical and comparative linguistics. Started in 1998 by Sergei A. Starostin and currently maintained by George Starostin, the site provides a great variety of resources for research on the evolution of human languages. Even though Starostin's original premise to identify and to confirm long-range relations is not universally accepted in the field, the portal has attracted the attention of numerous researchers using it to develop or to publish their data collections, but also, to make classical works available over the internet, e.g., Pokorny's Indo-European (Pokorny, 1959) and Vasmer's Russian Etymological Dictionaries (Vasmer, 1953).

This paper describes the conversion of the Starling etymological database to a Linked Data representation. Linked (Open) Data defines rules of best practice for publishing data on the web, and since Chiarcos et al. (2012), these rules have been increasingly applied to language resources, giving rise to the **Linguistic Linked Open Data (LLOD)** cloud.[4]

A *linguistically relevant* resource constitutes Linguistic Linked (Open) Data if it adheres to the following principles:

1. its elements are uniquely identifiable in the web of data by means of *URIs*,
2. its URIs should *resolve via HTTP*,
3. it can be accessed using *web standards* such as RDF and SPARQL, and
4. it includes *links* to other resources to help users discover new resources and provide explicit semantics.

It is Linguistic Linked *Open* Data (LLOD) if – in addition to these rules – it is published under an *open* license.

For language resources, Linked Data provides several important benefits as compared to legacy formalisms (Chiarcos et al., 2013):

Representation Represent linguistic data flexibly as linked graphs

Structural Interoperability Integrate data easily using RDF

Explicit Semantics Define RDF resources by referring to term bases

[1] http://www.brill.com/publications/leiden-indo-european-etymological-dictionary-series

[2] Here, liberation means to remove *technical* hurdles to access and to export content. Liberation in a strict sense also requires eliminating legal obstacles. While this is a desirable, long-term goal as well, we are bound by the copyright of the original data set.

[3] http://starling.rinet.ru

[4] http://linguistic-lod.org

Conceptual Interoperability Use and re-use shared vocabularies

Federation Combine data from multiple, distributed sources

Dynamicity Access the most recent edition live over the web

Ecosystem Benefit from widely available open source tools for RDF and linked data

The capability to refer to and to search across distributed data sets (federation, dynamicity, ecosystem) in an interoperable way (representation, interoperability) allows one to design novel, integrative approaches on accessing and using etymological databases, but only if common vocabularies and terms already established in the community are being used, re-used and extended.

So, in line with state-of-the-art practices in the LLOD community, we employ the lemon vocabulary[5] developed by the W3C OntoLex Community Group.[6] So far, however, the community behind the Linguistic Linked Open Data cloud has rarely addressed this type of language resource. Notable exceptions include Moran and Brümmer (2013), Chiarcos and Sukhareva (2014), Khan et al. (2014), de Melo (2014) and Declerck et al. (2015).[7] Till now these proposed did not cumulate in common specifications and recommendations regarding the representation of etymological resources in general: Moran and Brümmer (2013) established the lemon vocabulary (McCrae et al., 2011) for the representation of diachronic data sets used to automatically detect etymologically related cognates, but they did not discuss the explicit representation of etymological links. Khan et al. (2014) also adopted lemon, but focused on modeling the temporal extent of historical word senses rather than etymological information in a strict sense. Among those who did attempt to represent etymological relations, de Melo (2014) proposed a small special-purpose vocabulary consisting of 7 new object properties for representing etymologies harvested from the English Wiktionary. By grounding their model in a vocabulary commonly used by the LLOD community, Chiarcos and Sukhareva (2014) proposed a more sustainable solution as an extension of lemon (McCrae et al., 2011) which they applied to a collection of etymological resources and automatically generated translation pairs among older Germanic languages.

Here, we attempt to generalize over both proposals and illustrate the resulting vocabulary to the novel, and massive set of etymological data available from the Tower of Babel project.

[5] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification, https://github.com/cimiano/ontolex

[6] https://www.w3.org/community/ontolex/

[7] In addition, the World Loanword Database (Haspelmath and Tadmor, 2009, WOLD) provides similar information, although its current RDF export available from CLLD (Forkel, 2014) does not seem to comprise cognate information. Crist (2005) describes a pre-RDF approach to formalize cognates and etymological relations by means of feature structures.

2. The Tower of Babel

The *Tower of Babel* is a web based project on historical and comparative linguistics started by Sergei A. Starostin in 1998. It is widely recognized as a major resource for short- and long-range etymological relations. With more than 50 etymological dictionaries that cover all the world's major language families, it is an extensive resource of its kind. Its etymological databases can be browsed via a web interface, and dictionary data is also available for download. We illustrate its functionality for the *Turkic Etymological Dictionary* by Dybo et al. (2012).[8] Starling allows one to explore the dictionaries by means of faceted browsing using a coarse-grained phylogenetic tree (Figure 1.a).

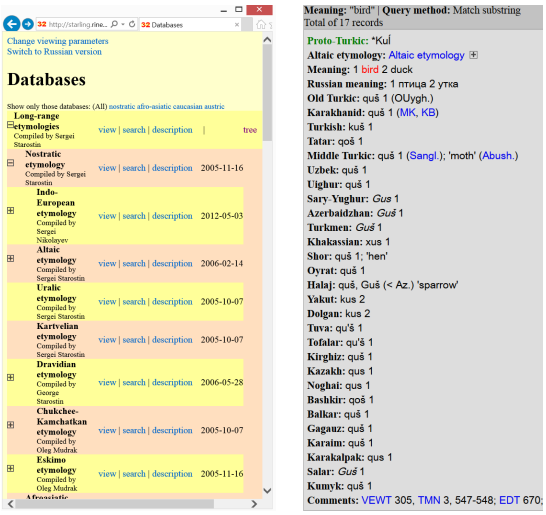


Figure 1: Starling: (a) phylogenetic tree for faceted browsing, (b) first query result for meaning "bird" in the Turkic etymological dictionary

Most nodes in this tree correspond to a single database for a language family or language (e.g., an Altaic etymological dictionary), sub-nodes represent independent databases for individual languages in a family or a sub-family (e.g., the Turkic etymological dictionary). In addition, a query interface is provided which allows filtering with respect to a number of predefined fields. These correspond to columns in the underlying relational database. Every dictionary seems to be stored in a different table as dictionaries provide different fields in search and visualization, but dictionaries are linked with each other through HTML hyperlinks. Figure 1.b illustrates an example entry from the Turkic etymological dictionary retrieved by a query for the meaning "bird": At the top of the result record the *Proto-Turkic* form is shown. It relates to a reconstructed form of the word.[9] The *meaning* of the proto-form is always given in English and Russian. Besides the comment field at the end of the record all other entries belong to a *cog-*

[8] http://starling.rinet.ru/cgi-bin/response.cgi?root=config&morpho=0&basename=\data\alt\turcet&first=1

[9] Reconstructed forms are marked by $*$.

Figure 2: Bibliographic information for **Abush.**

nate in a language of the *Turkic language family*.[10] A language entry can store different cognates which can be *indexed* with a natural number referring either to a given English or Russian meaning, glosses for a cognate appear as quoted English comment. A gloss can also refine the meaning as for the Middle Turkic word with the meaning 'moth' in Fig. 1.b. The entire record of a proto-form and its language-specific attestations is concluded with a comment field which contains hyperlinks to *bibliographical entries*. Also individual entries can be augmented with literature references (in parentheses), as the Middle Turkic entry in Figure 1.b and the abbreviation *Abush.* (Fig. 2).

3. Analyzing Starling Dictionaries

Along with its web version, the Tower of Babel also provides the offline tool *star4win*[11] which can be used to explore and to edit existing or create new etymological dictionaries under MS Windows. For this purpose, database dumps for many[12] of the dictionaries can be downloaded. Unfortunately, these remain in a proprietary legacy database format (dBASE) which cannot be directly processed without losing complex UTF8 characters. Instead, we used *star4win* to create an XML export for the individual databases. The XML export of the individual dictionaries represents the simple structure of a single relational table by distinguishing record IDs (primary keys) which are assigned attribute-value structures in different fields (columns). Focusing on etymological relations, we built a parser to extract all the complex information. As Figure 3 shows, an entry of the etymological database is based on a form in one (proto-) language (`PROTO`), and realizations in different related languages. A language field is introduced with a proprietary 3-Letter code, (e.g. TAT for Tartar) and includes the cognate(s) together with the respective meaning. This meaning is represented by a sense

[10]Old Turkic, Karakhanid, Turkish, Tatar, Middle Turkic (Chagatai), Uzbek, Uighur, Sary-Yughur, Azeri, Turkmen, Oyrat, Khalaj, Khakassian, Chuvash, Yakut, Shor, Dolgan, Tuva, Tofalar, Kirghiz, Kazakh, Noghai, Bashkir, Balkar, Gagauz, Karaim, Karakalpak, Salar, Kumyk.

[11]`http://starling.rinet.ru/download/star4win-2.4.2.exe`

[12]At the time of writing 40 etymological dictionaries are available for star4win. Another 11 can be downloaded in PDF and EXCEL format.

```
<record id="6">
  <field name="NUMBER">6</field>
  <field name="PROTO">*Kuĺ</field>
  <field name="PRNUM">1157</field>
  <field name="MEANING">1 bird  2 duck</field>
  <field name="RUSMEAN">1 птица  2 утка</field>
  <field name="ATU">quš 1 (OUygh.)</field>
  <field name="KRH">quš 1 (MK, KB)</field>
  <field name="TRK">kuš 1</field>
  <field name="TAT">qoš 1</field>
  <field name="CHG">quš 1 (Sangl.); 'moth' (Abush.)</field>
  ...
  <field name="REFERENCE">VEWT 305, TMN 3, 547-548; EDT 670;
   ЭСТЯ 6, 180-182, Лексика 168, Stachowski 162. Chuv. хьlat
  'hawk' &lt; Mong.
  </field>
</record>
```

Figure 3: XML representation for Fig. 1.b

ID which resolves to a substring in the `MEANING` (and `RUSMEAN`) field. The meaning may be complemented with optional gloss information written in parentheses or quotes (e.g., about its context or specific meaning facets). It should be mentioned that a cognate can also have several meanings attached to it as a list of IDs and that the number of cognates in a language is not restricted to one (Fig. 4).

```
<field name="JAK">bitij- 1, 2, bitī 'танец'</field>
<field name="KRG">bij 3, bij-le- 1</field>
<field name="KAZ">bij 3, bij-le- 1, büji-1 (dial.)</field>
<field name="NOGX">bĭji-</field>
```

Figure 4: Multiple meanings and cognates

Finally, the `REFERENCE` field (Fig. 3) provides bibliographic references to the data with author and additional gloss information.

While converting the XML structure is trivial, validating and interpreting the values of individual fields was a time-consuming task. One issue was that the data lacks a consistent encoding style and has a lot of variation in its notational conventions for free-text entries. Starling circumvents this problem by storing this complex semi-structured data as unanalyzed BLOB/CDATA content without any further analysis. This is a convenient solution that leads to simple data base structures, but important information remains inaccessible for automated processing. Considering Starling data from other dictionaries, this problem multiplies even further as a greater number of editors and their individual notation styles are involved. As a consequence, we face severe problems when parsing such semistructured content:

(1) Multiple alternative proto-forms are encoded in a single record (e.g. `<field name="PROTO">*göŕ (= *gör-s) / *gör-</field>`)

(2) Rare encoding variations in the data (e.g. `<field name="HAK">(pudurčun, püdürčün)</field>` with cognates exceptionally encoded with parentheses)

(3) Mixed references and comments in the `REFERENCE` field (see Fig. 3) `<field name="REFERENCE">EDT 371 (but not from *bas- 'press'!). Turk.`

```
; Mong.  basa (Clark 1980,
39).</field>
```

The Turkic dictionary contains a total of 2,017 records and 21,835 cognate entries for Proto-Turkic forms with their respective cognate(s) and meaning(s) in up to 30 Turkic languages. Out of these, 366 records (18.2%) show the problems of type (1), and 5109 cognate entries (23.4%) show problems of type (2). Both kind of data were excluded from automated conversion. Bibliographic references were converted in general, but using heuristic parsing rules that may lead to information loss in case of problems of type (3). On rare occasions, the source XML was manually edited to treat such cases.

4. Data Modeling

Following conventions from the LLOD community, we employ the lemon (*Lexicon Model for Ontologies*) vocabulary. Lemon consists of five modules, which comprise *Ontology-lexicon interface* (ontolex), *Syntax and Semantics* (synsem), *Decomposition* (decomp), *Variation and Translation* (vartrans) and *Linguistic Metadata* (lime). Originally, lemon has been developed to add linguistic information to existing resources developed by the Semantic Web community, hence its name. With the growth of the LLOD cloud, lemon has also been adopted by users from linguistics and the language resource community as a means to represent lexical resources as Linked Data and is now established as a de-facto standard for this purpose.

However, this linguistic use case tends to violate a number of lemon requirements. Very important in this context is the notion of an ontology. Historically, this is presupposed by lemon, but it does not exist for dictionaries in general. Accordingly, it needs to be forced on the data (e.g., with the DBpedia linking of the RDF version of PanLex),[13] artificially introduced (e.g., by duplicating senses, Moran and Brümmer, 2013), or just omitted (Chiarcos and Sukhareva, 2014).

Another aspect is that several properties need to be introduced to account for relations between etymological cognates.

4.1. Etymological Relations

De Melo (2014) introduced 7 properties for this purpose, which can be roughly classified into synchronic relations which connect words of the same language (`variant_orthography`, `derived`, `has_derived_form`, `is_derived_from`), and diachronic relations which connect words across different languages (`etymology`, `etymologically_related`, `etymological_origin_of`).

However, de Melo does not define these properties, so this is merely an interpretation. In fact, the differentiation between several etymological or derivational properties (e.g., `etymology` vs. `etymologically_related`) remains unclear.

More rigidly than de Melo (2014), lemon distinguishes lexical entries, lexical forms, and lexical senses. Whereas semantic relations operate on the level of lexical sense, etymological relations connect different lexemes – in addition to a relationship on the meaning level, this involves a systematic relationship between different forms. At the same time, etymologically related forms may differ in their meaning, hence, etymological relations do not exist on the level of form alone. Accordingly, etymological relations can only be represented as a property between one `LexicalEntry` and another. This also means that de Melo's synchronic relations (which operate on the level of lexical form or which involve morphological patterns rather than etymology) can be expressed with the core lemon vocabulary and are beyond the scope of a lemon extension for etymological relations.[14] We thus focus on de Melo's etymological relations which seem to differ by directionality (or the lack of knowledge about it).

If etymological cognates can be identified in different languages, it is not always clear whether one was the source of the other, whether both originate from the same source, and whether this source was a common ancestral language or a common sub- or adstrate. To express a generic etymological link without additional directionality information, we introduce a property `cognate`. This novel property is assigned the namespace `lemonet` (lemon with etymological extensions) previously declared by Chiarcos and Sukhareva (2014).[15] The property `lemonet:cognate` is symmetric (we do not know about directionality) and transitive (the relation may be indirect, e.g., through a common substrate, and can thus not be distinguished from relations inferred from the transitive closure of direct etymological links).

If source and target are known, a subproperty `lemonet:derivedFrom` is introduced[16]. Similar to `lemonet:cognate`, it is transitive, but it is not symmetric. In order to keep the *lemonet* vocabulary as minimalistic as possible, we do not explicitly represent the inverse of this property – it can nevertheless be queried by `lemonet:derivedFrom` using SPARQL 1.1 property paths. Taken together, we arrive at a sparse representation of etymological links which supports inferring general

[13] PanLex is available from `http://panlex.org`. A forced DBpedia linking is noisy and conceptually problematic because DBpedia is meant to provide information about concepts important in the world, but not to represent the lexical meaning of every possible word that can be used as a gloss. For example, there is no exact DBpedia concept for 'bird'. `http://dbpedia.org/c/9CB4KXNZ`

[14] `Derived` properties between morphemes and words correspond to the containment relation between `LexicalEntry` and `Affix`, etc., `derived` properties between lexemes correspond to different morphological realizations (`ontolex:isSenseOf`) of the same `LexicalSense`, `variant_orthography` can be represented by two different `ontolex:writtenRep` properties of the same lexical form.

[15] `lemonet:cognate` replaces the property `etym` of Chiarcos and Sukhareva (2014) as the name `etym` is redundant with the reference to etymology in the namespace.

[16] This corresponds to de Melo's `etymological_origin_of`. However, this name carries the connotation that the object of the property is *ultimate* origin of a word which may or may not be the case. Despite its similarity with de Melo's `derived`, etc., the lemonet property cannot be confused with morphological processes because it resides in an etymology-specific namespace.

cognate relations by subsumption and transitive/symmetric closure. Both can be queried easily and efficiently by combining RDFS reasoning and SPARQL 1.1 Property Paths. Using SPARQL 1.1 Property Paths, it is in particular possible to distinguish direct `derivedFrom` relations (querying for `derivedFrom`) from indirect ones (querying for `derivedFrom`*), so that this modelling does not lead to a loss of detail.

In addition to these two properties, Chiarcos and Sukhareva (2014) introduced a `lemonet:translates` property. With the novel lemon translation model, this is no longer necessary.

4.2. Possible Extensions and Alternative Views

Following the relational view on etymologies, we arrived at a minimal lemon extension to represent core information in etymological dictionaries as provided by Starling. The `lemonet` namespace now comprises exactly two properties, `cognate` and `derivedFrom`.

In the longer perspective, this may be complemented with means to represent the temporal and geographic scope of specific lexemes, possibly building on Khan and Frontini's earlier work. However, this is left for future research as we do currently not have any such specific data in a machine-readable fashion, neither in Starling, nor provided by de Melo (2014) or Chiarcos and Sukhareva (2014).

Another possible extension pertains to the degree of certainty about an etymological link. However, we see this as being closely related to the more general problem to represent uncertainty about RDF triples and advise to rely on existing solutions based on RDF reification such as FuzzyRDF (Straccia, 2009). Accordingly, an etymology-specific solution does not seem to be necessary.

The issue of reification also touches related research on representing etymological relations from monolingual (rather than etymological) dictionaries. This is exemplified in the treatment of etymological information by Declerck et al. (2015). Here, a `LexicalEntry` can receive a `hasEtymology` property whose object (the etymology) is the description of a language-specific etymology, complemented with additional properties for temporal extent (`hasCentury`), a string representation (`hasEtymologyForm`) and a language. This approach provides an elegant way to represent the period a loan word has entered another language, and it is an adequate representation of etymological information as occasionally found in comments of many dictionaries, i.e., as a historical remark about a particular lexical entry (i.e., a monolingual lexeme, as considered by Declerck et al., 2015). It is, however, less adequate for the information found in designated *etymological dictionaries* which aim to trace the entire history and/or distribution of a form and its cognates in a specific linguistic and/or cultural sphere – normally covering entire language families. For these, it is essential that references to other dictionaries are entities which can be *resolved* (e.g., we would like to follow a pointer from the Altaic etymological dictionary to a form in the Turkic etymological dictionary and further to its language-specific attestations). We thus require an object property in place of `hasEtymologicalForm`. Furthermore, etymological links do not exist on the level of forms, but the very notion of etymological *cognates* (inherited or loan words) involves an aspect of meaning (Crist, 2005).

For example, English *bank* is a loan word originating from a Proto-Germanic word which is the source of Modern English *bench*. However, its two meanings have very different histories. In the sense of 'river bank', it originates from a sea-faring Germanic people in the Middle Ages (either Low German or Scandinavian, cf. German *Sandbank* 'sandbank'). In the sense of 'financial institute', it originates from Italian *banca*, itself a loan from a Langobardian (Old High German) source with the meaning 'bench'. Both in a dictionary of Modern English as well as in a dictionary of Proto-Germanic, both could be grouped together under the same `LexicalEntry` (same form, for Proto-Germanic also same meaning), but their etymologies remain distinct and can only be distinguished on the sense level. For this reason, we prefer to see etymological links not as an association between strings,[17] but rather as relations between elements that combine form and sense information, i.e., as lexical entries.

For harmonizing both approaches, the relational representation of etymologies (as here, and found in designated etymological dictionaries), and the descriptive representation of etymologies (as described by Declerck et al., 2015, and found in many historical and dialectal, but monolingual dictionaries), a conventional solution would be to rely on standard RDF reification – which may represent a problem to decidability of subsumption inferences, though.

Within ontolex, however, RDF reification is not necessary in this case. The property `vartrans:lexRel`[18] from which `lemonet:cognate` and `lemonet:derivedFrom` are derived, is coupled with the class `vartrans:LexicalRelation` (with properties `vartrans:source` and `vartrans:target`, resp., their generalization `vartrans:relates`) as its reified representation.

By analogy, we propose two subclasses of `vartrans:LexicalRelation`, i.e., `lemonet:Cognate` and `lemonet:Derivation`. In practical applications, the relational representation can be transformed into the class-based representation by a single SPARQL UPDATE command (also the other way around, even though with possible information loss).

```
INSERT {
  ?derivation a lemonet:Derivation;
      vartrans:source ?s;
      vartrans:target ?t.
} WHERE {
  BIND(UUID() as ?derivation).
  ?t lemonet:derivedFrom ?s.
}
```

For the Starling data, we stay with the relational modelling as it yields a more compact representation (1 triple per etymological relation rather than 3).

[17] Or, more precisely, attribute-value pairs (data properties) assigned to lexical entries (resp. individuals bundling the etymological information about a particular lexical entry).

[18] From the lemon *vartrans* module

5. Conversion to RDF

The actual conversion of the Starling data is done with a stand-alone Java application which uses Apache-Jena[19] libraries for RDF modelling. We represent etymological relations with `lemonet:derivedFrom`. Most Starling dictionaries, including the Turkic etymological dictionary, are organized by underlying proto-forms, constructed from their cognates in descendant languages. For every proto-form p and its language-specific cognate c in a descendant language, we thus add the triple c `lemonet:derivedFrom` p. By subsumption inference, transitivity and symmetry of its superproperty, `lemonet:cognate` relations can be inferred automatically between all language-specific forms.

In the Starling data, the directionality of etymological links is generally known, we thus use `derivedFrom` properties between language-specific lexemes and proto-forms. While this is generally correct, more detailed relations between the forms of different descendant languages cannot be inferred from the structure of the database. For example, it is very likely that an Uighur word originates from a Proto-Turkic root *via* the Old Uighur (Old Turkic) form, or, likewise, that a modern Turkish word represents the immediate continuation of its Middle Turkic cognate rather than an independent reflex of the underlying Proto-Turkic root. This diachronic relation between the languages is not represented in the Tower of Babel data. However, an indirect `cognate` relationship can be inferred from the transitive closure of the superproperty `cognate`.[20] With additional information about the historical relation between different languages, a direct link can be reconstructed.

5.1. Implementing etymological relations

The proposed etymological properties can be implemented with the existing lemon vocabulary. We model our properties `lemonet:cognate` and `lemonet:derivedFrom` as subproperties of the lemon property `vartrans:lexicalRel`. We only have to add the *transitive* closure to both and the *symmetric* closure to `lemonet:cognate` to get the desired result.

Definition of cognate

```
lemonet:cognate
 a owl:ObjectProperty,
   owl:SymmetricProperty,
   owl:TransitiveProperty;
 rdfs:domain ontolex:LexicalEntry;
 rdfs:range ontolex:LexicalEntry;
 rdfs:subPropertyOf vartrans:lexicalRel;
 rdfs:label "etymological relation"@en;
rdfs:comment "The 'cognate' property
    relates two lexical entries that stand
    in some etymological relation."@en.
```

Definition of derivedFrom

```
lemonet:derivedFrom
 a owl:ObjectProperty,
   owl:TransitiveProperty;
 rdfs:domain ontolex:LexicalEntry;
 rdfs:range ontolex:LexicalEntry;
 rdfs:subPropertyOf lemonet:cognate;
 rdfs:label "directed etymological
     relation"@en;
 rdfs:comment "The 'derivedFrom' property
     relates two lexical entries that stand
     in a etymological relation where
     source and target are known. The
     subject position holds the source
     whereas the object position holds the
     derived word"@en.
```

5.2. Modelling with lemon

Lexical entries are organized into `Lexicon`s. For editing printed books, these correspond to the original source, and accordingly, they have been the traditional main locus of language information in an older lemon version, entailing that every `LexicalEntry` in a `Lexicon` is from the same language. In the current model, it is possible to assign lexical entries a language on their own, but we chose to follow the traditional approach as it yields a less redundant representation.[21] Accordingly, converting the Starling dictionaries produced a great number of heavily interlinked `Lexicon`s out of a single multilingual dictionary, e.g., 30 language-specific `Lexicon`s for the Turkic etymological dictionary.

For every `Lexicon`, then, its language was provided in its original string representation as `lime:language`[22] as well as a link to language identifiers from `lexvo.org` (de Melo, 2015, using `dct:language`).

A lexicon then contains all words found in a particular language as lexical entries. A lexical entry carries one or multiple senses (as a commented reference to the `meaning` element in Starling XML). As language-specific cognates refer to the spectrum of meanings of the proto-form (by coindexation in XML), the meaning of a cognate is thus always linked to the meaning of its proto-form as shown in the example below.

 # Lexicon definition

```
star:lexPT a lime:Lexicon ;
    lime:language "Proto-Turkic";
    lime:entry star:lexPT/KuÍ.
```

 # Lexical entry

```
star:lexPT/KuÍ a ontolex:LexicalEntry;
     ontolex:canonicalForm
             [ontolex:writtenRep "⋆KuÍ"].
```

[19] `http://jena.apache.org/`

[20] The 'direct' `cognate` links can be inferred using RDFS reasoning, the transitive closure can be queried with SPARQL 1.1 Property Paths.

[21] When representing language information with `lime:language` (as literal) and `dct:language` (as link to `lexvo.org`) for every `LexicalEntry`, this requires 35,702 triples for the Turkic dictionary. If, instead, this is assigned to `Lexicon`, it requires 60 triples only.

[22] From the lemon Linguistic Metadata (lime) module.

```
# Multiple senses for a proto lexical entry

star:lexPT/Kuĺ
    ontolex:sense star:lexPT/Kuĺ/sense1,
                  star:lexPT/Kuĺ/sense2.


# Definition of a sense

star:lexPT/Kuĺ/sense1 a ontolex:LexicalSense;
     skos:definition "bird"@en.


# Cognate sense linking

star:lexDLG/kus a ontolex:LexicalEntry;
     ontolex:canonicalForm
             [ontolex:writtenRep "kus"];
     lemonet:derivedFrom star:lexPT/Kuĺ;
     ontolex:sense
     [a ontolex:LexicalSense;
  ontolex:reference star:lexPT/Kuĺ/sense1].
```

Such cross-references between lexical entries associated with different `Lexicons` are an innovative aspect of our conversion. It is, however, not unusual in this community to refer to sense definitions from another dictionary. For example, Pokorny (1959) represents a classical work for Proto-Indo-European, and his lexical entries (proto-forms clustered according to their form and sense information) are referred from other dictionaries. Based on the English language description in the XML we identified ISO 639-3 language codes and links to lexvo.org.
In this process, we observed a number of issues: Starling abbreviations do not follow ISO 639 and needed to be interpreted with the help of experts (e.g., in order to identify Chalkan as one variety of North Altaic). Indeed, a direct mapping to ISO 639 was not always possible. For example, ISO 639 does not represent proto-languages. Even where an approximate mapping was possible, the granularity of ISO 639 is insufficient; for example, ISO 639 does not distinguish Central Asian Middle Turkic (Karakhanid) and Middle Turkic (both `xqa`).[23]

5.3. Results and Extensions

Applied to the Turkic Starling dictionary, the converter yields results as summarized in Table 1.
The conversion process described above for the Turkic etymological dictionary can analogously be applied to other Starling dictionaries because

- The XML structure of the etymological dictionaries is very similar.
- All dictionaries use a similar structure of single relational tables (with differently labeled columns).

[23]The more fine-grained Glottolog classification (`http://glottolog.org`) does not help in this case as it has a focus on modern and especially endangered languages. Accordingly, we employed language codes from multitree (`http://multitree.org`), i.e., `qqj` and `qjj` for the example.

XML proto-forms	
total	2,017
convertible	1,651
conversion rate	81.8%
RDF triples	
total	145,981
per lexicon	4,709
Lemon lexical entries	
total	17851
per lexicon	576
e.g.,	
Karakhanid (`xqa`)	699
Chagatai (`chg`)	412
Mod. Turkish (`tur`)	729
Turkmen (`tuk`)	820

Table 1: Extraction statistics for *Turkic Starling dictionary* and the 30 linked etymological lexicons built from it

The dictionaries differ in the syntax used for defining meanings (senses) and sense references, and in their conventions represent complex, multi-component information into a single cell in this table.
In order to account for these variations, a generic Starling converter requires refinement of its extraction rules, as it cannot rely on the regularities of a well-defined and verifiable data structure *within* the textual content of single cells from the original Starling tables. A generic converter which merely evaluates the original table structure is thus necessarily insufficient and requires resource-specific adjustments, e.g., to account for the resource-specific way that complex glosses are to be parsed out of the text representation. For that purpose our parser can be extended with specific extraction rules suited to cover irregular syntax as described above.
In order to assess the performance of the *unmodified* converter, it was tested on other Starling dictionaries of the Altaic language family.[24] Even without fine-tuning the parser, the results from Tab. 2 indicate relatively reliable extraction rates across different languages for both proto-form and cognate processing. However, this is partially due to the fact that the Turkic etymological dictionary is the most mature of these dictionaries, it is thus more likely to display particularly complex information.[25] Without further optimization, we thus have to expect a loss rate of 10-20% unparseable proto-forms, and 10-35% unparseable cognates when applying the parser to other Starling dictionaries.

Possible refinements and extensions include:

- Improve parsing for semi-structured records with multi-word definitions/glosses or multiple proto-forms. Multiple proto-forms are responsible for a loss rate of 18.2% of proto-forms and 23.4% of cognates.

[24]`http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl`

[25]It should be noted that an XML cognate field can contain a list of words which are then converted to multiple lexical entries (Fig. 3). Cognate numbers in Tab. 2 refer to successfully parsed XML fields, not the number of `lemonet:cognate` relations.

	XML proto-forms	*XML cognates*	*Triples*
Turkic	82% (1651/2017)	77% (16726/21835)	145,981
Japanese	83% (1410/1705)	91% (5487/6009)	45,873
Korean	91% (1101/1206)	84% (1697/2025)	19,016
Mongolic	83% (1799/2173)	68% (7756/11328)	74,171
Tungus	81% (1963/2435)	83% (8260/9902)	66,549

Table 2: Extraction statistics of Starling dictionaries for the Altaic language family

- Improve parser for bibliographic references.
- Link meanings to other LOD resources in the cloud.

The current RDF representation of the Turkic and other Starling dictionaries provides the sense definitions of proto-forms, but without formally defined semantics. Only the textual definition is preserved here, but it should ideally be augmented with a reference to other LOD resources. A naive approach would be to focus on single word definitions and to rely on DBpedia or BabelNet, and their respective linking services, DBpedia Spotlight and Babelfy.[26]
In the Semantic Web community, DBpedia would be preferred for this purpose, but we consider *lexical resources* more appropriate as they cover a greater portion of the vocabulary. BabelNet is a popular, large-coverage lexical resource, constructed from WordNet and automatically enriched from multi-lingual sources. Unfortunately, this automatic enrichment yielded data with a considerable level of noise – intolerable for philological and linguistic research.
We thus have to focus on large-coverage *manually constructed* lexical resources, and we propose to link to the upcoming LOD version of the WordNet Interlingual Index (Bond et al., 2016, ILI). In our approach, the ILI would act as a sense repository, much like the ontology prescribed by the lemon model. However, it should be noted that the status of WordNet as an ontology is controversial in the community, so that, again, this may violate existing lemon conventions.
In future research, we are going to experiment with conventional Word Sense Disambiguation (WSD) and Entity Linking techniques to map sense information to WordNet and the ILI. As both are complicated by the sparsity of sense and gloss definitions in our data, this linking represents a challenging task.
One appropriate solution requires the acquisition of parallel or glossed corpus data for major languages in the Turkic language familiy in order to assess the distributional semantics of the respective forms in at least some of the languages in the language family. With corpus-driven WSD for major languages such as Turkish, Azeri, or Kazakh, then, we can project sense linkings to other Turkic languages and Proto-Turkic.

6. Summary and Conclusion

We described the development of an RDF converter for the Starling etymological database, a massive, representative, and world-wide collection of etymological information created by experts in comparative and historical linguistics and from classical works in the field. Starling is currently provided in a human-readable web edition and is downloadable in the proprietary format of a relational legacy database only. With an RDF converter and by linking it with ISO 639-3 language identifiers from lexvo as well as by resolving cross-references between different etymological dictionaries provided as part of the Tower of Babel, we created a Linked Data edition of this information.

Creating a state-of-the-art machine-readable representation facilitates accessibility and usability of this data. If legal clearance for selected Starling dictionaries can be achieved, these will be published under an open license and become part of the LLOD cloud. The converter itself is available as open source from `http://acoli.informatik.uni-frankfurt.de/liodi`.

Beyond converting a particular resource, we see our paper as an important contribution to the establishment of conventions to represent etymological data in RDF. We gave an overview over the state of the art in representing etymological dictionaries as Linked Open Data. Based on this and on the data as encountered in Starling, we discussed problems of the conversion process as well as necessary extensions and adjustments. Building on earlier research, we propose a minimal inventory of lemon extensions for etymological relations, we discussed its status in relation to alternative views on etymological relations and we suggested a reification-based approach to capture etymological information from both the relation-based paradigm to representing etymological information exemplified here and the description-based paradigm described by Declerck et al. (2015).

Acknowledgments

We would like to thank three anonymous reviewers for insightful feedback and helpful comments, Monika Rind-Pawlowski and Irina Nevskaya for their help in deciphering Starling language identifiers and Irina Nevskaya and Anna Dybo for providing us with Starling data and guidance with respect to this. Our research was partially conducted in the context of the Junior Research Group 'Linked Open Dictionaries (LiODi)', funded by the German Federal Ministry of Education and Research (BMBF) since December 2015.

[26] `http://spotlight.dbpedia.org`, `http://babelfy.org`

References

Bond, F., Vossen, P., McCrae, J., and Fellbaum, C. (2016). CILI: The Collaborative InterLingual Index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, Bucharest, Romania.

Chiarcos, C. and Sukhareva, M. (2014). Linking etymological databases. a case study in germanic. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014): Multilingual Knowledge Resources and Natural Language Processing*, pages 41–49, Reykjavik, Iceland.

Chiarcos, C., Nordhoff, S., and Hellmann, S. (2012). *Linked Data in Linguistics*. Springer, Berlin.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Lexical Linked Data. In A. Oltramari et al., editors, *New trends of research in ontologies and lexical resources*. Berlin, Springer, Heidelberg.

Crist, S. (2005). Toward a formal markup standard for etymological data. In *Annual Meeting of the Linguistic Society of America 2005 (LSA-2005)*, Oakland, CA.

de Melo, G. (2014). Etymological Wordnet: Tracing the history of words. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, May.

de Melo, G. (2015). Lexvo.org: Language-related information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, 6(4):393–400.

Declerck, T., Wand-Vogt, E., and Mörth, K. (2015). Towards a pan-European lexicography by means of Linked (Open) Data. In *Proceedings of the 4th Biennial Conference on Electronic Lexicography (eLex-2015)*, Ljubljana, Slovenia.

Dybo, A. V., Starostin, S. A., and Mudrak, O. A. (2012). *Etymological Dictionary of the Altaic Languages*. Brill Academic Publishers, Leiden.

Forkel, R. (2014). The Cross-Linguistic Linked Data project. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014): Multilingual Knowledge Resources and Natural Language Processing*, Reykjavik, Iceland.

Haspelmath, M. and Tadmor, U. (2009). *World Loanword Database (WOLD)*. Max Planck Institute for Evolutionary Anthropology, Leipzig. `http://wold.clld.org/`.

Khan, F., Boschetti, F., and Frontini, F. (2014). Using lemon to model lexical semantic shift in diachronic lexical resources. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014): Multilingual Knowledge Resources and Natural Language Processing*, pages 50–54, Reykjavik, Iceland.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the Semantic Web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*, pages 245–259.

Moran, S. and Brümmer, M. (2013). Lemon-aid: Using Lemon to aid quantitative historical linguistic analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, Pisa, Italy.

Pokorny, J. (1959). *Indogermanisches Etymologisches Wörterbuch*. Francke, Bern.

Straccia, U. (2009). A minimal deductive system for General Fuzzy RDF. In *Proceedings of the 3rd International Conference on Web Reasoning and Rule Systems (RR-2009)*, pages 166–181, Chantilly, VA, USA.

Vasmer, M. (1953). *Russisches Etymologisches Wörterbuch*. Winter, Heidelberg.

Linked Open Data Dataset from Related Documents

Petya Osenova and Kiril Simov
Linguistic Modelling and Knowledge Processing Department
Institute of Information and Communication Technologies, BAS
Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.org

Abstract

In the paper we present a methodology for the creation of a LOD dataset over domain documents. The domain is European Law in its connection to national laws. The documents are interlinked over the web. They are linked also to other Linked Open Data datasets (such as GeoNames). The data includes five languages: Bulgarian, English, French, Italian and German. Thus, the paper discusses the first step towards the creation of a domain corpus with linguistically linked documents, namely - the reference linking among the documents and their linguistic processing.

Keywords: Domain document corpus, LOD, Linking, Multilinguality

1. Introduction

Many domains are represented by a set of documents which need to be available to the professionals in the area of their daily tasks. The access to the documents is of importance to the domain specialist. On the other hand, such a set of analyzed documents is a valuable language resource for the domain. The processing of legal documents for information retrieval, law studies and legal reasoning has a long history in the area of natural Language processing. Recently, legal processing employed the inventory of semantic technologies — ontology, ontological modeling, linked data. Here we mention a few of works in the area. The extraction of semantic information from law documents is presented in (Biagioli et al., 2005). The paper discusses some work in the area of provisions – modeling and extraction from documents. The Provision argument extraction is performed on the basis of document processing on syntax and semantic level. In this respect we have applied similar techniques for processing of texts. (Mimouni, 2013), (Kunkel, 2015) and (Casellas, 2012) discusses the modeling of legal vocabularies, thesauri and documents via ontologies and linked data. The common part of the works presented in these papers is the relation to language resources in the law domain, their usage for document processing and modeling the results by means of the semantic technology. Here we present a related work and discuss its relevance to Linguistic Linked Open Data.

The paper introduces the methodology and technical details behind the EUCases[1] Legal Linked Open Dataset (EUCases-LeLOD) and Web Interface Querying EUCases Linking Platform. Methodologically, the EUCases-LeLOD consists of a set of ontologies and RDF triples. EuroVoc and Syllabus are in the center of the model, since they are used as domain specific ontologies. Additionally, other supporting ontologies have been added, such as GeoNames for the named entities; PROTON as an upper ontology (Terziev et al., 2005), (Kiryakov, 2006); SKOS as a mapper between ontologies and terminological lexicons; Dublin Core as a metadata ontology. For an efficient reasoning, the so-called FactForge reason-able view was explored as an approach to linked data management.

Technically, EUCases-LeLOD is represented in RDF graphs and uses the SPARQL query language. The input documents have been encoded into the legal XML schema Akoma Ntoso. Thus, the schema has been converted into an appropriate RDF representation for the purposes of linking. For its Web Interface the EUCases Linking Platform relies on the customized version of the GraphDB Workbench, developed by Ontotext AD. In the underlying repository engine reasoning and query evaluation are performed over a persistent storage layer. Loading, reasoning and query evaluation proceed extremely quickly even against huge ontologies and knowledge bases. More specifically, for the project purposes GraphDB Standard Edition was selected as a scalable semantic platform. The address of the Web Interface to the EUCases Linking Platform is http://graphdb.eucases.eu/.

The approach taken for the creation of EUCases-LeLOD could be extended also to other collections of documents.

Our motivation for presenting this linked legal dataset of documents is as follows: we view the process of linking as a gradual one – first, the professionals and stakeholders are usually interested in referenced documents. However, even the step of correct referencing requires NLP processing. Thus, the linguistic information is in the data, but it remains hidden. We can imagine how many datasets like this have been produced in various domains. For that reason, as second, we would like to 'save' the linguistic information in providing further linking on paragraph, sentence, phrase and word level. After this the resource might be used also for linguistic research in the domain, construction of comparable corpora, extraction of parallel expressions, etc.

In this paper we focus on the first step in the context of the future linguistic linking.

The structure of the paper is as follows: in the next section the EUCases Document Modeling process is described. It includes the ontology modeling and the RDF representation. Section 3 discusses the reason-able view over the dataset. The last section concludes the paper.

[1]"EUCases: Linking Legal Open Data in Europe" is an European project: http://www.eucases.eu

2. EUCases Document Modeling

Each EUCases document is an XML document valid with respect to the legal XML schema Akoma Ntoso[2]. Here we present the main elements of the header of each EUCases document and its mapping to the ontologies selected for EUCases-LeLOD.

There are two types of EUCases documents: Act (covering legislative documents) and Judgment (covering case law documents). The structure of each EUCases document is divided into two: metadata and content. The metadata determines the type of the document, its life cycle, including the date of creation, the author(s), the place of creation, keywords, abstract, etc. The content of the document is the actual text of the document. The RDF representation of each EUCases document encodes information coming from both sources — the metadata and the content. The metadata is already represented explicitly during the creation of the XML representation of the document. The content of the document is described on the basis of the text annotation of the document via NLP and the linking tools developed within the project. Each annotation is represented as having a body and a target. The target of each annotation a document whose content is annotated with the corresponding information. The body is the additional information provided by the annotation. This additional information is represented as a URI pointing to a geopolitical entity described in the GeoNames dataset; an external document represented via the URL of the document; and/or a concept from an annotation ontology (in the project we exploit EuroVoc Thesaurus and Legal Taxonomy Syllabus).

2.1. Namespaces for EUCases project

All created instances of EUCases documents, metadata elements and annotations are represented as instances in the EUCases dataset. In order to represent them uniformly, we use for all EUCases instances the following namespace:

```
@prefix eucinst:
   <http://www.eucases.eu/lod/instances#> .
```

In our work we model all the necessary information using existing ontologies, but in case of necessity to extend some of the ontologies, new classes and properties will be defined within the following namespace:

```
@prefix eucont:
   <http://www.eucases.eu/lod/ontology#> .
```

The usages of these name spaces are given below.

2.2. Ontology modelling

In this section we describe the main classes of EUCases ontology exploited within the project to represent conceptually the EUCases documents. We start with PROTON ontology and extend it with a definition of new classes and properties where necessary or via mapping to the other ontologies mentioned above. In EUCases-LeLOD we represent the two types of documents: acts and judgments. For each of them we encode information coming from metadata section of the document and the content of the document.

[2]http://www.akomantoso.org/

The EUCases document is modelled as a sub-class of the PROTON class `ptop:Document`:

```
ptop:Document
  rdf:type owl:Class ;
  rdfs:comment "The information content of
                any sort of document.
                The tangible aspects are
                ignored. It is usually
                a document in free text
                with no formal structure
                or semantics."@en ;
  rdfs:label "Document"@en ;
  rdfs:subClassOf ptop:InformationResource .
```

`ptop:Document` has the following important for EUCases-LeLOD properties (directly defined for it or inherited): `ptop:documentAbstract`; `ptop:documentSubTitle`; `ptop:derivedFromSource`; `ptop:hasContributor`; `ptop:hasDate`; `ptop:hasSubject`; `ptop:inLanguage`; `ptop:informationResourceCoverage`; `ptop:informationResourceIdentifier`; `ptop:informationResourceRights`; `ptop:resourceType`; `ptop:title`. These properties comply with Dublin Core ones.

The EUCases document has two sub-classes for acts and for judgments. Fig. 1 represents the hierarchy of the documents.

On the level of `eucont:EUCDocument` class we define all the properties that are necessary for the representation of EUCases documents and that are shared by the two subclasses of documents — acts and judgments, for example `eucont:reference` which has domain and range `eucont:EUCDocument`. The properties specific for `eucont:Act` and `eucont:Judgment` are defined locally.

Document identifiers can be of several types: Akoma Ntoso identifier, National identifier, EUCases identifier, ELI identifier. Each identifier is an URI pointing to the different representation of the document. The first identifier is used as instance identifier of the document.

Language of the document is represented as language code with respect to ISO 639-2 standard. The language is represented by DC property `dcterms:language` which is mapped to PROTON property `ptop:inLanguage`. Behind the actual language of the document we also represent the original language of the document via the property `eucont:originalLanguage`.

A EUCases document could have several titles. The full title is represented by the DC property `dcterms:title`. The other types of title (short title, abbreviation, colloquial title) are represented by EUCases specific properties: `eucont:shortTitle`, `eucont:abbreviation`, `eucont:colloquialTitle`. All of them are subproperties of the PROTON property `ptop:title`.

The EUCases document type is represented by the DC property `dcterms:type` which is defined to be a subproperty of PROTON property `ptop:resourceType`.

The history of a EUCases document is presented as a sequence of events (or states) like document cre-

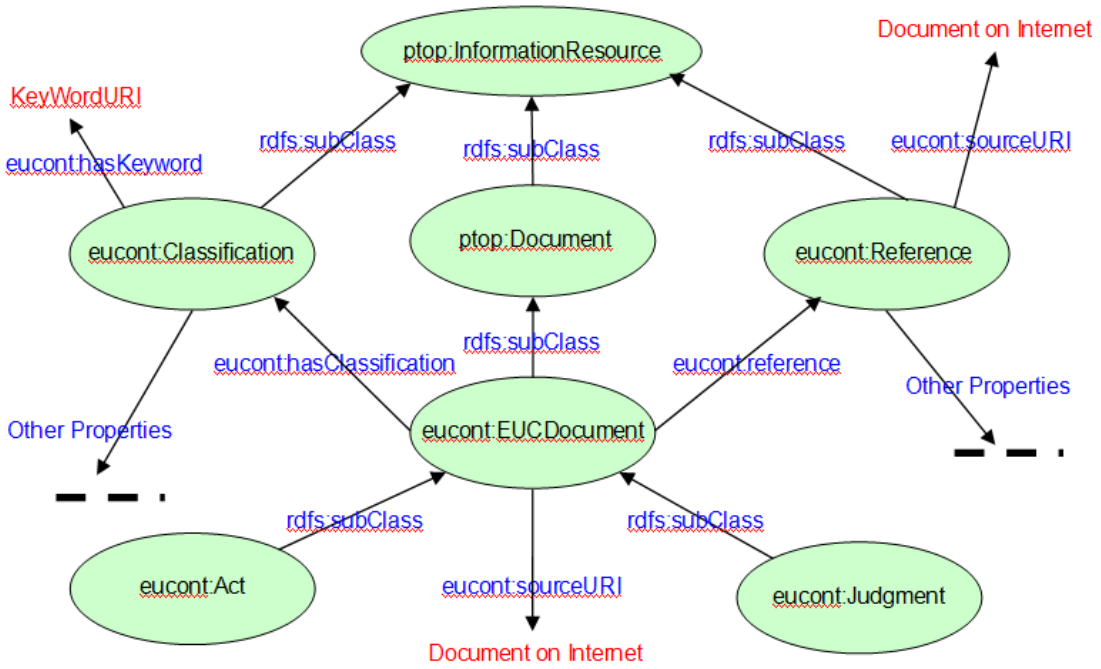


Figure 1: Ontology of EUCases documents.

ation, document publication, document signature, etc. Events and states are modelled by the PROTON class `ptop:Happening`. Each happening is determined by its beginning and end time moment as well as the participants. For example, the document creation is done in some period of time, by some legislative authority; a change of document also can be done by legislative authority in a given time interval. Although such document related events (or states) are well perceived by the users of EUCases-LeLOD, their representation in many cases is partial or unnecessarily complex. In order to avoid the complex nature of the document events in the current version of the EUCases ontology we encode only time stamps for some of the events. These time points or intervals are encoded by several EUCases properties whose names reflect the document events that happened at a given moment or interval. Such properties are `eucont:documentDate` - the date of the creation of the document; `eucont:publicationDate` - the date of the publication of the document; `eucont:effectDate` and `eucont:validityDate` - determine the period in which the document is in force. The time properties are subproperties of the PROTON property `ptop:hasDate`. In many cases the events and their participants are unique. In these cases the participants in the events can be also expressed by appropriate properties like `eucont:hasPublisher`. If in future it is necessary to extend the representation of provenance of EUCases documents to more detailed descriptions ontologies like PROV can be exploited.

Classification of a EUCases document is done via a set of keyword references. The keyword references point to terms in a thesauri like EuroVoc or Legal Taxonomy Syllabus in our case. Such a classification has a source which can be some real agent - Person or Organization; or software agent like EUCases NLP ToolKit. Each classification is represented by the EUCases class `eucont:Classification`. The class eucont:Classification is a subclass of the PROTON class ptop:InformationResource. The property `eucont:hasSource` represents the source of the classification. A classification is attached to a document by the property `eucont:hasClassification`. The property `eucont:hasKeyword` connects a classification with its keywords. Each keyword is an instance of the class `eucont:Keyword` which is a subclass of the PROTON class `ptop:Topic`. The SKOS class `skos:Concept` is a subclass of `eucont:Keyword`. Thus, using SKOS representation of EuroVoc thesaurus we can use EuroVoc terms as keywords in the EUCases classifications. The definition of a classification in EUCases ontology is given in Fig. 2.

On Fig. 3 the definition of a reference is given. We divide the reference in two parts: web address of the referred document and internal address. The web address is represented as the property `eucount:sourceURI`. It refers to the document on the web. Depending on the format of the web document the reference could be to the whole document or to an internal part of the document. When document can not be addressed internally via an URI, then the internal address is represented via some of the other properties defined in the ontology.

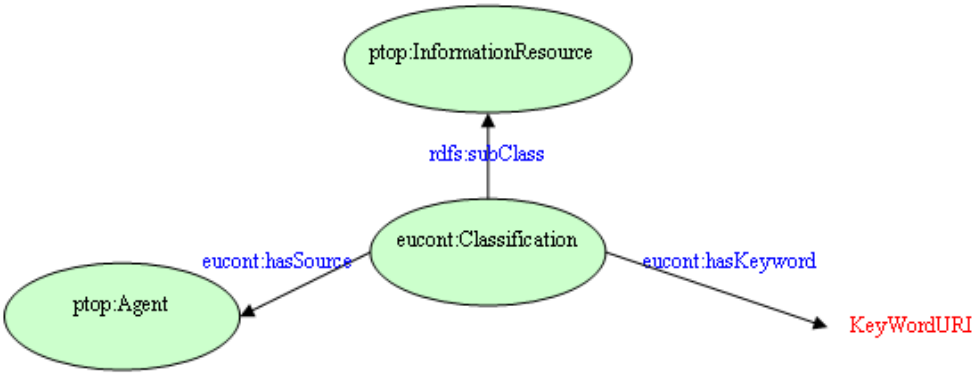


Figure 2: Definition of class Classification.

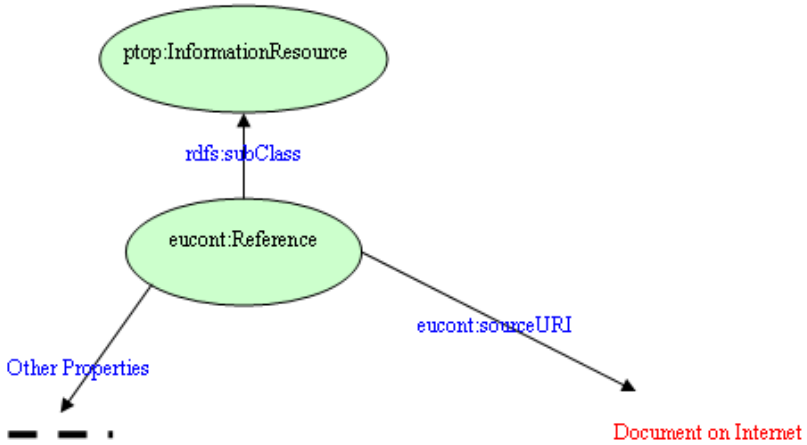


Figure 3: Definition of class Reference.

Geographical characteristics of the EUCases documents are modelled in two ways. First, the metadata geographical features determine the area in which the document is in force. These features are modelled by DC property dcterms:coverage with `rdfs:range` restricted to PROTON class `pext:PoliticalRegion`. The second kind of geographical information is extracted automatically from the content of the document. This kind of information is represented by property `eucont:hasLocationReference` which is also restricted to take instances of the concept `pext:PoliticalRegion` as values.

The summary of an EUCases document is extracted automatically by the EUCases NLP ToolKit. It is modelled by the DC property `dcterms:abstract`.

In order the different ontologies to be exploited together we also provide mapping rules between them. The mapping rules are expressed by RDFS properties `rdfs:subClassOf` and `rdfs:subPropertyOf`. Here we provide some examples of such rules.

Mapping from PROTON ontology to GeoNames ontology

The namespace for GeoNames is:

```
@prefix geo-ont:
     <http://www.geonames.org/ontology#> .
[ rdf:type owl:Restriction ;
   owl:onProperty geo-ont:featureCode ;
   owl:hasValue geo-ont:A.ADM4 ]
   rdfs:subClassOf   pext:PoliticalRegion .
```

The first part of the rule (with the brackets []) expresses the concept of political region in the terms of GeoNames. The same concept is modelled within PROTON as the class pext:PoliticalRegion. The PROTON Ontology is provided with a full set of such mapping rules to cover the whole GeoNames dataset.

Mapping from PROTON ontology to EUCases ontology

EUCases specific classes and properties are represented as an extension of the PROTON ontology. The rules for the mapping between them are of the following types:

```
eucontLEUCDocument
     rdfs:subClassOf ptop:Document .
```

```
eucont:shortTitle
     rdfs:subPropertyOf ptop:title .
```

The same kind of rules are used for mapping between the other ontologies. We are using subclasses and subproperties statements for the mapping in order not to enforce the equivalence between PROTON classes, properties and statements and the other ontologies. In this way we believe that the resulting EUCases-LeLOD dataset will not depend too much on PROTON ontology and, if necessary, other ontologies can be used instead of PROTON. In the next section we describe the procedure for converting EUCases documents from Akoma Ntoso XML representation into a set of RDF triples with respect to the EUCases ontology.

2.3. Creation of RDF representation of the EUCases documents

The process for RDFization of a given EUCases document comprises the following steps:

1. Reading and validation of an XML document. Each document in the EUCases database is represented in XML with respect to Akoma Ntoso legal XML schema. A module reads the document and checks its validity.

2. For each element of the document which provides information for the RDF representation one or more appropriate new XML elements are added.

3. For each added triple element the module computes the subject, predicate and object URIs. In some cases object can be a literal with a concrete value: text, number, date, etc.

4. The triples with defined subject, predicate and objects are extracted from the document and converted into actual RDF Triples.

5. The created sets of RDF triples for a given document are loaded into the EUcases RDF repository. During this process the new RDF triples are checked for consistency with the rest of the EUCases Dataset; new information following from the inference rules is also added to the repository.

During the addition of the RDF representation to the RDF repository the annotations from the document resolve their bodies to the corresponding instances that are already loaded into the repository, such as GeoNames instances, EuroVoc terms or Syllabus classes. Here are some examples of rules.

Creation of document instance
XML representation:

```
<FRBRthis value="32001r0044/main" />
```

RDF representation:

```
eucinst:32001r0044/main
    rdf:type  eucont:EUCDocument .
```

Language of the document
XML representation:

```
<FRBRlanguage language="bul" />
<FRBRlanguage eId="#OriginalLanguage"
                              language="bul" />
```

RDF representation:

```
eucinst:32001r0044/main
dcterms:language "bul"@en ;
eucont:originalLanguage "bul"@en   .
```

Document classification
XML representation:

```
<classification source="#nlp-toolkit">
   <keyword dictionary="eurovoc" eId="4622"
       showAs="transport user"
       value="eurovoc/4622/" />
   <keyword dictionary="eurovoc" eId="195"
       showAs="airport"
       value="eurovoc/195/" />
   <keyword dictionary="eurovoc" eId="497"
       showAs="damage"
       value="eurovoc/497/" />
</classification>
```

RDF representation:

```
@prefix eurovoc:
    <http://eurovoc.europa.eu/> .
eucinst:32001r0044/main
    eucont:hasClassification
    eucinst:32001r0044/classification001 .
eucinst:32001r0044/classification001
    rdf:type
        eucont:Classification .
eucinst:32001r0044/classification001
    eucont:hasSource
    eucinst:nlp-toolkit .
eucinst:32001r0044/classification001
    eucont:hasKeyword eurovoc:4622 ;
    eucont:hasKeyword eurovoc:195 ;
    eucont:hasKeyword eurovoc:497 ;
    eucont:hasKeyword eurovoc:1339 .
```

Here the actual terms in different languages will be available by the SKOS representation of EuroVoc.

3. Reason-able View over EUCases-LeLOD

The exploitation of linked data for data management is considered to have a great potential. On the other hand, several challenges need to be handled in order to make this possible. Reason-able views (Kiryakov and Momtchev, 2009) represent an approach for reasoning with and management of linked data defined at Ontotext and implemented in two systems, namely, FactForge (http://factforge.net) and LinkedLifeData (http://www.linkedlifedata.com). FactForge is based on general world knowledge LOD datasets like the DBPedia dataset. LinkedLifeData is domain oriented and it is used to support biomedical research. In the project we exploit FactForge as a source of background world knowledge. Reason-able view is an assembly of independent datasets, which can be used as a single body of

knowledge with respect to reasoning and query evaluation. The key principles can be summarized as the following instruction:

- Group selected datasets and ontologies in a compound dataset;
- Clean up, post-process and enrich the datasets if necessary. Do this conservatively, in a clearly documented and automated manner, so that (i) the operation can easily be performed each time a new version of one of the datasets is published and (ii) the users can easily understand the intervention made to the original dataset;
- Load the compound dataset in a single semantic repository and perform inference with respect to tractable OWL dialects;
- Define a set of sample queries against the compound dataset. These determine the level of service or the scope of consistency contract offered by the reason-able view. Each reason-able view is aiming at lowering the cost and the risks of using specific linked data datasets for specific purposes. The design objectives behind each reason-able view are as follows:
  - Make reasoning and query evaluation feasible;
  - Lower the cost of entry through interactive user interfaces and retrieval methods such as URI auto-completion and RDF search (a search modality where RDF molecules are retrieved and ranked by relevance to a full-text style query, represented as a set of keywords);
  - Guarantee a basic level of consistency — the sample queries guarantee the consistency of the data in the same way regression tests guarantee the quality of the software;
  - Guarantee availability — in the same way web search engines are usually more reliable than most of the web sites; they also do caching;
  - Easier exploration and querying of unseen data — sample queries provide re-usable extraction patterns, which reduce the time for getting to know the datasets and their interconnections.

The reason-able view is important for EUCases project as an approach to linked data, because the linked data extracted from the EUCases documents will interact with other LOD datasets. Our goal is to support the consistency of the domain specific data as much as possible. Cases of contradictory information will be also useful to the business scenarios of EUCases because they would result from different sources and opinions. The implementation of reason-able view is done within GraphDB Workbench using the RDF repository and inference supported by GraphDB.

In order to model the data extracted from EUCases documents and to construct a reason-able view, we have considered several ontologies to be put together:

- EuroVoc[3] and Syllabus are domain modeling ontologies which are used for the annotation of the content in the EUCases documents;
- GeoNames[4] ontology describes the structure of GeoNames LOD dataset;
- Dublin Core[5] provides a vocabulary for description of document metadata;
- PROTON[6] is a general ontology. It plays the important role of a joined ontology for the reason-able view;
- SKOS is a metaontology for mapping lexicons and ontologies.

SKOS is used in the distribution of EuroVoc for supporting lexicons for several languages aligned to the EuroVoc term identifiers. One alternative to SKOS ontology is Lemon Ontology[7]. We did not exploit Lemon for aligning the lexicons due to the following reasons: (1) SKOS is good enough for the goals of EUCases project; (2) The conversion and maintenance of EuroVoc is not our task and it is better to be done by the developers of EuroVoc; (3) We expect a converter from SKOS to Lemon ontology to be implemented soon. The usage of Lemon model will allow the representation of more linguistic features of the domain terms in the lexicon. These will include grammatical features, alternative forms, internal structures, discontinuity of the phrases. Also Lemon model provides links to other linguistic standards. This is useful in the process of creation of Linguistic LOD from document-based factual LOD datasets.

The linking between the different ontologies and LOD Datasets was implemented in two major ways:

- Ontology mappings
- Instance mappings

The rules for ontology mapping were given above in Sect. 2.2.. The ontology mappings ensure the usage of a single ontology for querying the LOD dataset. More on the different types of mapping and some problems can be found in (Simov and Kiryakov, 2015).

The instance mappings are provided by the NLP pipelines implemented in EUCases NLP pipelines. Besides the standard elements like tokenizators, POS taggers, lemmatization, the NLP pipelines for EUCases include modules like EUROVOC annotator and Geonames annotator. Both of them are working in two steps. First the candidate annotations are identified in the text of the documents. Then a procedure for selection of the correct annotation are applied. For example, for the Geonames annotation the following rules are applied. If there is not other indication we assume that the geolocation is in Europe. In this way we rule out many ambiguous cases that are not in Europe. The document specific location indicators are defined in different

[3] http://eurovoc.europa.eu/
[4] http://www.geonames.org/
[5] http://dublincore.org/
[6] http://www.ontotext.com/proton
[7] http://lemon-model.net/

contexts: the country of publication of the document, the last mentioned location which contains the current mention of a location, the current sentence, the current phrase.

4. Conclusion and Future Work

In the paper we presented a modeling over a set of documents in the legal domain with respect to LOD. The linking between the documents with other LOD datasets was provided via ontology and instance mappings. The ontology mappings have been done manually and the instance mappings have been done automatically via the respective NLP pipelines. The created LOD dataset is represented as a reason-able view and loaded in the Graph DB RDF repository. In this way, it is made available for professionals in the domain area.

In our view, the creation of Linked Open Data via NLP processing is and will be the main approach to the task. We consider this as a basis for creation of Linguistic LOD datasets. The document-based LOD dataset is a valuable Linguistic LOD dataset, since it allows for exploration of a corpus of domain documents linked to different vocabularies and gazetteers. But much more linguistic knowledge is made explicit during the creation of the dataset. This knowledge usually is lost after the project as much as it is not of interest to the professional users of the domain LOD dataset. In our case the NLP analyses are available and we are planning to add them to the RDF representation of documents. Availability of NLP analyses in addition to reference and factual data will support applications like studying of domain language, implementation of better processing pipelines, finding of comparable documents, extraction of parallel sentences and translation lexicons, extraction of domain phraseology and many others. The creation of LLOD from the linguistic analyses of the domain documents needs also usage of appropriate linguistic ontologies.

In future, we plan to annotate the documents with WordNet synsets. This will allow even better exploration of the data on semantic level. It will be useful also for professional users because it will provide them with possibilities for searching with everyday concepts, not just the domain terminology. Also WordNet is mapped already to many other Linked Open Data datasets — factual and linguistic. Thus, the annotation with WordNet synsets will provide additional mappings to other datasets.

5. Bibliographical References

Biagioli, C., Francesconi, E., Passerini, A., Montemagni, S., and Soria, C. (2005). Automatic semantics extraction in law documents. In *The Tenth International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 6-11, 2005, Bologna, Italy*, pages 133–140.

Casellas, N. (2012). Linked legal data: A SKOS vocabulary for the code of federal regulations. *Semantic Web Journal*.

Kiryakov, A. and Momtchev, V. (2009). Two reason-able views to the web of linked data. In *Proceedings of the Semantic Technology Conference 2009*, San Jose.

Kiryakov, A. (2006). Ontologies for knowledge management. In *Semantic Web Technologies: Trends and Research in Ontology-based Systems*, pages 115–138. John Wiley & Sons.

Kunkel, R. (2015). Using skos to create a legal subject ontology for defense agency regulations. *Legal Reference Services Quarterly*, 34(4):324–337.

Mimouni, N. (2013). Modeling legal documents as typed linked data for relational querying. In *Proceedings of the First JURIX Doctoral Consortium and Poster Sessions in conjunction with the 26th International Conference on Legal Knowledge and Information Systems, JURIX 2013, Bologna, Italy, December 11-13, 2013.*

Simov, K. and Kiryakov, A. (2015). Accessing linked open data via a common ontology. In *Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data*, pages 33–41, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Terziev, I., Kiryakov, A., and Manov, D. (2005). Base upper-level ontology (bulo) guidance. *Deliverable 1.8.1, SEKT project*.

Pinpointing the Eye of the Hurricane — Creating A Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data

Andrea Salfinger*, Caroline Salfinger*, Birgit Pröll†, Werner Retschitzegger*, Wieland Schwinger*
*Dept. of Cooperative Information Systems, †Inst. for Application Oriented Knowledge Processing,
Johannes Kepler University Linz,
Altenbergerstr. 69,
4040 Linz,
Austria
*{andrea.salfinger, caroline.salfinger, werner.retschitzegger, wieland.schwinger}@cis.jku.at, †bproell@faw.jku.at

Abstract

Crisis management systems would benefit from exploiting human observations of disaster sites shared in near-real time via microblogs, however, utterly require *location information* in order to make use of these. Whereas the popularity of microblogging services, such as Twitter, is on the rise, the percentage of GPS-stamped Twitter microblog articles (i.e., tweets) is stagnating. Geo-coding techniques, which extract location information from text, represent a promising means to overcome this limitation. However, whereas geo-coding of news articles represents a well-studied area, the brevity, informal nature and lack of context encountered in tweets introduces novel challenges on their geo-coding. Few efforts so far have been devoted to analyzing the different types of geographical information users mention in tweets, and the challenges of geo-coding these in the light of omitted context by exploiting situative information. To overcome this limitation, we propose a *gold-standard* corpus building approach for evaluating such situative geo-coding, and contribute a human-curated, geo-referenced tweet corpus covering a real-world crisis event, suited for benchmarking of geo-coding tools. We demonstrate how incorporating a semantically rich Linked Open Data resource facilitates the analysis of types and prevalence of geo-spatial information encountered in crisis-related tweets, thereby highlighting directions for further research.

Keywords: Corpus Building, Geo-parsing, Geo-coding, Toponym Resolution, Social Media

1. Introduction

Social Media for Crisis Management. Nowadays, timely situational update information on crisis events can frequently be retrieved from social media, such as eyewitness reports of disaster sites shared via microblogging[1] (Olteanu and others, 2015). Whereas initial prototypes have already demonstrated the potential of incorporating social media data, such as Twitter microblog articles (i.e., tweets), in emergency management systems (cf. surveys in (Imran and others, 2015; Salfinger and others, 2015a)), these utterly depend on *location information*, as emergency managers and first responders ultimately need to know *where* their assistance is required. Therefore, the actual sparsity of GPS-tagged tweets (Gelernter and Mushegian, 2011; Imran and others, 2015; Schulz and others, 2013) represents a major bottleneck for exploiting social media data for crisis management, requiring additional means to utilizing GPS tags to obtain essential location information. Apart from locations actually associated with the tweet's author (the user's current location and location profile (Ikawa and others, 2013)), location information can also be extracted from its *textual content*.

Extracting Location Information. However, the detection of place names in such free-form text (i.e., *toponym recognition* or *geo-parsing*), and the mapping of a place name to its corresponding geographic location by assigning appropriate coordinates (i.e., *disambiguation*, *toponym resolution*, *geo-coding* or *(spatial) grounding*) (Leidner, 2007; Martins and others, 2005), represents a non-trivial task due to both, *geo-/non-geo-ambiguity*[2], as well as *geo-/geo-ambiguity*[3] (Amitay and others, 2004). Whereas traditional news articles frequently allow to resolve these ambiguities, since these provide their reader with contextual information required to understand the situation described therein, the ultimate brevity and real-time nature of tweets introduce unprecedented challenges on their geo-coding: Tweets are typically written in informal, localized language and expose specific characteristics — for instance, location information may also be obscured in multi-word hashtags (e. g., "#Hawaiihurricane", "#bigislandoutage"). To meet the imposed length limitations[4], tweets frequently lack discourse context, as humans tend to omit contextual information that is shared between the correspondents[5]. Thus, this *context deletion* represents a severe obstacle for an automated extraction of (otherwise valuable) situational update information from social media. Consequently, geo-referencing of tweets needs to stretch beyond conventional topoynm recognition and resolution, as developed for news prose (e.g., (C. D'Ignazio and others, 2014; Leidner and Lieberman, 2011; Lieberman, 2012; Quercini and others, 2010; Samet and others, 2014)) and longer web documents

[1] Examples of microblogging platforms include Twitter (www.twitter.com), Tumblr (www.tumblr.com) and Sina Weibo (www.weibo.com).

[2] For example, "Jordan" may refer to a basketball player or a country.

[3] For example, "Sydney" may refer to a city in Australia or Canada.

[4] Up to 140 characters per tweet.

[5] For instance, Vieweg et al. could identify several tweets where people simply referred to "the river" when actually meaning the "Red river" in their studies of tweets on the Colorado flooding events in 2009 (Vieweg and others, 2010).

(such as Wikipedia or online news, e.g., (Amitay and others, 2004; Woodward and others, 2010)).

Disambiguation Context. Although geo-coding approaches fine-tuned towards the characteristics of tweets have been developed (Flatow and others, 2015; Gelernter and Balaji, 2013; Ikawa and others, 2013; Karimzadeh and others, 2013; Schulz and others, 2013), current approaches provide limited account for this *context deletion*: During the development of our social media-sensing Situation Awareness system for crisis management (Pröll and others, 2013; Salfinger and others, 2015b; Salfinger et al., 2016a; Salfinger et al., 2016b), we encountered many tweets that were not appropriately resolved by presently available geo-coding tools. From our empirical observations, we noted that such toponym resolution errors frequently could be attributed to the common error of not incorporating sufficient *context* for toponym disambiguation, which can be classified into the following two context classes: (i) "in-tweet-context", i.e., unambiguous toponym disambiguation *within* a single tweet is possible based on the joint context of all location mentions occurring in this tweet, and, (ii) "between-tweets-context" or "situative context", which refers to the event-level context of the monitored scenario - i.e., the associated event-context would allow to derive valuable disambiguation cues guiding toponym resolution, as proposed in (Salfinger et al., 2016b).

Ground-truth Data Sets. In order to examine and systematically study the challenges of toponym disambiguation, however, a *ground-truth data set* would be required, which reflects the way a human monitoring the crisis scenario would resolve encountered location descriptions by incorporating contextual reasoning. Although valuable work on corpus-building of geo-parsed and/or geo-referenced tweet corpora have been undertaken (which we will review in Sec. 2.), these mainly focus on general toponym recognition aspects, such as identification of proper place names (Wallgrün and others, 2014), or detection of locative expressions without considering the mapping of these to real-world locations (Liu and others, 2014). Little focus so far has been on studying toponym disambiguation problems, especially from the social media-specific *context deletion* perspective. Therefore, we set to systematize and share our experiences by creating a human-curated *ground-truth* data set suited to study such geo-coding challenges from a crisis management perspective.

Linked Open Data. However, the creation of such a shareable evaluation data set for toponym resolution tasks is complicated by the inter-dependency between the employed geographical reference frame, i.e., the topographical information used to determine the mapping from textual entities to geographical space, and the resulting geo-referenced corpus. Thus, corpora created with different geographical reference frames may not be directly comparable to each other (e.g., due to different toponym resolution granularity) (Leidner, 2006). Recently, however, the growth in Linked Open Data (LOD) initiatives provides a remedy towards this problem: Geographical ontologies, such as GeoNames[6], represent a semantically rich, comprehensive and global-coverage source of geographical knowledge, providing an extensive basis for geographical reference, and tend to become the de-facto standard for geographical reference sources utilized in geo-parsing tools (Wallgrün and others, 2014).

Contributions. Therefore, we introduce a gold-standard corpus building methodology involving publicly available annotation tools and LOD to create shareable language resources (LRs) for studying situative toponym resolution, and report on the resulting corpus building initiative. We propose an event-driven corpus sampling strategy to allow for incorporating situative context, an annotation schema involving a LOD resource which also comprises annotation types for assessing implicitly specified geographical information, describe the developed annotation process, and contribute the resulting human-curated, geo-referenced gold standard tweet corpus on a specific crisis event for benchmarking and training of geo-coding techniques. We further outline how the semantic richness of the employed LOD resource benefits the analysis of the resulting corpus, by examining the types and prevalence of geo-spatial information encountered in this corpus from a crisis management perspective. We specifically also assess *implicit* and qualifying geo-spatial information to outline which potentially valuable spatial cues could be exploited for crisis management applications, but which remain unused by presently available geo-coding tools, thereby indicating directions for further research. This is further underpinned by a comparative evaluation of state-of-the-art geo-parsing tools on this data set, which highlights current performance limitations. We hope that our proposed methodology encourages similar initiatives in creating sharable LRs supporting the analysis of situative geo-coding.

Structure of the Paper. In the next section, we compare our approach to related endeavors on gold standard corpus building for geo-coding purposes. In Sec. 3., we describe the set-up of our collaborative annotation project, before analyzing the resulting *gold standard corpus* in Sec. 4., and concluding our lessons learned in Sec. 5.

2. Related Work

In this section, we explain how our gold standard corpus creation extends valuable findings reported in other work. We first assess related tweet corpora, before discussing more widely related work on news corpora.

Social Media. Gelernter and Mushegian describe the building of a geo-annotated tweet corpus on the 2011 earthquake in Christchurch, New Zealand (Gelernter and Mushegian, 2011), thus, focusing on a specific crisis event, as in our study. Whereas they defined a location upon a diverse set of types (such as countries, buildings, street addresses) and also incorporated hashtags and abbreviations, as well as generic places, i.e., non-proper place names (e. g., "city", "house", "home"), they did not devise a specific annotation scheme for discriminating these types in order to study the distribution of encountered types, as in our approach. They neither did include place names being part of multi-word tokens, which we included to examine the frequency of place names encountered in multi-word hashtags.

Wallgrün et al. employed a crowd-sourcing approach to

[6] http://www.geonames.org

Table 1: Comparison of highly related approaches. Abbreviations: ? = not stated, K = 1000

	Annotations				Corpus Characteristics		
Approach	Toponym Recognition	Locative Expressions	Toponym Resolution	Employed Geographical Gazetteer	Event-specific	Annotators/Message	Volume
(Gelernter and Mushegian, 2011)	✓	?	?	—	✓	3	1.4K
(Liu and others, 2014)	✗	✓	✗	—	✗	3	1K
(Wallgrün and others, 2014)	✓	✗	planned	GeoNames	✗	5	6K
this work	✓	✓	✓	GeoNames	✓	3	4K

create a geo-annotated tweet corpus, and provided an extensive discussion of encountered annotator errors (Wallgrün and others, 2014). 6K tweets have been annotated for identified place names in a crowd-sourcing project on the Amazon Mechanical Turk platform, which, as opposed to our approach, were not confined to a specific event, but sampled according to different criteria. In the present work, we base upon their findings by incorporating their characterization of annotator errors into the definition of our annotation schema. Furthermore, Wallgrün et al. proposed to employ the GeoNames ontology for toponym resolution, which they planned to address in future work. Following their suggestion, our annotation schema thus encodes manually resolved toponyms by their Geonames identifiers (IDs). However, whereas Wallgrün et al. solely focused on proper place names, our annotation schema also involves annotation types for resolving implicitly stated geospatial information, since we also aimed at detecting implicit or vague spatial information in order to quantify the proportions and types of implicit information encountered in crisis-related tweets.

Liu et al. focused on the annotation of Locative Expressions (LEs) on corpora of different web document types (e.g., Twitter, Blogs, Youtube comments) (Liu and others, 2014), i.e., any expressions referring to a location (such as "in my cozy room", "at home" or "around the city"). Their manually annotated corpora provided the basis for comparing Precision, Recall and F-score of six different geoparsers. Their focus, however, has been on entity recognition, i.e., identifying the text chunks comprising LEs, not on their actual geo-coding (i.e., mapping to geographic coordinates).

The data sets for evaluating the geo-coding techniques proposed in (Flatow and others, 2015; Schulz and others, 2013) use the GPS locations of the user's device as geo-reference. However, the user's current location may be disparate from the *focused location* (Ikawa and others, 2013) of the tweet, i.e., the location the user writes about, which is actually the location of interest in our crisis management application domain.

News Articles. Extensive studies on toponym resolution have been conducted by J.L. Leidner, however, with a focus on news prose (Leidner, 2006; Leidner, 2007). The two human-curated gold standard datasets created in the course of this work (Leidner, 2006) therefore consist of news articles obtained from the REUTERS Corpus Volume I.

Since toponym recognition and resolution also represent core algorithmic tasks for Geographical Information Retrieval Systems (Geo-IR), the need for standardized evaluation procedures and appropriate benchmarking data sets also led to corpus building efforts in this research domain (cf. (Martins and others, 2005) for an overview), however, with a focus on newswire texts (e. g., Geo-IR evaluation tracks GeoCLEF[7] 2005, 2006, 2007 and 2008).

3. Methodology

In the present section, we describe our gold-standard corpus building methodology, for which we followed the best practice guidelines on collaborative annotation projects suggested in (Sabou and others, 2014).

Scenario. Due to our application domain of crisis management, we pursued an event-driven approach for corpus sampling, by assembling a corpus characterizing a specific real-world crisis event. Our initial tweet corpus has been retrieved with the aim of monitoring the effects of hurricanes Iselle and Julio on the Hawai'ian islands, in August 2014[8]. We recorded tweets matching keywords associated with that crisis[9] from the public Twitter Stream[10], yielding roughly 212 600 tweets collected between August the 9th to 21st, 2014. This event-driven approach allows us to study the challenges of geo-locating tweets within a real-world crisis context, as opposed to open-domain geo-coded corpora, such as created in (Wallgrün and others, 2014). We can thus specifically examine whether the studied tweets also contain *context-sensitive* geo-spatial information, i.e., information which cannot be interpreted if the general event context is lacking, thus making it impossible even for human annotators to understand. The selected data set conforms to a highly-localized event, involving small-scale locations on the Hawai'ian islands. Hawai'i furthermore proves to be challenging with respect to (w.r.t.) toponym resolution, since it comprises many ambiguous locations (i. e., multiple locations with the same name exist on different islands, e. g., Wailea, Wailua) which need to be properly resolved to aid crisis management tasks.

Corpus Sampling. In order to optimize the allocation of human work force, we designed a dedicated data preprocessing protocol to narrow down the data set to a manageable yet representative proportion of the collected tweets, and eliminate near-duplicate tweets. We restricted our data set to English-language tweets (provided by the tweet's language tag) in the time range between Aug., 9th - 16th, 2014, resulting into 137K tweets, 83K of those were actually textually distinct. In the first step, *background knowledge* regarding the monitored events was employed in or-

[7] http://ir.shef.ac.uk/geoclef

[8] www.latimes.com/nation/nationnow/la-na-nn-hawaii-storm-iselle-juliio-20140808-story.html

[9] tracked by a filter query leaving language and location deliberately unspecified and the following keywords: Hurricane, #HurricaneIselle, #HurricanePrep, #hiwx, #HIGov, Iselle, #updatehurricaneiselle, #Genevieve, #Iselle, #Julio

[10] Twitter Streaming API: https://dev.twitter.com/streaming/public

der to reduce the data set to presumably disaster-relevant tweets. This is due to the fact that keyword-based queries frequently return tweets not related to the disaster, i.e., in which the corresponding term is used in a different context (e. g., "#Mystery, #Romance #Humor a #Hurricane' what more could you want! http://t.co/13el5JKptR @rpdahlke #Bargain 99"). We also included several *location* terms, a-priori known to be crisis-relevant in the chosen scenario, in this initial filtering[11] to guarantee a high number of geo-referenceable tweets. Furthermore, we wanted to investigate user-generated content (i. e., tweets ideally written by human on-site observers), as opposed to the plethora of tweets containing news headlines, which refer to news articles and external web sources, since our major goal was studying the characteristics of social media and not - unintentionally - examining news prose. According to our empirical analysis, tweets referring to and advertising external content (e.g., consisting of news headlines and a URL to the corresponding news agency) tend to correlate with specific Twitter clients[12] (e. g., IFTTT and Hootsuite, a social media marketing tool for enterprises). We therefore filtered the tweets on Twitter clients that, according to our experience, more likely contain original content[13]. By focusing on content sent from mobile devices, we thus seek to increase the proportion of content published by end-users with a non-commercial focus. We further noted that even after filtering on textual distinctness, in a semantic sense, many duplicates remain. This is attributable to the different URLs generated from URL shorteners, which are commonly employed on Twitter to meet the strict length limitations. Thus, we receive many duplicates in terms of slightly modified and "broadcast" message content, such as "Pound of prevention' pays off for Hilo Medical Center during Iselle http://t.co/HSl9pX9JEI #hawaii" and "Pound of prevention' pays off for Hilo Medical Center during Iselle: Hilo Medical Center had to switch to gen... http://t.co/HHXodZT0O0". We therefore aimed at eliminating the effect of shortened URLs by replacing them with a specific token. Upon this URL-coding, we could discard tweets which have a too low string distance to other tweets, by using the `stringr` and `stringdist` R packages to filter out textually highly similar tweets (M.P.J. van der Loo, 2014).

Annotation Schema. We provided our annotators with a dedicated annotation schema for marking explicit and implicit spatio-temporal information encountered in tweets, and geo-referencing this information based on a semantically-rich LOD resource, the GeoNames ontology[14]. By linking to the corresponding ontology instances, we are not only able to unambiguously refer to a specific toponym and retrieve its geographic coordinates, but can also examine additionally provided geographic meta-data, such as a toponym's administrative division (allowing to discriminate coarse-grained — such as country-level — from fine-grained information, such as districts and villages). A screenshot showing the resulting annotation editor dialog is shown in Fig. 1. We incorporated the findings presented in (Wallgrün and others, 2014) into the definition of this annotation schema, which comprises the following annotation types:

◇ *Proper Place Name (PPN)* for marking named location entities, such as the names of populated places (i.e., countries, cities, etc.) or other geographical features (i. e., mountains, islands, etc.). This annotation type also includes several additional annotation features that should be specified, such as a free text field titled GeonamesID. By looking up recognized place names on the Geonames search interface, annotators should manually perform toponym resolution by identifying the appropriate location candidate from the resulting Geonames toponyms list, and enter its corresponding Geonames ID, which uniquely identifies this location. Since Geonames also allows a map-based inspection of its retrieved toponyms, annotators were encouraged to carefully analyze and disambiguate results. Furthermore, annotators were required to specify whether a location name is part of a single-word or multi-word hashtag (annotation feature "hashtag complete", e. g., "#hawaii", or "hashtag partial", e. g., "#bigislandoutage", respectively), whether it is used attributively (e. g., "One week later: This is how Hawaii Island residents have to live after #Iselle."), its name is specified informally (e. g., "#Iselle about to make landfall on Big Hawaiian Island."), or abbreviated (e. g., "Many still wo power in Puna District on Big Is."). Misspelled place names should be annotated (e. g., "Hawai"), but following (Gelernter and Mushegian, 2011; Wallgrün and others, 2014), we explicitly excluded place names part of an organization name (e. g., "Hawaiian Airlines") or a Twitter user handle (i.e., "mention", e. g., "@akeakamaihawaii"). Annotators should mark each occurrence of a PPN, even if specified multiple times in the same tweet.

◇ *Point of Interest (POI)* corresponds to distinctive locations that cannot be found on Geonames, but are known to a greater audience (e. g., "Iselle Relief: Plate lunches Available at Nanawale Community Center Today"), thus, mostly denote specific buildings or well-known spots.

◇ *Place Qualifier (PQ)* corresponds to a locative expression which further spatially restricts a given location (e. g., south California or upper Manhattan). Since, for crisis management applications, we are interested in the most fine-grained locative description possible, we are thus interested in examining such spatial restrictions, which would require spatial reasoning capabilities to be appropriately geo-coded in an automated fashion.

◇ *Non-proper Place Name (NPPN)* denotes general spatial

[11] Notably, filtering on the following terms: "Hawaii", "Pahoa", "Puna", "Kona", "Hilo", "iselle", "honolulu", "oahu", "maui", "kauai", "BigIslandOutage", "big island", "#HIwx", "HELCO"

[12] The Twitter client the tweet has been sent with can be retrieved from the tweet's meta-data.

[13] Twitter for iPad (http://twitter.com/#!/download/ipad), Twitter for iPhone (http://twitter.com/download/iphone), OS X (http://www.apple.com/), Twitter for Windows Phone (http://www.twitter.com), Twitter Web Client (http://twitter.com), Facebook (http://www.facebook.com/twitter), TweetDeck (https://about.twitter.com/products/tweetdeck), Twitter for Android (http://twitter.com/download/android), Twitter for Android Tablets (https://twitter.com/download/android), Instagram (http://instagram.com), Instagram (http://instagram.com), Mobile Web (M2) (https://mobile.twitter.com)

[14] www.geonames.org

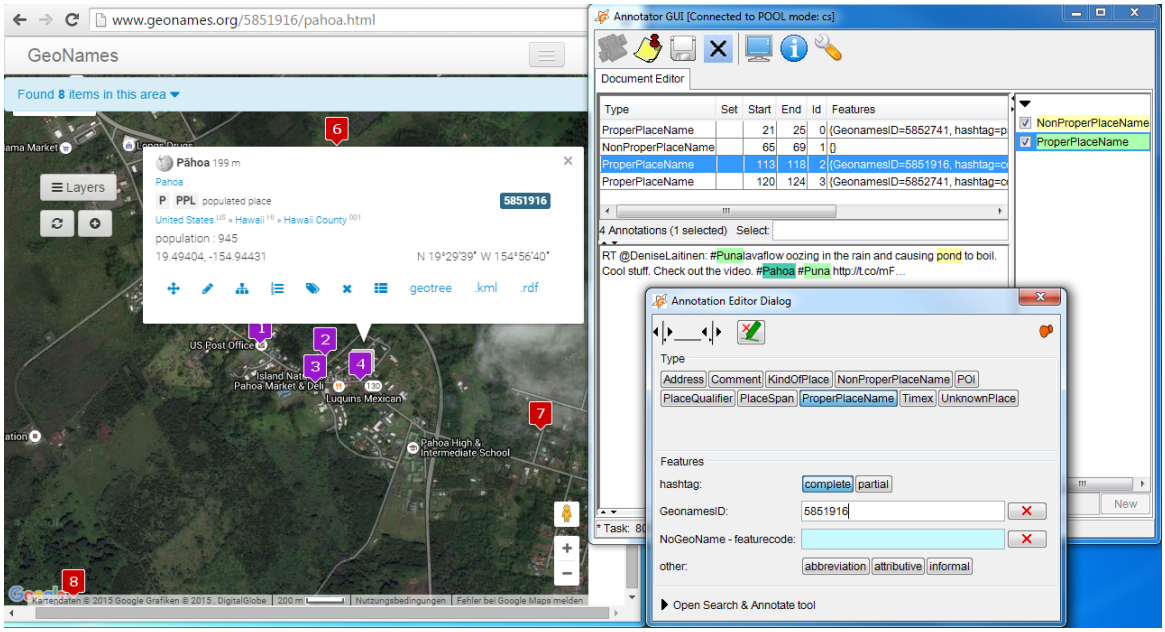


Figure 1: Screenshot showing the annotation task view.

descriptors (e. g., "south shore", "islands"), which lack an explicit identifier. However, these frequently correspond to the omission of discourse context encountered in tweets, and actually refer to specific locations known by the communicating people (cf. "the river" denoting the "Red river" example in Sec. 1.). Therefore, we demanded the annotation of such NPPNs, in order to examine whether the referred-to location could be inferred given the external *situative context* of the tweet (i.e., the general event context). Thus, if our annotators could infer the actual location due to their human intelligence and provided event context, the feature "Reference to Geoname ID" allowed to add the corresponding GeoNames ID. Thus, this annotation scheme provides a means to study geo-coding techniques for NPPNs which operate on techniques exploiting the situative context, as proposed in (Salfinger et al., 2016b), which sets our approach apart to other approaches discussed in Sec. 2..

◇ *Kind of Place (KOP)* represents a redundant description to a given PPN (e.g., "city of London"). As discussed in (Wallgrün and others, 2014), these are frequent causes for annotator disagreement, as it is difficult to discriminate in which cases a KOP expression is actually a part of the PPN itself. As Wallgrün et al. discussed, this is more frequently the case for physical features, such as mountains and lakes (e. g., "Lake Michigan", for which "lake" is a part of the PPN, in order to discriminate it from the state of Michigan), than for populated places (e. g., regarding "city of London", city is redundant information). In our guidelines, we were following the notion of (Wallgrün and others, 2014), i.e., KOP only corresponds to a separate annotation if it is clearly redundant, otherwise (e. g., if part of the PPN given on Geonames), the text should be included in the PPN annotation. Although we presented examples in the annotation guidelines to clarify on this, we still received many annotator disagreements, who often misconceived this annotation with others (such as NPPN and PQ).

◇ *Address (ADR)* comprises separate annotation features for marking street names, postcodes and house numbers.

◇ *Place Span (PSP)* represents a meta-annotation for marking a location span (e. g., "from Hilo to Pahoa"), in order to assess their frequencies.

◇ *Timex (TMX)* for annotating temporal expressions, which may be of sub-type "date", "time", or "other", if the previous two do not apply.

◇ *Comment (COM)* Annotators also were given the possibility of attaching their own, free-form remarks.

Project Execution. Regarding the technical setup, we employed the collaborative annotation platform GATE Teamware (Bontcheva and others, 2013), which provides a client-server-architecture for managing the set-up and distribution of collaborative annotation tasks. Annotators used a web interface to retrieve their assigned annotation tasks, which were executed using a GATE-based annotation editing interface (cf. Fig. 1) in conjunction with the GeoNames web interface. Regarding the assignment of annotation tasks, the sampled tweet corpus (in total 4 117 tweets) has been partitioned across twelve human annotators, who were conducting these tasks in the course of a summer internship at our institution (age range 15 - 19, two females, ten males). Our annotators had an educational background at high school level and were non-native English speakers, but have been learning English for at least five years, therefore, had a solid language level[15]. In an initial preparation meeting, our annotators have been presented with the required

[15]Place name identification has been recognized as relatively easy annotation task also for non-local users (Gelernter and Mushegian, 2011).

background information regarding the events covered in the data sets, i.e., have been given a summary on the key events, in order to understand the scope and situative context of the assigned tweets. We feel the option of realizing such introductory meetings presents an advantage of such lab-study based annotation projects over the use of online crowdsourcing platforms, as it allows the introduction of more complex annotation tasks by first providing the annotators with essential background knowledge. We also sought to address the frequently reported unfamiliarity problem (Gelernter and Mushegian, 2011; Wallgrün and others, 2014) (i. e., non-local annotators may overlook place names since they do not know the corresponding text fragment represents a location name, due to their unfamiliarity with the corresponding location), by pointing the annotators towards relevant geographic characteristics, locations and their popular abbreviations of the event site Hawai'i, thereby increasing their geographic awareness. Furthermore, annotators were given a set of guidelines regarding the devised annotation schema, involving detailed screenshots and examples. The annotation tasks were introduced by means of a small *pilot study*, in which annotators could get acquainted with the annotation interface (shown in Fig. 1) by experimenting with example tasks, and were encouraged to ask questions, before we assigned the actual annotations tasks. We demanded at least three annotators per tweet, thus corresponding to a total annotation effort of 12351 tweets. To avoid introducing any group bias, we split the entire data set into several batches, and permuted group composition of the three annotators allocated per batch across the different batches.

Data Evaluation. We assessed the annotators' agreement using GATE's Inter-Annotator Agreement plugin, measuring Precision, Recall and F1 in a strict sense. Whereas we received good Inter-Annotator Agreement for PPNs (F1: mean: 0.74, standard deviation: 0.07, measured strictly and incorporating equality of the specified Geonames ID), and acceptable results regarding the TMX (F1: mean 0.46, standard deviation: 0.05)), the agreement regarding the other annotation types was insufficient[16]. Thus, we can confirm the findings presented in (Wallgrün and others, 2014), who classified these annotation types into the most common cause of annotator errors, and furthermore, can show that even when provided with an annotation schema and guidelines addressing these error sources (as suggested by Wallgrün et al.), humans face difficulties in reaching an agreement w.r.t. the corresponding type. It also appears that human annotators have mainly focused on the detection of PPNs, as all other type have been frequently overlooked, which may explain the fact that other work solely focusing on the annotation of temporal expressions reported higher F1 scores.

Corpus Delivery. For ultimately aggregating the different annotators' mark-ups to the final gold standard data set, the following steps were performed: First, a majority voting component copied these annotations to the *consensus set*, if a majority of annotators agreed strictly (for which we required that the annotation span was equal, i.e., not overlapping, and *all* annotation features were equal). Second, one of the authors performed manual adjudication of the remaining annotations, using the GATE Developer Tool: By comparing the annotators' opinions using the annotation stack tool, the adjudication manager resolved conflicting cases by copying the correct annotation to the consensus set, annotating overlooked entities or merging differing annotations.

[16]F1, mean: NPPN: 0.10, PQ: 0.12, PSP: 0.44, POI: 0.15, KOP: 0.06

4. Discussion

In the present section, we outline how the semantic richness of the employed LOD resource enables a fine-grained analysis of the resulting corpus, by providing additional metadata allowing for a faceted analysis.

4.1. Characteristics of the Resulting Corpus

The resulting geo-referenced tweet corpus is publicly available for research purposes[17], in the widely used GATE document XML serialization format[18]. We furthermore also provide lists of the encountered annotated texts and their frequencies of identified PQs (mostly corresponding to orientation relations, such as cardinal directions), NPPNs and POIs, as well as the proposed annotation schema.

Finding the Needle in the Haystack. For 99% of PPNs, a corresponding GeoNames toponym could be identified, yielding in total 244 unique identified GeoNames references. However, whereas this may, at first sight, seem to benefit applications such as crisis management, an inspection of the most frequent toponyms in Tab. 2 also highlights disguised challenges: The - by far most frequent - toponym refers to the entire state of Hawai'i, which clearly is expected. For crisis management applications, however, this information is of limited use, as a more detailed localization of affected areas - such as severely hit cities and villages - would be required, which we indeed encounter on rank 3, 6 and 10 (Pahoa and its surrounding Puna District have been damaged the by hurricane). Thus, the *granularity* of provided spatial information (in terms of their corresponding administrative division — e.g., state-level information versus city-level information) should ideally be attributed with corresponding weights, rewarding highly localized information (e.g., Pahoa) with higher priority fur further processing than area-/country-level information (e.g., State of Hawai'i). However, this also induces the challenges on how to track such information on Twitter, which in times of such crisis is flooded by corresponding news headlines from all over the world, which, however, mostly contain coarse-grained information (e.g., that Hawai'i is threatened by a hurricane), but provide limited value for actual crisis management tasks.

Need for Hashtag Decomposition. 13% of PPNs are obscured in multi-word hashtags, thereby requiring geoparsers capable of extracting the toponym chunks from these.

[17]https://weizenbaum.tk.jku.at/owncloud/public.php?service=files&t=6076c0c9b7f3e03fc6204b1607a8b0e1

[18]Including the final, adjudicated gold-standard annotations, the annotations of each individual annotator, and the results obtained with analyzed tools, for reasons of reproducibility.

Table 2: Most frequent Toponyms.

Rank	Place Name	Geo Names ID	Freq.
1	State of Hawai'i	5855797	1996
2	Island of Hawai'i	5855799	482
3	Puna District	5852741	412
4	Maui County	5850871	152
5	O'ahu	5851609	134
6	Hilo	5855927	99
7	Kaui County	5848514	74
8	Honolulu	5856195	67
9	Hawaiian Islands	5855811	66
10	Pahoa	5851916	57

Table 3: Annotation Type Distribution.

Anno. Type	Total	Freq.	Feature	Total	Freq.
PPN	4177	67%			
			GeoN. ID	4155	99%
			Hashtag	1300	31%
			- complete	771	18%
			- partial	529	13%
			other	959	23%
TMX	1113	18%			
NNPN	500	8%	Ref. to Geo. ID	62	12%
PQ	202	3%			
POI	165	3%			
ADR	25	0%			
KOP	23	0%			
PSP	9	0%			

Low Frequencies of Other Spatial Information. Regarding annotation types other than PPN and NNPN, we observe low frequencies in this dataset. The rare occurrences of KOP annotations may be attributable to the length restrictions imposed by Twitter, as the limit of 140 characters per tweet probably forces users to eliminate redundant information such as KOP. However, the scarcity of qualifying spatial information (PQ), fine-grained spatial information such as POIs and ADR (which will most likely be provided by local users familiar with the geographical situation), and specification of place spans, demands further investigation. Intuitively, one would expect these types of spatial information to correlate with the provision of fine-grained situational update information (e.g., which areas may be severely affected, at which addresses shelters would be provided etc.). Therefore, further studies involving different crisis datasets would be required to analyze whether the observed low frequencies are attributable to the sampling strategy employed in the generation of the current corpus, or these annotation types are indeed generally rarely observed in crisis-specific data sets.

Need for Situative Context-Aware Toponym Resolution. Discriminating the most frequent toponyms, i.e., rank 1 and 2 in Tab. 2, represents a major challenge, since both are typically referred to by the text "Hawai'i" in the tweet, but correspond to different toponyms and spatial granularity: Rank 1 comprises the entire group of islands, whereas rank 2 solely denotes the largest Hawaiian island, making a key difference for crisis management purposes. Therefore, toponym resolution techniques capable of reasoning on the current situative context to extract the adequate toponym are required.

4.2. Benchmarking of Geoparsers

Ultimately, the most interesting question is how well existing geo-referencing tools perform on this ground-truth data set. We thus examined the performance of advanced state-of-the-art systems (C. D'Ignazio and others, 2014), notably CLAVIN-NERD[19], and GeoTxt[20] (Karimzadeh and others, 2013), cf. Tab. 4, both capable of resolving toponyms based on the GeoNames ontology. Both tools are built for recognizing and resolving PPN annotations only, therefore, the following experiments solely evaluate their performance on detecting and resolving PPN annotations. Since CLAVIN-NERD does not provide support for parsing hashtags, we thus preprocessed the tweet texts by replacing "#" tokens with blanks, which should — at least — enable it to resolve single-word hashtags accordingly. Following (Martins and others, 2005), we provide a separate evaluation of *toponym recognition* and *toponym resolution*, to pinpoint performance lacks to the corresponding phase. Whereas these tools yield high Precision, Recall is below 50%, thus, the majority of geo-spatial information actually contained in tweets (from a human's perspective) remains unused.

Toponym Resolution Errors. We furthermore analyzed the most frequently incorrectly disambiguated toponyms, cf. Tab. 6. As assumed in Sec. 4.1., the resolution of small-scale locations tends to be problematic, which, however, is crucial for application domains such as crisis management. To examine this assumption, we conducted another experiment to separately evaluate the tools' performance on such small-scale locations, by excluding annotations corresponding to a populated place of an administrative division 1 and 2, as provided by the Geonames ontology, which indeed yields a severe drop in Recall (cf. Tab. 4). A closer analysis of the mapped locations suggests that incorporating a geo-spatial reasoning aware of the situative context could potentially improve toponym resolution, as several of the toponyms selected by these tools are located highly disparate from the actual event location (notably, are located even at different continents).

5. Conclusions and Lessons Learned

In the present work, we contributed a geo-referenced, manually curated tweet corpus, described the employed corpus building methodology, and provided an analysis of the resulting corpus. We examined the availability and prevalence of geospatial information in tweets from the requirements perspective of a crisis management application, thereby identifying several research challenges for future work. Our evaluation of state-of-the-art geo-parsing

[19]https://clavin.bericotechnologies.com, https://github.com/Berico-Technologies/CLAVIN-NERD

[20]http://www.geotxt.org

Table 4: Corpus statistics, **A** = gold standard data set, **B** = results obtained with the geo-referencing tool listed in the left-most column.

Tool (**B**)	Match (Correct)	Only **A** (Missing)	Only **B** (Spurious)	Overlap (Partial)	Prec. **B/A**	Rec. **B/A**	F1	Match (Correct)	Only **A** (Missing)	Only **B** (Spurious)	Overlap (Partial)	Prec. **B/A**	Rec. **B/A**	F1
	All PPN annotations.							**Small-scale PPN annotations.**						
	Toponym Recognition — F1.0-score strict on PPN annotations, without considering Geonames ID.													
CLAVIN-NERD	1963	2052	221	162	**0.84**	0.47	0.60	618	1235	175	58	**0.73**	0.32	**0.45**
GeoTxt Stanford h	2234	1700	441	243	0.77	**0.53**	**0.63**	640	1162	397	109	0.56	**0.33**	0.42
GeoTxt Gate h	no results retrieved													
	Toponym Resolution — F1.0-score strict on PPN annotations, incorporating Geonames ID.													
CLAVIN-NERD	1727	2431	600	19	**0.74**	0.41	0.53	408	1499	439	4	**0.48**	0.21	**0.30**
GeoTxt Stanford h	2023	2136	877	18	0.69	**0.48**	**0.57**	453	1452	687	6	0.40	**0.24**	**0.30**

Table 5: Toponym Resolution errors, GS = gold standard data set, F. = Frequency.

GS	Clavin-Nerd	F.
Island of Hawai'i (5855799), HI, US	Big Island (4747418), Virginia, US	90
Island of Hawai'i (5855799), HI, US	Republic of Estonia (453733)	27
Puna District (5852741), HI, US	Pune, India (1259229)	20
Kailua-Kona (5847504), HI, US	Cona, Italy (3178217)	11
Island of Hawai'i (5855799), HI, US	Hawaii, FL, US (6463769)	7

Table 6: Toponym Resolution errors, GS = gold standard data set, F. = (Total) Frequency.

GS	GeoTxt Stanford h	F.
Puna District (5852741), HI, US	Pune, India (1259229)	64
Pacific Ocean (2363254), HI, US	Pacific, MO, US (4402300)	28
Island of Hawai'i (5855799), HI, US	Big Island (4747418), Virginia, US	18
Kailua-Kona (5847504), HI, US	Cona, Italy (3178217)	13
State of Hawai'i (5855797), HI, US	Hawaii, FL, US (6463769)	6

tools' performance on our gold standard corpus revealed that further research on tackling the specifics of tweets is utterly needed, as current tools provide unsatisfactory Recall, especially regarding small-scale locations. Thus, only a fraction of geo-spatial information can be used at the moment, hindering valuable use cases for Twitter data, such as benefiting crisis management. Since Recall can only be measured given a comprehensive ground truth data set, we therefore hope that the contribution of our gold standard corpus may aid in the development of effective location entity recognition and geo-coding techniques for tweets.

Naturally, our current gold standard corpus is limited in terms of generalizability, since only a single crisis event is covered and we only incorporated English-language tweets. For future work, we thus seek to extend our corpus building endeavor towards other crisis events and languages, allowing to further examine potential country- or language-specific characteristics in social media usage. By devising and describing a corpus building methodology involving publicly available annotation tools and LOD resources, we hope to encourage other research groups to join these efforts in creating shareable, inter-operable LRs for studying situative geo-coding, similarly to related efforts for collaboratively created ground truth data sets for examining social media characteristics across crises, such as the extensive CrisisLex26 data set for informativeness classification of crisis-related tweets (Olteanu and others, 2015).

6. Acknowledgements

This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FFG BRIDGE 838526, FFG TALENTE 850859, 852071, 850855 and 850594, and WTZ AR10/2015. We further thank Paula Maria Fischer, Theresa Gierlinger, Josef Haider, Philipp Hofer, Paul Huber, Leo Jungmann, Felix Kastner, Erik Reichl, Lorenz Reichl, Lukas Riegler, Maximilian Schartner and Martin Stoiber for performing the annotation tasks.

7. Bibliographical References

Amitay, E. et al. (2004). Web-a-where: Geotagging Web Content. In *Proc. of the 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 273–280. ACM.

Bontcheva, K. et al. (2013). GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

C. D'Ignazio et al. (2014). CLIFF-CLAVIN: Determining geographic focus for news. In *NewsKDD: Data Science for News Publishing*.

Flatow, D. et al. (2015). On the Accuracy of Hyper-local Geotagging of Social Media Content. In *Proc. of the Eighth ACM Int. Conf. on Web Search and Data Mining*, WSDM '15, pages 127–136. ACM.

Gelernter, J. and Balaji, S. (2013). An Algorithm for Local Geoparsing of Microtext. *Geoinformatica*, 17(4):635–667.

Gelernter, J. and Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, 15(6):753–773.

Ikawa, Y. et al. (2013). Location-based Insights from the Social Web. In *Proc. of the 22Nd Int. Conf. on World Wide Web Companion*, WWW '13 Companion, pages

1013–1016. Int. World Wide Web Conferences Steering Committee.

Imran, M. et al. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys*, 47(4).

Karimzadeh, M. et al. (2013). GeoTxt: A Web API to Leverage Place References in Text. In *Proc. of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pages 72–73. ACM.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11.

Leidner, J. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, 30(4):400–417.

Leidner, J. L. (2007). *Toponym resolution in text*. Ph.D. thesis.

Lieberman, M. D. (2012). *Multifaceted Geotagging for Streaming News*. Ph.D. thesis, College Park, MD, USA.

Liu, F. et al. (2014). Automatic Identification of Locative Expressions from Social Media Text: A Comparative Analysis. In *Proc. of the 4th Int. Workshop on Location and the Web*, pages 9–16. ACM.

Martins, B. et al. (2005). Challenges and Resources for Evaluating Geographical IR. In *Proc. of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pages 65–69. ACM.

M.P.J. van der Loo. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1):111–122.

Olteanu, A. et al. (2015). What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proc. of the 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 994–1009. ACM.

Pröll, B. et al. (2013). crowdSA - Crowdsourced Situation Awareness for Crisis Management. In *Proc. of Social Media and Semantic Technologies in Emergency Response (SMERST)*.

Quercini, G. et al. (2010). Determining the Spatial Reader Scopes of News Sources Using Local Lexicons. In *Proc. of the 18th SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, GIS '10, pages 43–52. ACM.

Sabou, M. et al. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proc. of the Ninth Int. Conf. on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*, pages 859–866.

Salfinger, A. et al. (2015a). Crowd-Sensing Meets Situation Awareness - A Research Roadmap for Crisis Management. In *Proc. of the 48th Annual Hawaii Intl. Conf. on System Sciences (HICSS-48)*.

Salfinger, A. et al. (2015b). crowdSA - Towards Adaptive and Situation-Driven Crowd-Sensing for Disaster Situation Awareness. In *Proc. of IEEE Int. Multi-Disciplinary Conf. on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA 2015)*.

Salfinger, A., Retschitzegger, W., Schwinger, W., and Pröll, B. (2016a). Towards a Crowd-Sensing Enhanced Situation Awareness System for Crisis Management. In Galina L. Rogova et al., editors, *Fusion Methodologies in Crisis Management*, pages 177–211. Springer International Publishing.

Salfinger, A., Schwinger, W., Retschitzegger, W., and Pröll, B. (2016b). Mining the Disaster Hotspots - Situation-Adaptive Crowd Knowledge Extraction for Crisis Management. In *Proceedings of the 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 219–225. IEEE.

Samet, H. et al. (2014). Reading News with Maps by Exploiting Spatial Synonyms. *Commun. ACM*, 57(10):64–77.

Schulz, A. et al. (2013). A Multi-Indicator Approach for Geolocalization of Tweets. In *Int. AAAI Conf. on Weblogs and Social Media*.

Vieweg, S. et al. (2010). Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '10, pages 1079–1088. ACM.

Wallgrün, J. O. et al. (2014). Construction and First Analysis of a Corpus for the Evaluation and Training of Microblog/Twitter Geoparsers. In *Proc. of the 8th Workshop on Geographic Information Retrieval*, GIR '14, pages 4:1–4:8. ACM.

Woodward, D. et al. (2010). A Comparison of Approaches for Geospatial Entity Extraction from Wikipedia. In *Semantic Computing (ICSC), 2010 IEEE Fourth Int. Conf. on*, pages 402–407.

lemonGAWN: WordNet Gaeilge as Linked Data

Jim O'Regan, Kevin Scannell, Elaine Uí Dhonnchadha

Trinity College Dublin, Ireland
Saint Louis University, Missouri, USA
Trinity College Dublin, Ireland
jaoregan@tcd.ie, kscanne@gmail.com, uidhonne@tcd.ie

Abstract

We introduce *lemonGAWN*, a conversion of WordNet Gaeilge, a wordnet for the Irish language, with synset relations projected from EuroWordNet. *lemonGAWN* is linked to the English WordNet, as well as the wordnets for four Iberian languages covered by Multilingual Central Repository (MCR), and additionally contains links to both the Irish and English editions of DBpedia.

1. Introduction

WordNet (Miller et al., 1990; Fellbaum, 1998) is a lexical database for English. It relates words to *lexical senses*, which represent different senses of those words, and groups those senses into *synsets*, and provides sets of relations between these synsets (additionally, a number of lexical relations are provided between senses). The synsets and their relations form a semantic graph of English.

Initially developed at Princeton to model psycholinguistic theories of human lexical memory, it has found uses in a number of areas, including various areas of natural language processing; its usefulness in several areas have lead to the creation of wordnets for several other languages, such as the wordnets in the EuroWordNet project (Vossen, 1998), typically linked to Princeton WordNet (PWN). Through these links, the relations can be projected, applying the (largely language independent[1]) semantic graph to the other language; the other uses of wordnet aside, this semantic graph, when connected to the lexical units of a language, is a valuable linguistic resource.

WordNet Gaeilge (described in section 2.) is a wordnet for Irish *(Gaeilge)*, linked to PWN. To make the data more accessible, we are making it available as linked data (section 3.); in addition to providing a pre-generated version of PWN's semantic graph, applied to Irish, we provide links to a number of other wordnets.

2. WordNet Gaeilge and LSG

WordNet Gaeilge is based on *Líonra Séimeantach na Gaeilge* (LSG), an Irish wordnet originally created in 2006 by Scannell[2]. The synsets in the LSG map to PWN synsets in a two-step process. The first step uses English "glosses" in the lexical database[3]. Where the English glosses are unambiguous, they are mapped directly: *stáplóir* is glossed as "stapler" and this lies in a unique PWN synset.

The second step disambiguates the remaining glosses using a sentence-aligned corpus of English texts and their Irish translations, for example, the word *bruach* which has "bank" as one of its glosses. Irish sentences containing *bruach* (or inflected forms, such as *bhruach*, *mbruach*, etc.) are extracted along with the corresponding English sentences. Some of the English sentences will contain the word "bank", and the additional context provided by these sentences is used to decide which is the correct sense of "bank" using standard techniques in word-sense disambiguation. To ensure enough data are available, the bilingual corpus is quite inclusive: the Irish words and their glosses are included, even though they do not form complete sentences, as the glosses alone are often sufficient to determine the correct sense. This fact is well-known to lexicographers, including Ó Dónaill (1977), who gloss words like *feileastram* with two ambiguous English words ("flag, iris") but with no fear of confusion.

Wordnet Gaeilge synsets	77814
Missing from PWN	28356
Missing Irish label	5936
Nouns	49889
Verbs	11548
Adjectives	15250
Adverbs	1127

Table 1: WordNet Gaeilge synsets

WordNet Gaeilge does not link directly to PWN synets, instead using the "sense keys" which identify lexical units, because for many words, the sense distinctions made by the Princeton lexicographers are too fine even for intelligent non-lexicographers to make reliably, and certainly too fine for statistical methods. In addition, there are many distinctions made in Irish that are not made in English (e.g. *dearg* ("red") vs. *rua* ("red", in reference to hair) in Irish would map to a single Princeton synset) and these are precisely the distinctions one does not want to give up in a monolingual Irish language resource. For these reasons, a separate layer – an "intermediate wordnet" – was added between Irish and English, with mappings in both directions. It's still really an English wordnet, but one that is tailored to the needs of Irish. de Bhaldraithe (1959) was very useful in constructing this; the senses given under each English word give a rough first approximation of the sense inventory of the intermediate wordnet.

[1] EuroWordNet, and other similar wordnet projects, use a set of relations that are modified specifically for language independence.

[2] See http://borel.slu.edu/lsg/

[3] These are usually one- or two-word definitions like those found in Ó Dónaill (1977).

LSG is developed as an Open Source project[4], available under the terms of the GNU Free Documentation License, as are the resources described in the present paper. Table 1 gives the current status[5] of synsets in WordNet Gaeilge.

3. Linked Data

Linked Data (Bizer et al., 2009) builds on the technologies of the Web, such as URIs and the HTTP protocol, as a means of creating typed links between data from different sources, using RDF (Resource Description Framework)[6]. The W3C outlines four rules for making data available:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

(Berners-Lee, 2011)

Linked Data has been embraced in recent years by creators of linguistic data as a means of overcoming various problems relating to the inter-operability of disparate data sources Chiarcos et al. (2012). *lemon* (McCrae et al., 2012) describes a model for "ontology lexica" that, among other lexical resources, is being used as the basis for several RDF conversions of wordnets, such as the Chinese WordNet (Lee and Hsieh, 2015).

3.1. Wordnets as Linked Data

PWN, in various versions, has been converted to RDF several times. van Assem et al. (2006) describes a conversion of PWN 2.0, provided by W3C, which has served a central role in the Semantic Web for several years. In our conversion of WordNet Gaeilge, we considered three conversions:

- McCrae et al. (2014) describe a version of PWN 3.1 using the *lemon* model. As the data is hosted by Princeton, this can be considered the canonical version.
- VU Amsterdam provide a conversion of PWN 3.0[7], following the W3C model van Assem et al. (2006).
- IULA Universitat Pompeu Fabra provide a *lemon*-based conversion[8] of Multilingual Central Repository (MCR) (González et al., 2012), a EuroWordNet-based collection of wordnets, including English (based on PWN 3.0), Spanish, Basque, Catalan, and Galician.

As Princeton's RDF is based on PWN 3.1, and not version 3.0 used as a basis for WordNet Gaeilge, we chose not to use it as the basis for our conversion at the present time: each new edition of PWN adds, merges, splits, and deletes synsets, and mapping between versions is non-trivial[9]. However, as it implements the W3C's Best Practices for Converting WordNets to Linked Data[10], we followed its example in how to model our conversion.

As Princeton's RDF is the canonical edition, we wish to both introduce links to it. Although a set of synset mappings from PWN 3.0 to PWN 3.1 is available, as WordNet Gaeilge is based primarily around word senses, not synsets, this is not a complete solution; on the other hand, updating the sense links affects the WordNet Gaeilge database, while a guiding goal in our RDF conversion was to not alter the underlying data.

The RDF conversion of WordNet Gaeilge is being performed in the context of a project to create an Irish-English machine translation system, based on the Apertium platform (Forcada et al., 2011). The Multilingual Central Repository (MCR), as it is based on EuroWordNet, includes links to EuroWordNet's Top Ontology (Vossen et al., 1998), which include classifications of nouns that are necessary to Irish-English translation, in particular "Human" (in English to Irish translation, to select the correct numeral) and "Occupation" (in Irish to English translation, to disambiguate between a subject location *(x is in his house)* and a subject attribute *(X is a teacher)*, which share the same syntactic structure in Irish, but have a different semantic structure in English, e.g., *tá sé ina theach* ("he is in his house") and *tá sé ina mhúinteoir* ("he is a teacher")[11]). For this reason, we chose MCR as the basis for our projection of synset relations. Table 2 contains the number of relations obtained through projection.

VU Amsterdam's conversion follows the W3C's edition of PWN 2.0, and uses the PWN-specific model of that edition. McCrae et al. (2014) make the case that *lemon*'s open model is more suited for interlinking with other resources, so we chose not to use it for our primary conversion. On the other hand, the closed model is more amenable to testing for constraint violations, so our scripts generate a second conversion following this model. In addition, the VU Amsterdam conversion includes semantic relations (Fellbaum et al., 2009) which are not available in other conversions, that further classify the derivational relations in PWN.

3.2. Linking to other resources

Our primary targets in generating links to other datasets have been to other wordnets, currently, VU and the five languages available as part of MCR. In addition, a number of other lexical resources include their own conversions of PWN, such as Lexvo (de Melo, 2013) and Uby (Gurevych

[4]https://github.com/kscanne/wordnet-gaeilge

[5]Git revision a7078d6173735107e838938b3a36360a4da6f9a7, 2015-09-10.

[6]http://www.w3.org/TR/rdf-concepts/

[7]http://datahub.io/dataset/vu-wordnet

[8]https://github.com/martavillegas/EuroWordNetLemon

[9]The website for Princeton's RDF claims to include synset identifiers from PWN 2.0 and 3.0, but at the time of writing, these were not functional.

[10]https://www.w3.org/community/bpmlod/wiki/Converting_WordNets_to_Linked_Data

[11]A further ambiguity exists with states, but it is less easily resolved using wordnet data.

has_hyperonym	14965
has_hyponym	14965
has_mero_madeof	151
has_mero_member	295
has_mero_part	1617
has_subevent	155
is_caused_by	55
is_derived_from	360
is_subevent_of	155
near_antonym	1818
near_synonym	2946
region	155
region_term	155
related_to	16590
see_also_wn15	1684
state_of	472
sumo_at	1364
sumo_equal	749
sumo_plus	25503
topConcept	79757
usage	160
usage_term	160
verb_group	250

Table 2: Synset relations obtained by projection from EuroWordNet (MCR).

et al., 2012), via its RDF conversion, lemonUby (Eckle-Kohler et al., 2014). Table 3 contains the number of sense links to other data sets, table 4 contains the number of synset links.

DBpedia (Auer et al., 2007), an effort to extract Linked Data from Wikipedia, has emerged as a central hub for Linked Data, due to the broad topic coverage of the underlying data. As well as the data from the English edition of Wikipedia, a number of internationalization chapters have been set up, to extract DBpedia data from various language editions of Wikipedia. A chapter for Irish[12] has been set up, though is not currently hosting the extracted data (that is, although data for Irish is available, the URIs it contains are not currently available via HTTP). We provide links to the Irish edition, in anticipation of their availability; we additionally provide links to the English edition of DBpedia, to be immediately useful.

The links to DBpedia were primarily generated via a set of mappings from Google's (now defunct) FreeBase[13]. Although these links were validated by humans, by presenting the PWN gloss and the Wikipedia page related to the FreeBase topic, a number of errors have crept in. In addition to that, the links date from 2012, and do not reflect changes made in either FreeBase or Wikipedia. As part of closing down FreeBase, Google made their data available to the Wikidata project; we plan to regenerate our links based on those which have been validated by Wikidata contributors as the data becomes available.

Logainm (Grant et al., 2013) is a bilingual Linked Data resource for Irish placenames. At present, we only have 5 links from WordNet Gaeilge to Logainm. This is perhaps due to the relatively low coverage of Irish placenames in PWN. We plan to investigate if further links can be found in the Irish-specific synsets.

VU PWN 3.0	97639
Lexvo	27254

Table 3: Word sense links to other data sets

VU PWN 3.0	37607
MCR (English)	97639 (34116)
MCR (Basque)	97639 (34116)
MCR (Catalan)	97639 (34116)
MCR (Spanish)	97639 (34116)
MCR (Galician)	97639 (34116)
lemonUby	37743
DBpedia (en)	7167
DBpedia (ga)	2197
Logainm	5

Table 4: Synset links to other data sets. The number of unique synsets, where relevant, is given in brackets.

4. Future work

The primary goal of future work around *lemonGAWN* is in making it available. Although it is planned to make the data available as "5-star Linked Open Data" (Berners-Lee, 2011), practical concerns have delayed this; in the meantime, the data is being provided via Github[14] under the same terms as WordNet Gaeilge. In addition, scripts used in preparing the data are also being made available[15].

EuroWordNet contains a number of synsets not present in PWN, as does WordNet Gaeilge. As there is likely to be overlap between these synsets, future work will focus on introducing links between them: where an English label is available in both wordnets, we will use the method outlined in section 2.; for the remainder, we will investigate using the other lexical resources available via Lexvo and lemonUby in a similar manner.

SentiWordNet (Baccianella et al., 2010) extends PWN for use in opinion mining and sentiment analysis, by attaching sentiment scores to each synset. We have projected these scores, in the same manner that other links were projected. Ongoing work aims at validating the resulting scores, for use as a sentiment analysis lexicon for Irish. Work on creating chatbots for Irish (Ní Chiaráin and Ní Chasaide, 2016) is incorporating this sentiment analysis lexicon.

Although much work on building wordnets focuses on synset-level relationships, PWN additionally provides lexical links, which provide more fine-grained information about particular words, such as derivational relationships, or connecting verbs to their particles. As a pilot for adding

[12] http://ga.dbpedia.org/

[13] Downloaded from https://code.google.com/p/mlode/downloads/list.

[14] https://github.com/jimregan/lemonGAWN.

[15] https://github.com/jimregan/gawnrdf.

derivational relationships to our conversion, we have focused on adding lexical antonyms; by selecting words from each synset and the antonymic synset, and checking for common prefixes indicative of negation, we have added an initial set of 275 lexical antonyms. Ongoing work in this area concentrates on extracting other derivational relationships, using pairs of affixes across WordNet Gaeilge and PWN, while future work will aim at extracting non-derivational lexical relationships, using corpus-based methods, based on the observation that words collocate with their antonyms.

References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives, 2007. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudr-Mauroux (eds.), *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 722–735.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani, 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC 2010*, volume 10.

Berners-Lee, Tim, 2011. Linked data-design issues (2006). http://www.w3.org/DesignIssues/LinkedData.html. [Accessed September 16th, 2015].

Bizer, Christian, Tom Heath, and Tim Berners-Lee, 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann (eds.), 2012. *Linked Data in Linguistics*. Springer.

de Bhaldraithe, Tomás, 1959. *English-Irish Dictionary: With Terminological Additions and Corrections*. An Gúm.

de Melo, Gerard, 2013. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*:7.

Eckle-Kohler, Judith, John P. McCrae, and Christian Chiarcos, 2014. lemonUby–a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, submitted. special issue on Multilingual Linked Open Data*.

Fellbaum, Christiane (ed.), 1998. *WordNet: An Electronic Lexical Database*. Wiley Online Library.

Fellbaum, Christiane, Anne Osherson, and Peter E. Clark, 2009. Putting Semantics into WordNets "Morphosemantic" Links. In Zygmunt Vetulani and Hans Uszkoreit (eds.), *Human Language Technology. Challenges of the Information Society*, volume 5603 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 350–358.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez Felipe Sánchez-Martínez, and Francis M. Tyers, 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

González, Aitor, Egoitz Laparra, and German Rigau, 2012. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC'12)*.

Grant, Rebecca, Nuno Lopes, and Catherine Ryan, 2013. Report on the Linked Logainm project. Technical report, Dublin: Royal Irish Academy and National Library of Ireland; Galway: NUI Galway.

Gurevych, Iryna, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth, 2012. Uby: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lee, Chih-Yao and Shu-Kai Hsieh, 2015. Linguistic Linked Data in Chinese: The Case of Chinese Wordnet. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*. Beijing, China: ACL.

McCrae, John, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al., 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

McCrae, John P., Christiane Fellbaum, and Philipp Cimiano, 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*. Reykjavik, Iceland: ELRA.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller, 1990. Introduction to WordNet: An On-Line Lexical Database. *International journal of lexicography*, 3(4):235–244.

Ní Chiaráin, Neasa and Ailbhe Ní Chasaide, 2016. Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for Irish. In *Proceedings of LREC 2016 (to appear)*.

Ó Dónaill, Niall, 1977. *Foclóir Gaeilge-Béarla*. Oifig an tSoláthair.

van Assem, Mark, Aldo Gangemi, and Guus Schreiber, 2006. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06), Genoa, Italy*.

Vossen, Piek, 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. *Computers and the Humanities*, 32(2-3).

Vossen, Piek, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters, 1998. The EuroWordNet Base Concepts and Top Ontology. Technical report, Paris, France.

Linking Verb Pattern Dictionaries of English and Spanish

Vít Baisa[1], **Sara Može**[2], **Irene Renau**[3]
[1]Masaryk University, Czech Republic, Brno
[2]University of Wolverhampton, United Kingdom
[3]Pontificia Universidad Católica de Valparaíso, Chile
xbaisa@fi.muni.cz,S.Moze@wlv.ac.uk,irene.renau@pucv.cl

Abstract

The paper presents the first step in the creation of a new multilingual and corpus-driven lexical resource by means of linking existing monolingual pattern dictionaries of English and Spanish verbs. The two dictionaries were compiled through Corpus Pattern Analysis (CPA) – an empirical procedure in corpus linguistics that associates word meaning with word use by means of analysis of phraseological patterns and collocations found in corpus data. This paper provides a first look into a number of practical issues arising from the task of linking corresponding patterns across languages via both manual and automatic procedures. In order to facilitate manual pattern linking, we implemented a heuristic-based algorithm to generate automatic suggestions for candidate verb pattern pairs, which obtained 80% precision. Our goal is to kick-start the development of a new resource for verbs that can be used by language learners, translators, editors and the research community alike.

Keywords: Corpus Pattern Analysis, Linked Data, Lexicography, Lexical Semantics, Bilingual resources

1. Introduction

This paper presents the results of a preliminary study in cross-linguistic pattern linking based on existing monolingual verb pattern dictionaries for English and Spanish, which are the outcomes of two separate research projects aiming to create freely available monolingual resources. The *Pattern Dictionary of English Verbs* (PDEV)[1] currently covers over 1,700 English verbs, whilst the *Pattern Dictionary of Spanish Verbs* (PDSV)[2] contains around 300 verbs (100 of which are currently available online). Both dictionaries were conceived as inventories of semantically motivated syntagmatic patterns, i.e. sentence structures and the semantic categorisation of the verb's arguments. Consider the example below:

1. [[Human | Institution]] avoids [[Eventuality]]
   Example: *The Government must avoid war.*

A common use of the verb 'avoid' has to do with a [[Human]] or an [[Institution]] trying to prevent an [[Eventuality]] from occurring. The capitalised words displayed between double square brackets are not lexical items, but 'semantic types', i.e. mnemonic labels that best describe the semantic features shared by the nouns that typically occur in a given argument slot. Syntactically, the verb occurs in a monotransitive construction. The observed sense of the verb 'avoid', i.e. *to prevent from occurring*, can only be activated by this specific combination of obligatory syntactic arguments (subject, direct object) and their corresponding semantic types. As a result, patterns allow us to unambiguously map word meanings onto their syntagmatic context, offering rich syntactic and semantic information about the verb's behaviour whilst providing exhaustive evidence from the corpus.

The present study represents our first attempt at linking equivalent verb patterns found in two or more languages. For instance, the pattern of the English verb *avoid* shown in 1 is equivalent to the following pattern exhibited by the Spanish verb *evitar*:

2. [[Human | Institution]] evitar [[Eventuality]]
   Example: *El Gobierno debe evitar la guerra.*

The two patterns are identical in that they are both transitive and use the same semantic categories ('semantic types') to describe their arguments. The meaning of both patterns is also the same. In this paper, we propose to match patterns English and Spanish verb patterns automatically by applying a heuristic-based algorithm that calculates the similarity between patterns. If successfully implemented, the algorithm will allow us to start building a bilingual lexical resource efficiently using PDEV and PDSV.

In recent years, multilingual lexical resources have been mushrooming all over the globe. Despite their coverage and suitability for different tasks and purposes, there resources have yet to successfully tackle the complexities of verb behaviour. A multilingual resource such as the one we propose here will have a number of potential applications in Natural Language Processing and language learning, and will provide empirically sound lexical data that can be used in theoretical and applied cross-linguistic studies.

The paper is structured as follows: Section 2. provides information on the theoretical and methodological background underpinning the proposed research and describes the two pattern dictionaries in more detail; section 2.2. features a short overview of related work in the field, and the following two sections focus the on the manual (Section 3.) and automatic (Section 4.) linking methods developed in this study. Finally, our plans for future are discussed in the Conclusion.

2. Background

2.1. Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) (Hanks, 2004a) is a corpus-driven technique that aims at mapping word meaning onto specific syntagmatic patterns exhibited by the target word

[1] www.pdev.org.uk
[2] www.verbario.com

in any type of text. Based on Theory of Norms and Exploitations (TNE) (Hanks, 2004b; Hanks, 2013), CPA aims at identifying patterns of normal usage (norms) and investigating the way the very same patterns are exploited creatively (exploitations) by means of in-depth, labour-intensive lexical analysis of corpus data. By doing so, it provides a window into the normal, every-day phraseology, which makes it particularly well-suited for both lexicographic and NLP tasks.

TNE and CPA are influenced by a large amount of cognitive, pragmatic and corpus linguistics studies interested in investigating how words interact in creating meaning and how this connection can be demonstrated using empirical data (see (Hanks, 2013) for a theoretical overview). CPA has been developed especially with lexicographical resources in mind, providing a solid alternative to 'classical', introspection-based analyses of meaning, which focus on words in isolation rather the way they behave in specific contexts. In CPA, meaning is pattern-based, not word-based. For instance, consider example 1 again:[3]

3. 1 [[Human | Institution]] avoids [[Eventuality]]
   2 [[Human | Animal]] avoids [[Physical Object]]

There are no syntactic differences between the two patterns - both are transitive, but the semantic types assigned to the subject and direct object do not match, hence the difference in meaning. More specifically, the first pattern refers to an action, process or state a human being or an institution tries to keep from occurring so that it does not affect them, whereas the second pattern refers to the reaction of a human or an animal trying not to physically interact with an object.

2.2. Multilingual Lexical Resources

The compilation of large, freely available multilingual lexical resources by means of linking pre-existing data has been gaining considerable traction in recent years, and justifiably so - once a monolingual resource has been created, it makes perfect sense to reuse and transform the data for different purposes. Bringing together compatible resources for different languages is particularly popular, as demonstrated by the existence of two major international projects in lexical analysis: WordNet, which allows researchers to connect and share their work through the Global Wordnet Association, see (Vossen, 2002),[4] and FrameNet (Fillmore and Baker, 2010),[5] whose infrastructure and data are being used by hundreds of researchers from all across the globe. PDEV and PDSV differ from the lexical resources developed in these two projects in that they do not share the same object of study: WordNet studies concepts linked to groups of verbs named *synsets*, FrameNet is centred around semantic frames, and CPA is corpus-driven and pattern-based. As a result, they can only be considered as complementary resources. As already pointed out, an important advantage of CPA is that it is particularly well-suited for verbs, as it allows researchers to perform fine-grained syntactic and semantic analysis of any verb's argument structure.

BabelNet[6], Omega Wiki[7], and Wiktionary[8] are other multilingual projects to be mentioned, which represent a step in the right direction in that they use word senses rather than words (or lemmas) to interlink the vocabulary of a number of different languages. Nevertheless, they lack an empirical basis, that is, they are not linked to corpus evidence.

Finally, in Language Learning, new tools are being created and offered online. A good example is Linguee[9], a tool that combines pairs of bilingual dictionaries in many languages with a parallel corpus showing the use of the target word in context. Another example is the Interactive Language Toolbox (Buyse and Verlinde, 2013), which was developed for second language learners. These are only two examples of how a bilingual or a multilingual dictionary can adapt to new technologies and users' needs and combine with non-lexicographical resources to provide an enhance user experience. Our proposal can be considered as a step in the same direction.

2.3. CPA Projects

The **Pattern Dictionary of English Verbs** (PDEV) is a publicly available resource developed in the DVC (Disambiguating Verbs by Collocation) project by Patrick Hanks' team at the University of Wolverhampton. The dictionary provides information on all the typical patterns associated with a verb, their definitions, and the corresponding corpus examples. For each verb, a corpus sample of 250 concordance lines is extracted from the British National Corpus (Leech, 1992), and tagged with pattern numbers using Sketch Engine (Kilgarriff et al., 2014). Depending on the semantic and syntactic complexity of the verb, the sample can be incrementally augmented to 500 or 1,000 concordance lines. Patterns are identified mainly through lexical analysis of corpus lines, complemented by the information found in the automatic collocations profile, *word sketches*[10], a feature available in the Sketch Engine, and are described using the CPA Editor (Baisa et al., 2015) and CPA's shallow ontology of semantic types[11]. Implicatures (pattern definitions) are written; register, domain, and idiom/phrasal verb labels are added, and links to FrameNet (Ruppenhofer et al., 2006) are created, linking the two complementary lexical resources. Dictionary entries also include quantitative information: for each separate pattern, a percentage is calculated based on the pattern's frequency in the annotated data (Figure 1 shows PDEV entry for *harvest*). PDEV-lemon, a linked data implementation of PDEV is available (Maarouf et al., 2014).

The **Pattern Dictionary of Spanish Verbs** (PDSV) is currently being developed within the *Verbario* project at the Pontifical Catholic University of Valparaíso, Chile. The goal of the project is two-fold: 1) to perform manual analysis of the most frequent Spanish verbs using CPA as a working methodology, and 2) to develop and implement procedures aimed at automatizing the creation of new patterns

[3] For the full list of patterns, see: `http://pdev.org.uk/#browse?q=avoid;f=A;v=avoid`

[4] `globalwordnet.org`

[5] `framenet.icsi.berkeley.edu`

[6] `babelnet.org`

[7] `omegawiki.org`

[8] `wiktionary.org`

[9] `linguee.com`

[10] `en.wikipedia.org/wiki/Word_sketch`

[11] `pdev.org.uk/#onto`

(Nazar and Renau, in press). The project uses the same version of the CPA shallow ontology as the DVC project, as it proved to be equally valid for Spanish. In addition, the CPA ontology also serves as a top ontology in the creation of a new automatic taxonomy of Spanish nouns, which is being applied to the task of labelling verb arguments with semantic types (Nazar and Renau, 2016). PDSV is being built following the same guidelines as PDEV and with the continued support from the English team, which ensures compatibility between the two projects and ensuing lexical resources. The Spanish team uses the same database structure and corpus interface as the English team (i.e. the Sketch Engine), but they focus on high-frequency verbs (as opposed to the predominantly medium-frequency verbs currently contained in PDEV), and typically annotate slightly larger corpus samples (i.e. between 250 and 1,500, depending on the verb).
Finally, a considerable amount of work has been conducted in the application of CPA to Italian (Ježek et al., 2014), which resulted in the creation of a parallel Pattern Dictionary of Italian Verbs (PDIV).

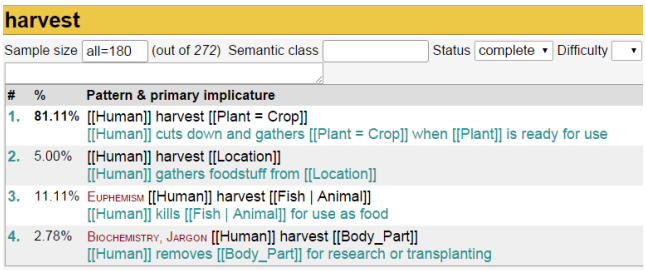


Figure 1: The dictionary entry for *harvest* in PDEV, as shown in the CPA Editor.

PDEV and PDSV are highly compatible in that they are being compiled using the same tools and methodology, making them perfect candidates for cross-linguistic pattern linking. In addition, CPA-based monolingual pattern dictionaries are developed independently of each other by different teams of lexicographers, which prevents dictionary data from being skewed due to possible interferences between languages. Corresponding pattern pairs in two or more languages can simply be linked to create a multilingual lexical resource based on their shared syntactic and semantic features. If successful, the proposed linking technique could make a significant contribution to the development of a new generation of multilingual lexical resources that focus explain meaning through patterns of real language use rather than abstract lists of word senses.

3. Manual Pattern Linking

In an effort to identify potential issues in the future, we decided to link patterns of a small subset of English and Spanish verbs. We selected 87 Spanish verbs with one or more English equivalents (126 in total), focusing on verb pairs such as *acusar* (accuse) and semantically equivalent groups of near synonyms such as *enfadar* (annoy/anger/infuriate/enrage). Pattern pairs identified through the manual linking procedure were later used as a gold standard in evaluation of the automatic linking task (Section 4.). Only verbs exhibiting up to 15 patterns were included in this pilot study, because highly polysemous verbs require specific strategies due to their grammatical complexity.
The study allowed us to identify the following methodological and practical issues that prevented us from finding full matches for all the patterns studied:

1. Both dictionaries differ significantly in terms of coverage: PDEV covers mainly low-to-middle frequent verbs, whereas PDSV contains middle-to-high frequent verbs. This reduces the number of potential matches; for instance, *golpear* (to hit) and to stab are often listed as translation equivalents in bilingual dictionaries despite the fact that their semantic overlap is very low.

2. The lack of full equivalence between languages, also known as anisomorphism (Yong and Peng, 2007). The following types of semantic anisomorphism were identified:

   (a) Lack of 1:1 correspondence: highly polysemous verbs typically exhibit a range of meanings and syntactic structures that differ significantly from their closest translation equivalents; in some cases, a pattern in a language might correspond to multiple patterns in the other language; e.g., for the previous example of *golpear*, a pattern such as '[[Human]] stabs (Physical Object 1) (at Physical Object 2)' could be considered equivalent to '[[Human]] golpear [[Physical Object]]', but the last one is too general to be matched to the English pattern.

   (b) Zero equivalence: some patterns simply do not have a corresponding pattern in the target language due to cultural, social, cognitive or pragmatic reasons. Idioms and other phraseological units are particularly problematic in that respect; e.g. the Spanish expression *sin comerlo ni beberlo* ('without being responsible for the damage caused to somebody'), which is listed as a pattern under the entry for *beber* (to drink), cannot be linked to any pattern for the verb *to drink*.

   (c) Syntactic differences: semantically equivalent pattern pairs often differ significantly in terms of their syntactic structure. A good example is the causative-inchoative alternation—a considerable portion of the verb pairs we studied showed that corresponding verbs often differ in the syntactic alternations they exhibit. For instance, the Spanish verb *agravar* exhibits both alternations, whereas its closest equivalent in English, *to aggravate*, can only be used in a causative construction.

4. Automatic Pattern Linking

To speed-up the labour-intensive procedure of manual linking, we decided to implement a heuristic-based algorithm

for automatic linking of pair candidates. Since the number of manually linked pattern pairs was very limited, it was not possible to train a machine learning system for the task. The small set of annotated manually pairs was used as a gold standard for evaluation of the method. Manual links are considered to be correct and the output of the automatic method will have to be constantly revised by lexicographers.

4.1. Algorithm

For each of the 490 Spanish patterns, we computed a similarity score for all its possible translations into English (i.e. verbs and their patterns, which resulted in a total of 5,067 Spanish-English pattern pairs). Candidate English patterns were then sorted by the score and the top pair was put forward as the best candidate for pattern linking.

The **similarity score** was computed by comparing pattern structures. Since this is a preliminary work, our analysis focused only on the three main syntactic arguments: subject, direct object and indirect object. An argument can have more than one semantic type associated with it, e.g. [[Human]] and [[Institution]] often occur together, as shown in Example 1. Whenever there was a non-empty intersection of semantic types in a given argument, each matched semantic type received one score point (only [[Human]], the most frequent semantic type, was assigned 0.5). If both given arguments were empty (also a match, mainly in the case of intransitive verbs), 0.5 score points were assigned. When the arguments contained different semantic types, the algorithm used the CPA ontology to check if the two types are in a hypernym relation (e.g. [[Event]] is the hyponym of [[Eventuality]]). If, for instance, [[Event]] appears as the direct object in the Spanish pattern and [[Eventuality]] in its English counterpart, we can use the CPA ontology to get a partial match). Each hyponym or hypernym got score points based on the distance in the CPA ontology tree (the further apart they are located, the fewer score points they gain, measured in powers of 0.5). Scores for the three slots (subject, direct indirect object) were summed and the final score was assigned to the given pattern pair (cf. Table 1). All candidate pairs were sorted by the score and the top ranking pattern was returned.

Spanish	English	Scr	Comment
Entity|Eventuality	Human	1/8	Human < Animate < Physical_Object < Entity, distance = 3
Human	Human	1/2	Human is almost in all patterns so the score was only 0.5
Artifact	Eventuality	0	No relation in ontology

Table 1: Examples of ontology matches and the resulting scores (Scr) for pattern arguments.

The first column contains Spanish semantic types in a pattern argument. Since both PDEV and PDSV contain verbs with patterns containing semantic type Human in subject argument, the algorithm considers it as a weaker sign of equivalence. When two different semantic types S and E are in the same argument in a Spanish and an English pattern, CPA ontology (which is shared between PDEV and PDSV) is queried. If S is hypernym/hyponym of E or vice versa, the score is computed as 0.5^N where N is the distance in the ontology hierarchy (a tree in the case of CPA ontology).

Not all possible pattern pairs were considered, only patterns of equivalent English and Spanish verbs were taken into account. We have used a statistical English-Spanish dictionary derived from a parallel English-Spanish corpus. It is important to note that even if a verb in one language has more than one translation equivalent in the other language, the comparison of pattern structures should narrow the number of all possible pattern pairs—a pattern express one of possible meanings of a verb and it is reasonable to expect equivalent patterns to have the same or similar structure.

To evaluate the method, we created a random sample of 50 Spanish-English verb pairs. We excluded all cases in which a Spanish pattern cannot be matched against an English pattern in the sample, although we are fully aware of the fact that a matching English pattern could potentially be found outside the sample (we calculated that this happens in around 40% of the cases in our sample). Despite our work being at an early preliminary stage, the proposed method shows promising results, achieving 80% precision: 40 of the 50 pairs were correctly suggested as candidates and the rest was incorrect.

5. Conclusion

The paper presented the results of a pilot study on linking verb patterns across languages. Despite the fact that our work is currently at an early preliminary stage, the study clearly demonstrated the advantages of linking methodologically compatible, monolingual pattern dictionaries through a combination of both manual and automatic procedures. The algorithm developed for the task performed remarkably well considering the size of our gold standard dataset. There is plenty room for improvement—the manual task will have to be further refined, and the algorithm's performance improved by augmenting the size of the training data. Nonetheless, the work presented here will serve as a solid basis for the future development of the proposed methodology and ensuing lexical resource. Our immediate plans for the future include the creation of larger gold standard datasets of manually linked pattern pairs, as well as the adaptation of the software in a way that will allow lexicographers from different teams to manually specify links between two or more patterns contained in CPA-based pattern dictionaries. Our ultimate goal is to create a valuable, multilingual, corpus-driven lexical resource for verbs that reflects real language use and can therefore be used by language learners, language professionals (e.g. translators, editors) and the research community alike.

6. Acknowledgements

This work has been partly supported by the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic, AHRC grant [DVC, 47. AH/J005940/1, 2012-2015] and by the Conicyt-Fondecyt project "Detección automática del significado de los verbos

del castellano por medio de patrones sintáctico-semánticos extraídos con estadística de corpus" (nr. 11140704, lead researcher: Irene Renau), which is funded by the Chilean Government.

7. References

Baisa, V., El Maarouf, I., Rychlý, P., and Rambousek, A. (2015). Software and data for Corpus Pattern Analysis. In Horák, A., Rychlý, P., and Rambousek, A., editors, *Ninth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 75–86, Brno. Tribun EU.

Buyse, K. and Verlinde, S. (2013). Possible effects of free on line data driven lexicographic instruments on foreign language learning: The case of Linguee and the interactive language toolbox. In *Procedia: Social and Behavioral Sciences*, volume 95, pages 507–512. Elsevier BV.

Fillmore, C. J. and Baker, C. (2010). A frames approach to semantic analysis. *The Oxford Handbook of Linguistic Analysis*, pages 313–339.

Hanks, P. (2004a). Corpus Pattern Analysis. In *Euralex Proceedings*, volume 1, pages 87–98.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274.

Hanks, P. (2013). *Lexical Analysis: Norms and exploitations*. Mit Press.

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., and Popescu, O. (2014). T-pas: A resource of corpus-derived types predicateargument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 26–31.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Leech, G. (1992). 100 million words of English: the British National Corpus (BNC). *Language Research*, 28(1):1–13.

Maarouf, I. E., Bradbury, J., and Hanks, P. (2014). PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL): Multilingual Knowledge Resources and Natural Language Processing at the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Nazar, R. and Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia.

Nazar, R. and Renau, I. (in press). A quantitative analysis of the semantics of verb-argument structures. In *Collocations and Other Lexical Combinations in Spanish. Theoretical, Lexicographical and Applied Perspectives.*, pages 92–108. Ohio University Press.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice.

Vossen, P. (2002). WordNet, EuroWordNet and Global WordNet. *Revue Française de Linguistique Appliquée*, 7(1):27–38.

Yong, H. and Peng, J. (2007). *Bilingual Lexicography from a Communicative perspective*, volume 9. John Benjamins Publishing.

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.*

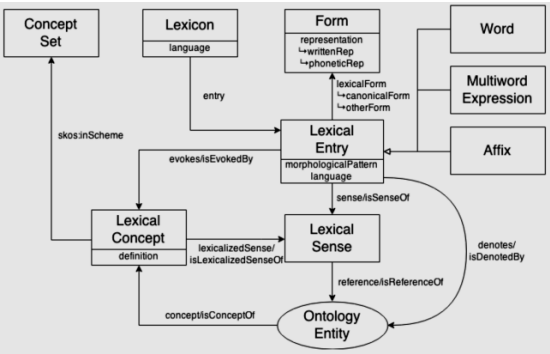


Figure 1: The core model (*ontolex*)
Figure created by John P. McCrae for the W3C Ontology-Lexica Community Group.

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.

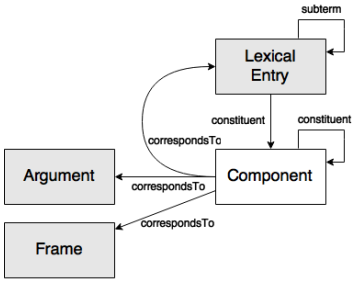


Figure 2: The relation between the decomposition module and the LexicalEntry class of *ontolex*.
Figure created by John P. McCrae for the W3C Ontology-Lexica Community Group

.

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

Building an Ontological Model of the BLL Thesaurus First Steps Towards an Interface with the LLOD Cloud

Vanya Dimitrova, Christian Fäth, Christian Chiarcos, Heike Renner-Westermann, Frank Abromeit
Goethe Universität Frankfurt am Main, Germany
{v.dimitrova|h.renner-westermann}@ub.uni-frankfurt.de,
{faeth|chiarcos|abromeit}@informatik.uni-frankfurt.de

Abstract

This paper describes ongoing efforts to position the Bibliography of Linguistic Literature (BLL) within the wider context of Linguistic Linked Open Data (LLOD), and to enhance the functionality of the *Lin|gu|is|tik* portal, a virtual library for the field of linguistics, with an LOD interface on this basis. Being the connecting point between the portal and LLOD cloud resources, the BLL Thesaurus has to fulfill novel and specific formal and conceptual requirements.

Remodelling the BLL Thesaurus as an OWL ontology is the main subject of the paper. We sketch specifics of the thesaurus, its scope and nature, and describe our general methodological approach, design decisions and a usage scenario. We present the basic ontological framework and discuss concrete challenges encountered in the conversion process. Concrete examples from the domains of morphology and syntax demonstrate the complexity of the task. Additionally, we elaborate on the next steps towards an LOD interface and long-term perspectives.

Keywords: Bibliography of Linguistic Literature (BLL), thesaurus of linguistic terminology, ontological modelling, Linguistic Linked Open Data (LLOD), Ontologies of Linguistic Annotation (OLiA)

1. Introduction

The Bibliography of Linguistic Literature (BLL) is one of the most comprehensive linguistic bibliographies worldwide. It covers general linguistics with all its neighboring disciplines and sub-domains as well as English, German and Romance linguistics. Dating back as far as 1971, the BLL lists over 452,000 references and has an annual growth of about 10,000 entries. The BLL does not only represent a significant source of bibliographic data, but it also provides a hierarchically categorised bilingual thesaurus of domain-specific index terms.

The historical development of the BLL Thesaurus represents a prototypical example for the way a resource of this kind evolves over time, how it can be expanded to cover novel application scenarios and what challenges to expect in the process. Its primary use case is the indexing of bibliographic records, but recently, it is also applied for indexing online resources in the context of the virtual library under `http://www.linguistik.de`, henceforth *Lin|gu|is|tik* portal. At the moment, the BLL Thesaurus is being extended with a linking to terminological repositories in the Linguistic Linked Open Data (LLOD) cloud and will be employed to develop an (L)LOD-based search facility within the *Lin|gu|is|tik* portal.

These novel use cases, introduced in Sect. 2, pose new conceptual and technical requirements. Section 3 describes the ongoing process of remodeling the BLL Thesaurus as an ontology; we discuss peculiarities of the thesaurus, our general approach, methodological and practical challenges as well as its potential for developing an ontology-based search functionality. Section 4 uses concrete examples to demonstrate design decisions, challenges and strategies in building the BLL ontology. Finally, Sections 5 and 6 present the prospective linking with LLOD resources such as the Ontologies of Linguistic Annotation and the development of an (L)LOD-based search functionality on this basis.

2. The *Lin|gu|is|tik* portal and LLOD

The *Lin|gu|is|tik* portal represents a virtual library that provides an integrated access to scientific information on every subject of linguistics, ranging from general and comparative linguistics over the documentation and study of minor, threatened or ancient languages to the investigation of larger European languages (Renner-Westermann, 2013). Funded by the German Research Foundation (DFG), the *Lin|gu|is|tik* portal is an ongoing cooperation between Goethe University Frankfurt (represented by the University Library and the Applied Computational Linguistics lab), the Institute of German Language (IDS Mannheim),[1] and the LINSE Linguistik-Server of the University Duisburg-Essen with its link database LinseLinks.[2]

The virtual library comprises six main modules: Five directories of online resources (links, journals, databases, dictionaries, and research projects) and a Catalogues module with an integrated search function for numerous sources. This includes the catalogues of the University Library and the IDS Mannheim, the Online Contents Linguistik, all directories of online resources, diverse open access documents as well as the Bibliography of Linguistic Literature (BLL).[3] To a large extent, this data is indexed against subject terms that are organized in the BLL Thesaurus.

The *Lin|gu|is|tik* portal is designed as a hub for linguistically relevant scientific information. Implementing an LLOD interface provides a natural way to increase its scope and capabilities: The portal will be extended with an LOD-

[1] `http://www.ids-mannheim.de/org`
[2] `http://www.links.linse.uni-due.de`
[3] `http://www.blldb-online.de`

based search facility to immediately retrieve LLOD resources. The BLL Thesaurus will serve as a pivot connecting the *Lin|gu|is|tik* portal and the LLOD cloud.
Linguistic Linked Open Data[4] is a movement about publishing open language resources for different use cases in academic research, applied linguistics or natural language processing. Currently, it comprises 126 resources, including lexical-conceptual resources (dictionaries, knowledge bases), corpora, terminology repositories (thesauri, ontologies and registries for linguistic concepts, features, and terms), and metadata collections (language resource metadata, bibliographies). Since its first instantiation in September 2012, the LLOD cloud continues its rapid growth because it provides important benefits as compared to legacy formalisms: flexible representation, structural interoperability, explicit semantics, conceptual interoperability, federation, dynamicity, and ecosystem (Chiarcos et al., 2013). Conceptual interoperability is particularly fruitful in the context of a virtual library: By linking the BLL Thesaurus to LLOD terminologies, BLL records immediately become interoperable with other LLOD resources such as the World Atlas of Language Structures,[5] the Phonetics Information Base and Lexicon,[6] or the Glottolog/LangDoc bibliography.[7] These links and the use of shared vocabularies allow us to automatically access and index LLOD language resources and thereby to develop a (linked) language resource search as part of the *Lin|gu|is|tik* portal.

The implementation of an LLOD interface includes the following steps:

1. Converting the BLL Thesaurus to SKOS/RDF
2. Remodelling the SKOS edition of the BLL Thesaurus as an ontology
3. Converting bibliographic records and their indexation (BLL Thesaurus subject terms) to RDF
4. Linking the BLL Ontology with LLOD terminology repositories
5. Developing a search algorithm, data storage solutions and a query interface

In the current phase, we focus on remodelling the BLL Thesaurus as an OWL2/DL ontology. Even though a naïve SKOS representation may already seem sufficient to establish a linking with selected LLOD resources, we aim to provide a formally consistent and re-usable resource with rigidly defined data categories for our domain, linguistic thesauri. As such, OWL provides description logical operators (conjunction/intersection $\sqcap$, disjunction/join $\sqcup$, negation/complement $\neg$) to represent and to (partially) resolve conceptual overlap and ambiguity as observed in the BLL Thesaurus (Sect. 4.). Furthermore, rigidly applying OWL constraints to re-modelling an existing terminology resource often helps to uncover problematic modelling decisions and thereby facilitates developing a more consistent representation of domain terminology.
Moreover, OWL2/DL has the potential to develop a search functionality that makes use of automated reasoning technologies as described in Sect. 3.3. . While this is beyond the scope of our current project, it is a concrete possibility in combination with other OWL2/DL-based resources such as the Ontologies of Linguistic Annotation (Chiarcos and Sukhareva, 2015, OLiA).[8] Indeed, our efforts to create an BLL Ontology take an initial focus on the domains of syntax and morphology, and for these, OLiA represents a central terminological hub within the LLOD cloud. OLiA adopts OWL2/DL as its primary modelling framework, so that integrating the BLL Thesaurus in the modular OLiA architecture, requires an OWL modelling to establish valid links with the OLiA Reference Model (see Sect. 5.). OLiA provides this 'Reference Model' as an intermediate representation between domain-specific vocabularies such as the BLL Thesaurus and several LLOD terminology repositories in the LLOD cloud, thereby facilitating the interoperability of linked resources with not only one LLOD terminology repository, but with multiple repositories developed and maintained by different communities and for different purposes (Chiarcos and Sukhareva, 2015).

3. The BLL Thesaurus as an ontology

This section describes specific peculiarities of the BLL Thesaurus as determined by its original use case, our general approach towards the construction of an ontological model and its prospective potential to provide a novel, ontology-based search functionality.

3.1. BLL Thesaurus: Facts and features

The BLL is based on the acquisitions of the Special Subject Collection area *General Linguistics* of the DFG.[9] Since its first publication, the classification and the subject terms used for indexing in the bibliography have been continuously enhanced. By the time of writing, they comprise a thesaurus of 7,481 hierarchically organised index terms.
At the top level, the BLL Thesaurus is organized into several 'branches'. The main branch *Levels*[10] includes the levels of language description (e.g., *Syntax*, *Phonology*) and consists of 1,983 subject terms. With respect to *Languages*, 2,141 index terms are defined for the indexing of language varieties, including dialects, reconstructed and artificial languages. The branch *Domains* covers the subdisciplines of linguistics (e.g., *Psycholinguistics*, *Sociolinguistics*) and lists 3,050 subject terms. In addition to independent terms, the BLL Thesaurus provides 7,700 *see references*, i.e., cross-references that indicate synonymous expressions or aliases such as '*Multitasking* see *Cognitive complexity*', '*Grounding (pragm.)* see *Mutual knowledge*').

[4] `http://linguistic-lod.org`
[5] `http://wals.info`
[6] `http://phoible.org`
[7] `http://glottolog.org/langdog`
[8] `http://purl.org/olia`
[9] The Special Subject Collections were parts of a system established at different university libraries in Germany after WWII, and designed to support the acquisition of the international literature in every specific field of research.
[10] Thesaurus subject terms are represented in *italics*, and ontological classes in `typewriter font`.

The primary use case of the BLL Thesaurus is indexing linguistic literature, and this function determines its current scope and structure. By continuous accommodation to the ongoing development in the field of linguistics, the BLL Thesaurus evolved over time, and its future growth and adaptions will be determined by recent publications in the field. It will thus reflect the terminological progress and thematic specializations of the field over time. This is an important difference in comparison to other terminological repositories and ontologies of linguistic terminology. Beyond giving an account of the basic categories, the BLL Thesaurus also captures terms from areas of intensified research which are not widely accepted or may even be a topic of controversial debate, and thus cannot be considered "standard knowledge" of the field. Such "peripheral phenomena" include, for example, *Split topicalization* and *Inflected infinitive*, and neither of these can currently be found in any LLOD term base.[11]

In order to properly capture the focus of the indexed literature and cover novel subjects and topics, numerous new subject terms and *see references* are included year after year. In 2014, 235 subject terms were added to the thesaurus, for example *Argument sharing*, *Parasitic participle*, and *Whispered interpreting*.

While the BLL Thesaurus will continue to grow, deletion of subject terms happens extremely seldom and only if related subject terms are merged into a new category. In 2014, for example, the subject terms *Geography (technical language)* and *Geodesy (technical language)* were combined to form the new subject term *Earth sciences (technical language)*.

Yet, the continuous adjustment to the development in the field should not unnecessarily complicate its application or affect the integrity of the bibliographic database. It is particularly important to note that such deletions do not lead to inconsistencies. Internally, BLL subject terms are identified by unique and stable IDs. In case of subject term deletion, the respective IDs are blocked and cannot be reused. Accordingly, these IDs can be used a basis to generate persistent URIs for Linked Data editions of the BLL Thesaurus: Any statement involving a deleted subject term (resp., its URI) will remain valid, only that it will be marked as being deprecated if a deletion occurred.

For developing an LLOD edition of the BLL Thesaurus, we decided to model it as a resource in its own right which can subsequently be linked with existing terminology repositories. Also in the LLOD community, different theoretical approaches lead to deviating views, dissimilar conceptualisations and incompatible models of the linguistics domain. It is thus not possible to directly design the LLOD edition of the BLL Thesaurus as a sub-module of an existing ontology, e.g., as a Community of Practice Extension (Farrar and Lewis, 2007, COPE) of the General Ontology of Linguistic Description (Farrar and Langendoen, 2003, GOLD).

Indeed, as Farrar and Langendoen (2010, p.47) point out, a single and uniform "ontology for all of linguistics is, at this point, unachievable and would require deep consensus as to how language is conceptualized". As an example, the General Ontology for Linguistic Description (GOLD) was designed as an ontology for descriptive linguistics for the needs of the language documentation community. It gives a formal and partially axiomatised account of the most basic categories and relations used in the scientific description of human language (Farrar and Langendoen, 2003), and its main objective is to secure interoperability of language data.

The BLL Thesaurus, on the other hand, is used for the indexing and documentation of linguistically relevant publications, and it aims at covering the field *in its entirety*: languages, domains, subdomains, subdisciplines as well as theoretical frameworks. The scope of the BLL Thesaurus and the quantity of subject terms are incommensurable with those of general terminological repositories or ontologies that tend to focus on a few well-documented domains (e.g., morphosyntax). However, the way linguistic concepts are represented in the BLL Thesaurus is less axiomatic or definitional. The hierarchical organisation of the linguistic concepts is based mainly on thematic and lexical associations, not necessarily on proper conceptial subsumption.

3.2. Remodelling the thesaurus: General methodology

An introduction of the ontology building methodology can be found in Farrar (2007) and Farrar and Langendoen (2010). However, our purpose is not to provide a new general ontology for the domain of linguistics, but to reorganise and remodel an existing classification into an ontology in order to use it as a part of a data integration process in the wider context of LLOD. Thus, in our specific case, a different path must be followed.

The BLL Thesaurus not only provides a list of domain-specific subject terms, but also represents these terms in a hierarchical tree structure. So, apparently, the first steps in the process, namely enumerating the basic categories found in the given domain and developing class taxonomies (Farrar, 2007), had already been done. Still, the hierarchical relations that exist in the thesaurus only partially fulfill the criteria of a formal ontology.

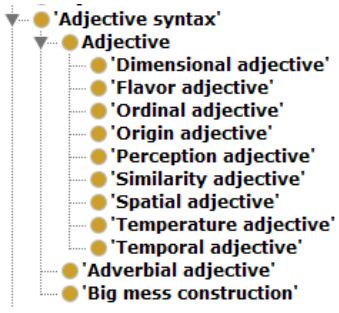


Figure 1: Hierarchical structure of the BLL Thesaurus

[11] In addition to OLiA, we checked TDS (`http://languagelink.let.uu.nl/tds/ontology/LinguisticOntology.owl`), GOLD (`http://linguistics-ontology.org/gold`), ISOcat (`http://isocat.org`) and DBpedia (`http://dbpedia.org`).

The BLL Thesaurus is internally represented in OCLC PICA,[12] and its semantics are based on lexical associations rather than the object-oriented view underlying OWL and the `rdfs:subClassOf` property. Figure 1 shows the subject term *Adjective* with its BLL parent, siblings and subcategories: While the subcategories of *Adjective* can indeed be regarded as subclasses in an ontological sense, it is hard to say that "Every adjective **is an** adjective syntax", and so, the exact interpretation of *Adjective syntax* and the relation of *Adjective* and its sibling concepts is problematic. Because of the nature of hierarchical relations in the BLL Thesaurus, a "naïve" automated conversion of its hierarchical structure from OCLC PICA to `rdfs:subClassOf` properties and a full-fledged OWL model will not produce a valid ontological representation. Therefore, we employ a *two-layered* approach: We first convert OCLC PICA to RDF/SKOS in fully automated fashion, and then, this RDF/SKOS representation is reassessed and thoroughly revised in order to establish a valid ontology.

As a result of the first step, the BLL hierarchy is expressed by `skos:broader` relations, which are less rigidly defined than `rdfs:subClassOf` and recommended for modelling thesauri (Pastor et al., 2009). The automatically created SKOS file is then imported into an OWL editor[13] and all BLL concepts are manually classified and organised to build the actual BLL Ontology.

Since the URIs are generated from stable internal IDs, any future SKOS export of the BLL Thesaurus will produce identical concept URIs, so that concepts from earlier SKOS editions of the thesaurus that were previously defined as part of the BLL Ontology will maintain their ontological rendering. Newly added subject terms will need an ontological rendering, i.e., manual classification. In case of subject term deletion, the affected ontological class will be automatically recognized as deprecated and marked as such.

Even though a list of domain-specific subject terms is already available from the SKOS export, we still have to pose the question which notions depicted by these terms are indeed categorical. Since different theoretical frameworks are also represented as a part of the Thesaurus, one encounters subject terms such as *Minimalist morphology* and *Distributed morphology* (subcategories of *Theory (morph.)*). Are these categorical by nature, or instances of the concept 'Morphological theory'?

In order to decide whether a particular subject term is an ontological class or an individual instance, the potential applications of the ontology must be taken into consideration. In general, this decision is determined by meta-ontological criteria and "to some extent, this is an arbitrary modelling choice" (Farrar, 2007, p.177). Our approach regarding the decision where classes end and instances start is grounded in the original use case of the BLL Thesaurus and the functionality we aim for.

Glottolog, for example, employes a set-theoretic approach and models languages as concepts: A language is defined by the set of documents which describe it (Nordhoff and Hammarström, 2011). Every languoid (a cover term for dialect, language, and language family) is seen thus as a set and modelled as a SKOS concept. As a consequence, genealogical relations can be modelled in an elegant fashion by means of $\sqsubseteq$ (`skos:broader`) relations. An account of the further advantages of this method can be found in Nordhoff (2012).

In a similar vein, we see BLL subject terms as classes with bibliographic entries from the BLL as their instances in the BLL Ontology. The subject terms, empirically grounded in the BLL Bibliography, represent ontological classes which are defined as collections of references to (bibliographical or related) resources, their hierarchical structure can thus be used for subsumption inferences. For example, the previously mentioned theoretical frameworks are represented in the BLL Ontology as subclasses of `TheoreticalFramework` under the top-level concept `OtherLinguisticTerm` (Figure 2). If a given publication is indexed for being concerned with *Categorial syntax*, it can be retrieved by queries for either `SyntacticTheory`, or `TheoreticalFramework`, but also `OtherLinguisticTerm`.

With an OWL2/DL model, even more complicated querying operations are possible.

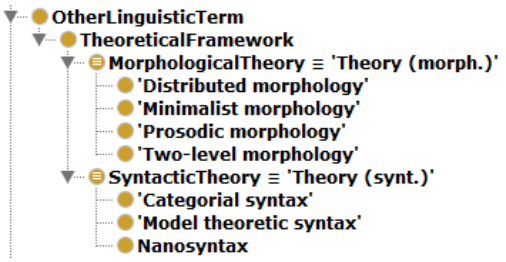


Figure 2: Hierarchical organisation of *Theoretical frameworks* in the BLL Ontology

3.3. Use case: Ontology-based querying

Before presenting ongoing developments regarding the development of an interface between the BLL Thesaurus and the LLOD cloud, we would like to illustrate possible applications of the BLL Thesaurus as an ontology.

One practical motivation to provide an OWL2/DL model as an additional layer over the original SKOS representation of the BLL Thesaurus is to validate and revise the resource from a top-down perspective. More importantly, however, is that it opens the possibility to formulate *ontology-based queries* over the bibliographical resources it is associated with. With an OWL2/DL ontology, a reasoner can be applied and thus, description logical queries can be formulated to retrieve results with a much more expressive formalism than currently possible within the *Lin|gu|is|tik* portal.

At the moment, the extended search functionality of the *Lin|gu|is|tik* portal (Fig. 3) allows a user to provide and to combine up to five criteria to filter matches to a query. In addition to keywords (i.e., BLL concepts), these include free text search, title matches, creator/publisher and year. These filters can be combined with the operators `AND`,

[12] `http://www.oclc.org`

[13] Protégé 5.0, `http://protege.stanford.edu`

OR and AND NOT. The operators are combined in a left-associative fashion.

Imagine now a linguist is interested in cross-lingual manifestations and determinants of grammatical voice. Grammatical voice refers to alternations in the syntactic realisation of semantic arguments, e.g., whether the PATIENT of a given predicate is realized as direct object (in an active clause) or as subject (in a passive clause). Obviously, this involves a semantic component, but as a trigger of either active or passive constructions, contextual factors such as the surrounding discourse play a role. The nature and the effect of these discourse factors have been a matter of intense research (Givón, 1994) and continue to be an active research area in comparative linguistics.

Unfortunately, querying for the keyword 'voice' with the current search functionality conflates different keywords from phonetics and syntax, respectively, but a refined search for 'grammatical voice' produces only seven matches. Our researcher might thus want to broaden the scope of this query by circumscribing the phenomena she is interested in. This involves two aspects, the morphosyntactic realisation of arguments (by grammatical case) and verbs (by voice marking), and their underlying determinants (i.e., semantic roles or discourse factors).

This can be immediately expressed in the following OWL2/DL class description.

$$(\texttt{CaseFeature} \sqcup \texttt{VoiceFeature}) \sqcap (\texttt{DiscourseFeature} \sqcup \texttt{SemanticRole})$$

Such class descriptions can be directly used for querying bibliographical references with the BLL Ontology. Note that such a complex query could not be expressed with the current search functionality as this is left-associative. It can, however, be directly fed into a reasoner and tested already with off-the-shelf tools like Protégé. One possible practical application of the OWL version of the BLL Thesaurus can thus be seen in an additional search functionality that may be added to the current keyword search.

Unlike the current keyword search, it allows to combine an arbitrary number of criteria, it is neither limited to five keywords, nor to left-associative operator grouping. With other parts of the BLL Thesaurus (e.g., languages) formalized and linked, more complex queries become possible and may even extend beyond search in bibliographical data. With the syntactic and morphosyntactic components of the BLL Thesaurus linked with LLOD vocabularies such as OLiA (Sect. 5.), it is possible to run the same query not only over BLL-linked publications, but also over (metadata for) linguistic corpora, and corpora with annotations for case (or voice) and semantic roles (or discourse phenomena) may be retrieved in this way, e.g., the German SALSA corpus (Burchardt et al., 2006). Using such a corpus, it would be possible to quantitatively study correlations and deviations between grammatical case (or morphological voice) and semantic roles, and their contextual determinants.

Such an extended search functionality is a prospective goal of our work, but a mid-term goal only. At the moment, we focus on more elementary questions regarding the modelling of the BLL Ontology and its linking with terminological resources from the LLOD cloud.

4. Challenges of ontology modelling

In order to demonstrate the complexity of the task to create a formal ontology from a SKOS thesaurus in our domain, we provide concrete examples with classification proposals and formal solutions. The ontological remodelling of the BLL Thesaurus initially focuses on the levels of linguistic description, more precisely on the branches *Syntax* and *Morphology*, consisting of 289 and 191 subject terms, respectively.

For the ontology development, different strategies can be employed. The top-down approach starts with redefining the most general concepts before descending into their subconcepts, whereas the bottom-up approach would begin with the most granular index terms combining them into higher level concepts later on. Both methods have certain disadvantages when remodelling existing hierarchical order: they are "blind" to the levels above or below a given concept.

Therefore, for the remodelling of the BLL Thesaurus and the reorganisation of the taxonomies, we use a combination of both strategies, we start with the most salient concepts. Since subject terms are already hierarchically structured, we apply initially a top-down approach. When encountering ambiguous concepts or problematic hierarchical relations that do not fulfill the ontological criteria, we examinate the concepts again in a bottom-up manner. This is a labourous task that may even require to inspect the referenced bibliographic entries.

4.1. Defining the top-level structure

When developing the ontology, we follow the general principle "define first the salient concepts and then generalize and specify them appropriately". So, *Verb* and *Adverb* are defined as morphosyntactic categories, *Case* and *Tense* as morphosyntactic features and *Word formation* and *Inflection* as examples for a morphological process.

We establish a basic class structure and top-level concepts by grouping BLL subject terms around the elementary notions linguistic **category**, linguistic **feature**, linguistic **process**, and linguistic **relation**. Currently, the primary top-level concepts and their immediate subclasses include:

```
MorphologicalTerm
   MorphologicalCategory
   MorphologicalProcess
MorphosyntacticTerm
   MorphosyntacticCategory
   MorphosyntacticFeature
SyntacticTerm
   SyntacticCategory
   SyntacticConstruction
   SyntacticFeature
   SyntacticProcess
   SyntacticRelation
```

Entities that clearly fall in one of these groups are categorical by nature and constitute an ontological class. Yet, a number of concepts are harder to classify, and these are currently assembled under OtherLinguisticTerm, a sibling concept to MorphologicalTerm, etc.

Figure 3: Extended search in the *Lin|gu|is|tik* portal

The BLL Ontology specifies classes for linguistic categories (e.g., `Adjective`, `Determiner`, subsumed under `MorphosyntacticCategory`), as well as for grammatical features (e.g., `CaseFeature` and `AspectFeature` as subclasses of `MorphosyntacticFeature`). Note that the BLL Ontology differs from GOLD or OLiA in that no object properties are defined that would assign features to (instances of) categories, etc. This is because such relations are not necessary in our domain, the structuring of bibliographical records.

Considering the nature of the BLL Thesaurus and its usage, even the division into XY`Feature` and XY`Category` is partially an artificial one. This is not strictly required for terms used to describe linguistically relevant publications, but follows other LLOD vocabularies and primarily serves as a mechanism to assist the ontological restructuring here.

4.2. Reorganising subject terms

Reorganising thesaurus concepts into this top-level structure requires verification of the existing definitions for a given linguistic concept. As these definitions are often implicit in the original BLL Thesaurus data, a disambiguation of subject terms is only possible through scrutinising the indexed bibliographic entries. Different scenarios emerge according to the nature of the BLL concept.

In principle, the requirements for a consistent ontological structure can often be fulfilled by clarifying and adjusting name and labels of a concept. *Syntax* as a cover term for concepts like *Noun phrase* is obviously a misnomer for an ontological model, but could be easily replaced by a designation that clearly states that it covers to *elements of* syntactic analysis. A natural approach to implement such renaming proposals would be to assign a different `skos:prefLabel`. Unfortunately, this is not possible in our setting, as this property is reserved for a different purpose in the SKOS edition of the thesaurus.[14]

Instead, we create a new ontological class that follows conventional naming/labeling strategies, `SyntacticTerm` in the example, we define it as being equivalent (i.e., co-extensional, $\equiv$) with the problematic BLL concept *Syntax* as automatically exported from the BLL Thesaurus and mark the latter as being deprecated. Similarly, the subject term *Case (morph.)* is identified with a novel concept `CaseFeature`, etc. The reorganisation of the taxonomic structure is thus facilitated by creating ontological classes without a corresponding BLL subject term, e.g., `SyntacticConstruction`, `MorphosyntacticFeature`.

In many cases, the existing hierarchical relations can be preserved, e.g., the subclasses of *Adjective* (Fig. 1) or most subcategories of *Word formation* (*Apheresis*, *Contamination*, *Derivation*, etc.) fulfill the requirements for an ontological subclass. Yet, occasional changes in the hierarchical organisation are required because of the nature of the phenomenon designated by a BLL subject term. For example, *Embedding*, a BLL subcategory of *Subordinate clause*, cannot be an ontological subclass of `Subordinate clause` in the ontology. *Embedding* refers to a process rather than a construction (as evident from its grammatical form as a gerund ending in *-ing*), in the ontological model, it is thus a subclass of `SyntacticProcess`. Similarly, `Denominal adjective` is a subclass of `Adjective`, but not of *Derivation*, the original mother-concept in the Thesaurus. *Derivation*, on the other hand, preserves its hierarchical position:

```
MorphologicalProcess
    Word formation
        Apheresis
        Backformation
        Composition
        Contamination
        Derivation
```

In the BLL Thesaurus, a single linguistic phenomenon is often described by multiple subject terms. This reflects

[14] Within the BLL Thesaurus, this is used for name extentions which specify the context of usage or the perspective of analysis and to distinguish homonyms (e.g., *Adjective (lex.)*).

the multi-stratal nature of language and the different perspectives of analysis (e.g., *Case (morph.)* and *Case (synt.)*; *Anaphora (synt.)*, *Anaphora (pragm.)*, *Anaphora (psycholing.)*). Here, the name extentions (in brackets) distinguish different points of view, e.g., an analysis from the perspective of morphology, syntax, pragmatics, or psycholinguistics. In the BLL Ontology, such phenomena are grouped together in a single ontological class, e.g.:

`CaseFeature` ≡ `Case(morph.)` ≡ `Case(synt.)`

4.3. Complex classes (⊓, ⊔)

Some BLL subject terms fall in the conceptual overlap of existing classes (e.g., *Adverbial adjective*, *Compound adjective*). In the BLL Ontology, these are modelled as subclasses of the intersection of the overlapping classes. *Verbal compound*, for example, is defined as a subclass of both `Verb` (subclass of `MorphosyntacticCategory`) and (⊓) `Compound` (subclass of `MorphologicalCategory`):

`Verbal compound` ⊑ `Verb` ⊓ `Compound`

This example illustrates that the BLL Ontology allows us to exceed beyond the rigid structural limits of a classical thesaurus: A single subject term can be assigned to multiple mother concepts. Indeed, the semantics of BLL classes as aggregates of bibliographical records means that all classes can share instances. Thus, we abstain from disjointness axioms in the BLL ontology.

Ambiguity represents another source of complex class definitions. In a few cases, one BLL concept covers two different linguistic phenomena, e.g., *Topicalisation* and *Inversion* both denote a linguistic process as well as the outcome of this process. So, a disambiguation is required. Some ambiguous subject terms can be disambiguated by renaming or by changing their hierarchical position, but others require different formal and conceptual solutions.

Compounding, for example, is applied for the morphological process of composition as well as for compounds as a morphological category. To resolve this, it is first defined as a subclass of `AmbiguouslyDefinedConcept`, a newly defined ontological top-level concept specifically designed to assemble concepts with such a problematic definition. Additionally, two new ontological classes are created: `Composition` (a subclass of `MorphologicalProcess`) and `Compound` (a subclass of `Morpheme`). The BLL subject term *Compounding* is subsequently equated (≡) with the disjunction of the newly introduced classes:

`Compounding` ≡ `Compound` ⊔ `Composition`

Several BLL subject terms denote an opposition that has to be resolved. The subject term *Mass noun/count noun* is modelled in a way similar to the ontological representation of *Compounding*: It is defined as a subclass of `AmbiguouslyDefinedConcept` and equated by an *EquivalentClasses* axiom to the disjunction of the newly introduced classes `MassNoun` and `CountNoun`:

`Mass noun/count noun` ≡ `MassNoun` ⊔ `CountNoun`

Alternatively, such oppositions can be resolved by referring to a generalisation that covers both aspects of the opposition. This was the modelling we chose for the BLL subject term *Inclusive/exclusive*. Here, a new class `ClusivityFeature` (a subclass of `MorphosyntacticFeature`) is defined, and the BLL term is equated to it by a simple equivalence:

`Inclusive/exlusive` ≡ `ClusivityFeature`

Handling ambiguous subject terms by using disjunction constructs brings certainly both advantages and disadvantages. Using such constructs allows to more precisely capture the intended meaning and application of an ambiguous concept within the Thesaurus: So, we stay close to the primary data (the bibliographic entries), and resolve cases of terminological overload. On the other hand, disjunction constructs somewhat complicate the structure of the graph by breaking the chain of `subClassOf` and `equivalentClass` relations thus making it less trivial to create SPARQL queries.

4.4. Unclassified subject terms

Because of the peculiarities of the BLL Thesaurus and its domain, it is inevitable to encounter subject terms that do not fit in the defined ontological scheme. For example, the BLL concept *Transitional probability (synt.)* from the *Syntax* branch of the thesaurus is still not classified. Similarly, *Pluralis majestatis* and *Analogy (morph.)* from the *Morphology* branch can not find adequate ontological rendering for the time being. These are tentatively presented as subclasses of the top-level concept `UnclassifiedConcept` introduced to enable a possible future rendering expected with the inclusion of additional branches.

To summarize this section, it is evident from these examples that the nature of the BLL Thesaurus prohibits generic solutions for challenging cases: Only by thorough case-by-case clarification we can stay close to the primary BLL meaning. The further refinement of the various ontological entities depends on the particular application of the ontology. "In general axioms should be limited to asserting what must be the case versus what can be case." (Farrar, 2007, p.178). The ongoing work on the remodelling of the BLL Thesaurus follows this principle. Our main objective is to create an ontological representation sufficiently detailed to facilitate the implementation of a LOD search functionality within the *Lin|gu|is|tik* portal.

5. The BLL Ontology as Linked Data

After the BLL Thesaurus has been remodelled as ontology, it will be linked with LLOD terminology repositories. For our initial focus areas, morphosyntax and syntax, the Ontologies of Linguistic Annotations (OLiA) represent the central terminology hub in the LLOD cloud (Chiarcos and Sukhareva, 2015). For other levels of linguistic analysis

beyond morphology, syntax and discourse, links with other LLOD vocabularies with the respective specialization are possible, as well, e.g., `http://lexvo.org` and `http://glottolog.org` for language identifiers, `http://phoible.org` for phonological features, etc. These are, however, beyond the scope of our initial work on morphology and syntax.

OLiA introduces a 'Reference Model' to mediate between resource-, domain- or language-specific 'Annotation Models', and several 'External Reference Models', i.e., community-maintained terminology repositories. OLiA Reference Model concepts are linked with externally provided terminology repositories, most notably GOLD (Farrar et al. 2010),[15] ISOcat (Kemps-Snijders et al. 2009)[16] and the TDS ontology.[17] Accordingly, any resource provided with an Annotation Model and linked with the Reference Model can also be interpreted in terms of these External Reference Models.
In this modular architecture, the BLL Ontology will be integrated in the same way as a conventional OLiA Annotation Model. This means that a new ontology is created, a 'Linking Model', which imports the BLL Ontology and the OLiA Reference Model and then assigns BLL Ontology concepts corresponding superconcepts from the OLiA Reference Model by means of `rdfs:subClassOf` properties. As a result, BLL concepts will immediately become interoperable with OLiA, GOLD, ISOcat, TDS, etc.
The automated conversion, the manual remodelling process and the linking of the BLL Thesaurus with OLiA and similar LLOD terminology repositories within the LLOD cloud results in multiple layers of interlinked ontologies:

1. The original SKOS export remains available along with the revised BLL Ontology, and the original structure of the thesaurus will be preserved in `skos:broader` relations.

2. The BLL Ontology with its manually remodelled class hierarchy provides a (partial) ontological interpretation of the BLL Thesaurus in its SKOS edition. It is physically separated from the original SKOS export but refers to the same URIs.

3. The BLL Linking Model connects the BLL Ontology with the OLiA Reference Model. Again, this is physically separated from both ontologies.

4. The OLiA Reference Model, as available from the LLOD cloud, provides reference semantics for morphological, morphosyntactic and syntactic concepts.

5. By means of OLiA Linking Models, multiple OLiA Annotation Models are linked with the OLiA Reference Model and provide resource-specific terms as used in annotated corpora or for grammatical features in lexical resources. Following the path from BLL over the OLiA Reference Model to these Annotation Models, we can scan LLOD resources for features corresponding to BLL concepts.

6. By means of additional Linking Models, the OLiA Reference Model is linked with GOLD, ISOcat, TDS, etc. Like OLiA, also these are used to facilitate conceptual interoperability between language resources, and as with OLiA Annotation Models, it is possible to identify concepts corresponding to BLL subject terms and then to use this information to retrieve LLOD resources that refer to these repositories to define their annotations.

[15] `http://linguistics-ontology.org`
[16] `http://isocat.org`
[17] `http://languagelink.let.uu.nl/tds/`

6. Status and prospects

As a dataset, the SKOS export of the BLL Thesaurus is already available, its current edition covering 5,340 subject terms, 2,141 language identifiers and a total of 55K SKOS triples. The BLL Ontology developed on top of it has been completed for the *Syntax* and *Morphology* branches, so far. Its OLiA linking is currently under development. We are in the process of clarifying details of a persistent hosting service and plan to publish the linked BLL Ontology under a Creative Commons licence in mid-2016, both for practical use in the *Lin|gu|is|tik* portal and for inclusion in the LLOD cloud.

In addition to this, we will create tools for the conversion of the bibliographic entries into RDF and develop a LOD-based search algorithm. Since the LLOD cloud consists of a vast variety of dispersed data sources, direct federated queries are likely to become inefficient. Therefore, we will implement an LLOD crawler and a database to aggregate and store BLL-related data from the cloud on a regular basis. A detailed description of the connection between the *Lin|gu|is|tik* portal and the LLOD cloud can be found in Chiarcos et al. (2016).

By implementating an interface between the BLL Thesaurus and the LLOD cloud, we will be able to take advantage of the immense opportunities the Semantic Web offers for linguistic research. By using LLOD vocabularies and term bases the *Lin|gu|is|tik* portal will gain access to an ever-growing pool of linguistic resources on the web, but also vice versa: LLOD cloud users will benefit from a significant new source of information.

Acknowledgments

We would like to thank three anonymous reviewers for insightful feedback and helpful comments. The research described in this paper was supported by the German Research Foundation (DFG) in the context of the project 'Virtual Library for General Linguistics' (Virtuelle Fachbibliothek Allgemeine Sprachwissenschaft) in the funding period 2015-2016.

References

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The SALSA corpus: A German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May.

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards Open Data for Linguistic: Linguistic Linked Data. In A. Oltramari, et al., editors, *News Trends of Research in Ontologies and Lexical Resources. Theory of Applications of Natural Language Processing*, pages 7–25. Springer, Heidelberg.

Chiarcos, C., Fäth, C., Renner-Westermann, H., Abromeit, F., and Dimitrova, V. (2016). Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Protorož, Slovenia.

Farrar, S. and Langendoen, D. T. (2003). A Linguistic Ontology for the Semantic Web. *GLOT International*, 7(3):97–100.

Farrar, S. and Langendoen, D. T. (2010). An OWL-DL implementation of Gold. An ontology for the Semantic Web. In A. Witt et al., editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*, pages 45–66. Springer, Dordrecht.

Farrar, S. and Lewis, W. (2007). The GOLD Community of Practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, 41(1):45–60.

Farrar, S. (2007). Using 'Ontolinguistics' for language description. In A. C. Schalley et al., editors, *Ontolinguistics. How Ontological Status Shapes the Linguistics Coding of Concepts*, pages 175–191. Mouton de Gruyter, Berlin + New York.

T. Givón, editor. (1994). *Voice and Inversion*. Typological Studies in Language, 28. John Benjamins, Amsterdam.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Writh, S. E. (2009). ISOcat: remodelling metadata for language resources. *International Journal of Metdata, Semantics and Ontologies*, 4(4):261–276.

Nordhoff, S. and Hammarström, H. (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the 1st International Workshop on Linked Science (LISC 2011), held in conjunction with the 10th International Semantic Web Conference (ISWC 2011)*, Bonn, Germany.

Nordhoff, S. (2012). Linked Data for Linguistic Diversity Research: Glottolog/Langdoc and ASJP Online. In C. Chiarcos, et al., editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 191–200. Springer, Berlin + Heidelberg.

Pastor, J. A., Martinez, F. J., and Rodriguez, J. V. (2009). Advantages of thesaurus representation using Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, 14(4). `http://InformationR.net/ir/14-4/paper422.html`, paper 422.

Renner-Westermann, H. (2013). Lin|gu|is|tik - Portal für Sprachwissenschaft. Webis. Aktuelles über Sammelschwerpunkte an deutschen Bibliotheken, `http://blogs.sub.uni-hamburg.de/webis/2013/08/01/linguistik-portal-fuer-sprachwissenschaft`.

Enhancing the Quality of Metadata by using Authority Control

Thorsten Trippel, Claus Zinn
Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
thorsten.trippel@uni-tuebingen.de, claus.zinn@uni-tuebingen.de

Abstract

The *Component MetaData Infrastructure (CMDI)* is the dominant framework for describing language resources according to ISO 24622 (ISO/TC 37/SC 4, 2015). Within the CLARIN world, CMDI has become a huge success. The Virtual Language Observatory (VLO) now holds over 800.000 resources, all described with CMDI-based metadata. With the metadata being harvested from about thirty centres, there is a considerable amount of heterogeneity in the data. In part, there is some use of controlled vocabularies to keep data heterogeneity in check, say when describing the type of a resource, or the country the resource is originating from. However, when CMDI data refers to the names of persons or organisations, strings are used in a rather uncontrolled manner. Here, the CMDI community can learn from libraries and archives who maintain standardised lists for all kinds of names. In this paper, we advocate the use of freely available authority files that support the unique identification of persons, organisations, and more. The systematic use of authority records enhances the quality of the metadata, hence improves the faceted browsing experience in the VLO, and also prepares the sharing of CMDI-based metadata with the data in library catalogues.

Keywords: Metadata quality, bibliographic metadata, authority records

1. Motivation

The Virtual Language Observatory (VLO) offers a faceted browser that helps users exploring linguistic resources at grand scale. At regular intervals, the VLO uses the OAI-PMH protocol to fetch metadata descriptions from about thirty partner organisations, ingests them into a single database, and offers a unified access to over 800.000 resources. While all partner organisations offer their metadata in CMDI, a huge data curation process is required to harmonise all data. Despite the common format, this data curation is by no means trivial. While some data providers make use of controlled vocabularies, for instance, to refer to country or language names, others use simple strings for this. Moreover, there are some data descriptors where strings are used by all parties, namely when referring to a person (say, as the creator of the resource) or an organisation (say, to describe where the resource has been created). Consider a user who uses the VLO to identify a resource in terms of the organisation it might be originating from. When the user asks the faceted browser to display a full list of organisations, the window in Fig. 1 (left) shows up. The user is confronted with, *e.g.*, four different spellings for the *Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)*. Whenever the user selects one of the four values, the metadata behind the other three spellings is automatically excluded from the search results, which is a rather unsatisfying user experience. As the screenshot indicates, this is not an isolated case. In the VLO, there are still hundreds of duplicates in the categories "organisation", "language" or "country", despite ongoing curation efforts such as CLAVAS [U3].

In the Library Sciences, where catalogues often contain millions of records from many different fields, the use of *authority files* is crucial. It allows librarians to associate alternative names with preferred ones. Fig. 1 (right) shows the authority record (taken from the *Gemeinsame Normdatei* of the German National Library, see below) for the BBAW. The record has a unique resource identifier, lists the organisation's preferred name, its alternative names, and other information such as the organisation's geographic location, or information about its predecessor and history. With an organisation being identifiable with a uniform resource identifier (URI) (such as `http://d-nb.info/gnd/2131094-4`), its name spelling becomes secondary. If all CMDI metadata providers would complement an organisation's name with a similar URI, the quality of all aggregated data would be greatly enhanced.

The unique identification of entity names is also highly relevant when linking CMDI-based metadata to external sources such as library catalogues and the linked open data initiative. It makes it possible to link the publications of a linguist with the linguistic resources he or she created.

2. Background

The German NaLiDa project[1] operates at the interface between subject field specific research infrastructures such as CLARIN and core infrastructures of research organisations such as libraries and computing centres. One aim is to define processes for ingesting metadata of linguistic resources (a subset of the VLO data) to the Tübingen Library Catalogue. In this process, we need to bridge the metadata standards used in the linguistics community with the dominant standards in the library world. The use of common URIs to refer to persons and organisations is such as bridge.

[1] NaLiDa is a project acronym abbreviating "**Na**chhaltigkeit **Li**nguistischer **Da**ten" (Sustainability of linguistic data).

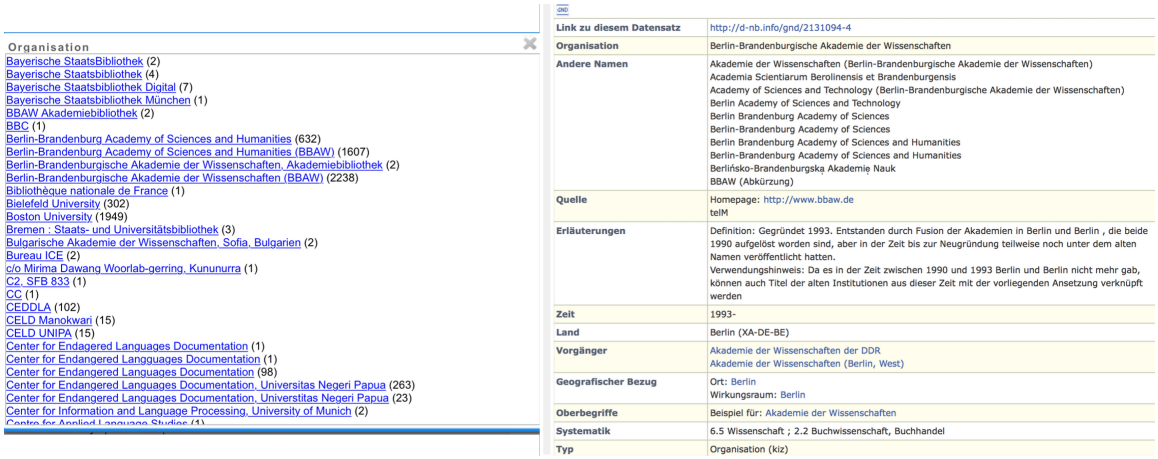


Figure 1: Organiation Duplicates in the VLO (left), the GND entry for the BBAW (right).

2.1. Metadata for language resources according to ISO 24622-1

The *Component MetaData Infrastructure (CMDI)* is a framework for the creation and use of metadata formats (CLARIN-D, 2012, page 19ff). Its abstract model follows an element-in-element, lego-brick approach to metadata modelling where schemas are defined by the selection and combination of predefined *data categories* and *components*. Data categories correspond to basic metadata elements or fields and are defined in the concept registry [U1], whereas components are hierarchically organized structures of data categories and components and are defined in the component registry [U2].

CMDI is the dominant framework for metadata in the CLARIN world, but it is also used by META-SHARE and other communities. At the time of writing, the CLARIN concept registry has about 1500 metadata terms and the CLARIN component registry offers over 1100 components with nearly 180 different public profiles (schemas).

With the rising numbers of resources described in CMDI, it is now time to adopt the use of authority files to uniquely describe the entities associated with the creation of the resources, and to hence complement string-based names for persons, corporate bodies *etc*. with authority records commonly used in the library world. This addresses the data heterogeneity issues described before.

2.2. The Use of Authority Files

In the library world, the use of authority files is good practise to identify persons, corporate bodies, but also subject headings. An authority file record gives a name in a standardised representation. It usually lists a person's (or organisation's) preferred name and complements it with alternative forms. Often, an authority record is associated with a unique resource identifier.

The **Integrated Authority File** (German: *Gemeinsame Normdatei (GND)*) is an international authority file for the organisation of persons, corporate bodies, conferences and events, geographic information, topics and works [U4]. It is maintained by the German National Library, and it has about 10 million entries, which includes over 2.5 million person names. The database is used widely in libraries, archives, and museums. It has a Creative Commons Zero (CC0) license.

The German National library also feeds the **Virtual International Authority File (VIAF)**, which is a joint project of several national libraries and is operated by the Online Computer Library Center (OCLC), see [U5]. The aim of VIAF is to link together the national authority files of all project members to a single virtual authority file. Each VIAF record is associated with a unique resource identifier and aggregates the information of the original authority records of the member states.

The **International Standard Name Identifier (ISNI)** is the "ISO certified global standard number for identifying the millions of contributors to creative works and those active in their distribution, including researchers, inventors, writers, artists, visual creators, performers, producers, publishers, aggregators, and more", see [U6]. It holds nearly 9 millions identities, including over 2.5 million names of researchers, and more than 500.000 organisation ids.

The US-American Library of Congress is another established authority file provider, see [U7]. More recent initiatives include the Open Researcher and Contributor ID (ORCID), see [U8], and ResearcherId, see [U9]. For geographical places, the GeoNames geographical database is widely used, see [U10].

All of these authority agencies attach a unique resource identifier to their records. Also, all agencies provide a RDF representation of records, so that it is possible to link together many data sources via the common format and the common use of identifiers. Note, for instance, that many Wikipedia biographical articles refer to the URIs of the aforementioned authority agencies.

3. Adding Authority Information to CMDI-based metadata descriptions

CMDI is a flexible framework making it easy to add provisions for authority records. For this, it is nec-

(a) GND record: Thorsten Trippel (b) GND record: Claus Zinn

Figure 2: The authors' names in the GND.

essary to use data descriptors in the CLARIN concept registry: `/issuingAuthority/` (added to the concept registry) and `/id/`. Values for the concept `/issuingAuthority/` must stem from a controlled vocabulary referring to the authority institutions that we currently support. Currently, we include VIAF, GND, ISNI, ORCID, LC and `geonames.org`.

In the CLARIN component registry, we define the component `/AuthoritativeID/`, which holds the aforementioned two data descriptors, and `/AuthoritativeIDs/`, which brackets one or more occurrences of `/AuthoritativeID/`.[2] References to authority files are modelled as pairs of unique resource identifier and authority, where we use the controlled name of the authority registering the identifier. Modelling the authority reference as a pair of identifier and issuing institution makes it easy to add other authorities when required.

Fig. 3 depicts the use of the new descriptional means when referring to a person. In the given case, we associated three different authority records to the string denoting the person *Erhard Hinrichs*. It shows that about 60% of all names occurring in our local CMDI instances can be complemented with information from authority records stemming from GND, VIAF or ISNI. Notably, all researchers with a PhD are covered. Note that all organisations in our local CMDI instances have corresponding GND, VIAF, or ISNI records.

In sum, the curation effort is manageable. Having modified the CMDI profiles, the metadata instances must be adjusted to adhere to their new profiles. First, all instances must now have a reference to the modified profile. Second, when persons, organisations and locations are given (usually as strings), those are complemented with corresponding references to their respective authority records. Given there is a (hand-made) table associating name strings with authority records, an XSLT style-sheet can be written to mechanise the updating of the CMDI instances.

Having associated authority file information with person names (*i.e.*, strings), some other bits of information often included in the CMDI metadata may become redundant, but not necessarily so. In fact, the affiliation of a person given in the original CMDI metadata may well be different to the affiliation of this person given in the authority record. The first affiliation indicates where the person worked when the resource in question was created; this is often more relevant than the person's affiliation given in the authority record.

[2]Due to the recursive nature of profiles, about a dozen of other CMDI components that contain references to names (such as `/Contact/` or `/Funder/`) were modified to include the new component. Note that, at the time of writing, the new concepts and components reside in the private space of the CLARIN registries.

4. Discussion

The use of authority files greatly improves the quality of the CMDI-based metadata. Persons and corporate bodies are now uniquely identifiable. When CMDI data providers adopt authority files, we will see two main benefits: (i) an improvement in search through aggregated data sources within the CLARIN Virtual Language Observatory (especially wrt. organisations), and (ii) a better linking to library catalogues which use the same authority file information. The latter makes it possible to find a researcher's entire work (traditional publications and research data) with a single query. Data sharing at the URI level pays off.

We have seen that authority records may contain information about a person's birthdate, sex, academic degree, or profession. The record may give a reference to a geographical location (where the person works or has worked). Some of this information will not be up to date, see for instance,

```
<Person>
    <firstName>Erhard</firstName>
    <lastName>Hinrichs</lastName>
    <Role>Projektleiter</Role>
    <AuthoritativeIDs>
        <AuthoritativeID>
            <id>http://viaf.org/viaf/37069402</id>
            <issuingAuthority>VIAF</issuingAuthority>
        </AuthoritativeID>
        <AuthoritativeID>
            <id>http://d-nb.info/gnd/143840657</id>
            <issuingAuthority>GND</issuingAuthority>
        </AuthoritativeID>
        <AuthoritativeID>
            <id>http://isni.org/0000000118749683</id>
            <issuingAuthority>ISNI</issuingAuthority>
        </AuthoritativeID>
    </AuthoritativeIDs>
</Person>
```

Figure 3: Example fragment for the new encoding for a person name.

the informaton given in Fig. 2(a).[3] Here, exitisting CMDI metadata may well overwrite or complement the information associated with a person's authority record.
With over 2.5 million person names in the GND, there are entries that share the same name. By coincidence, a GND search for "Claus Zinn" shows two entries. The second entry is less specific than the one given in Fig. 2(b); it may well be an unwanted duplicate. In fact, associating the correct authority file with a given name is often facilitated by the additional information the record contains, in particular, the person's profession or associated publications.
To our knowledge, only libraries can directly enter or update existing authority file information. Note that the German National Library allows users to easily request a correction or actualisation of their GND entries (each GND record is displayed with an action "request correction").

So far, the CMDI community makes little use of metadata standards and controlled vocabularies used elsewhere. There are three major avenues to develop CMDI toward other metadata standards, and to bring CMDI closer to the library world, and subsequently toward the Semantic Web:

- making available tools that map CMDI to other metadata standards, in particular, towards the dominant standards in the library world such as Dublin Core [U11] and MARC 21 [U12].
- making available conversion tools that convert the XML-based CMDI representations to RDF-based representations where all information is expressed in terms of RDF triples.
- using unique resource identifiers to refer to persons, corporations, and geographical places.

Existing work tackles the first and second aspect of opening up CMDI to the metadata world. (Ďurčo and Windhouwer, 2014) propose a conversion from CMDI to RDF, and (Zinn et al., 2016) propose crosswalks between CMDI-based profiles and the library metadata standards MARC 21 and Dublin Core. In isolation, none of the work yields results that a librarian will be entirely happy with. Having a CMDI-based record converted to MARC 21 helps its ingestion in the library catalogue, but without authority information the new information is not linked to any prior information in the catalogue (*e.g.*, common author or common publisher). Similarly, while having a CMDI-based record be expressed in RDF has a number of advantages (*e.g.*, common data format with other data sources, RDF-based technology for storing or querying data sets), a true conversion of CMDI-based RDF data requires data sharing at the URI level.
In this paper, we have addressed the third aspect, incorporating authority records into CMDI-based metadata descriptions. This vastly improves the conversion to MARC 21 and RDF, and it also strengthens the links to other datasets. We encourage other CMDI metadata providers to follow our steps.

[3]The GND record has Trippel's geographic location given as *Bielefeld*, which was true at a time when he wrote his PhD thesis.

We also encourage the CLAVAS initiative, which seeks to produce a curated list of organisations based on the CLARIN VLO, see [U3], to associate with each organisation a reference to the respective record from the GND database. In fact, the many alternative names present in an organisation's GND record could be used to partially automate the mapping process.

For the long term, the CLARIN consortium may want to consider a metadata policy that propagates (or even enforces) the use of authority records in CMDI-based metadata.

Note. All our CMDI metadata, enriched with authority file information, will soon be available in the CLARIN VLO.

Acknowledgments. The NaLiDa project has been funded by the German Research Foundation, reference numbers DO 1346/4-2, WA 3085/1-2, and HI 495/4-2.
We would like to thank the anonymous referees for their comments, which helped improve this paper considerably.

Web Resources

[U1] The CLARIN Concept Registry, see `openskos.meertens.knaw.nl/ccr/browser`
[U2] The CLARIN Component Registry, see `catalog.clarin.eu/ds/ComponentRegistry`
[U3] The CLAVAS OpenSKOS Vocabulary Service, see `openskos.meertens.knaw.nl/clavas`
[U4] The Integrated Authority File of the German National Library, see `www.dnb.de/EN/Standardisierung/GND/gnd_node.html`
[U5] The Virtual International Authority File, see `viaf.org`.
[U6] The International Standard Name Identifier, see `isni.org`.
[U7] The Library of Congress Control Number, see `id.loc.gov/authorities/names.html`.
[U8] The Open Researcher and Contributor ID, see `orcid.org`.
[U9] ResearcherId, see `www.researcherid.com`.
[U10] The GeoNames database, see `geonames.org`.
[U11] The Dublin Core Metadata Initiative, see `www.dublincore.org`.
[U12] The MARC 21 standard, see `www.loc.gov/marc/bibliographic`.

5. Bibliographical References

CLARIN-D. (2012). The CLARIN-D user guide. `http://media.dwds.de/clarin/userguide/text`.
ISO/TC 37/SC 4, (2015). *ISO 24622-1:2015(en) Language resource management Component Metadata Infrastructure (CMDI) Part 1: The Component Metadata Model*. International Organization for Standardization.
Ďurčo, M. and Windhouwer, M. (2014). From CLARIN component metadata to linked open data. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 24–28. Co-located with LREC 2014. 26.-31 May 2014, Reykjavik. ELRA.
Zinn, C., Trippel, T., Kaminski, S., and Dima., E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In *Proceedings of LREC 2016, Portorož*. ELRA.

Developing and Using the Ontologies of Linguistic Annotation (2006-2016)

Christian Chiarcos, Christian Fäth, Maria Sukhareva
Goethe-Universität Frankfurt am Main, Germany
{chiarcos|faeth|sukharev}@informatik.uni-frankfurt.de

Abstract

This paper describes the Ontologies of Linguistic Annotation (OLiA) as one of the data sets currently available as part of Linguistic Linked Open Data (LLOD) cloud. Within the LLOD cloud, the OLiA ontologies serve as a reference hub for annotation terminology for linguistic phenomena on a great band-width of languages, they have been used to facilitate interoperability and information integration of linguistic annotations in corpora, NLP pipelines, and lexical-semantic resources and, they mediate their linking with multiple community-maintained terminology repositories.

This paper summarizes a decade of research on OLiA (2006-2016), it provides an overview over design, progress, recent applications and prospective developments, and introduces two novel applications of OLiA.

Keywords: linguistic terminology, annotation interoperability, ontology-based approaches

1. Background

The heterogeneity of linguistic annotations has been recognized as a key problem limiting the (re-)usability of NLP tools and language resources. Since the early 2000s, several repositories of linguistic annotation terminology have been developed to facilitate annotation interoperability by means of a joint level of representation, most notably ISOcat (Kemps-Snijders et al., 2009), the General Ontology of Linguistic Description (Farrar and Langendoen, 2010, GOLD), and the TDS ontology (Saulwick et al., 2005). However, these have been maintained by separate communities and addressed their respective needs only, thereby achieving a limited degree of interoperability only: ISOcat accompanied the development of ISO TC37/SC4 standards and thus a technological bias. It was heavily used in the CLARIN project, so that it attracted especially the European community. GOLD, on the other hand, was originally developed for the language documentation community and is currently maintained by the Linguist List. It thus attracted researchers from linguistics rather than NLP, and particularly from the American community. Finally, TDS was created for a different, – and specialized – use case, the search in typological databases.

The Ontologies of Linguistic Annotation (OLiA) have originally been created as a mediator between these and other terminology repositories on the one hand and linguistically annotated resources on the other hand (Schmidt et al., 2006). However, with GOLD and TDS development stalled since 2010,[1] and ISOcat development frozen as of December 2014,[2] OLiA becomes increasingly important as a terminology repository in its own right for both Natural Language Processing and linguistics.

OLiA applies Linked Data principles to leverage several, distributed terminology repositories. It thus represents a central terminology hub for annotation terminology within the Linguistic Linked Open Data (LLOD) cloud already since its conception in 2010.

We provide an overview over the status and use cases of OLiA since its original publication (Schmidt et al., 2006), with a special emphasis on developments since Chiarcos (2010a). In addition, we elaborate on two novel applications of OLiA. We describe the general architecture (Sect. 2.) and elaborate on OLiA-specific design issues of the ontology (Sect. 3.), followed by a discussion of earlier, recent and prospective applications of OLiA with respect to three major functions: terminology management and documentation (Sect. 4.), resource modeling and access (Sect. 5.), and natural language processing (Sect. 6.).

2. Architecture

The **Ontologies of Linguistic Annotations** (Chiarcos, 2008) represent a modular architecture of OWL2/DL ontologies that formalize the mapping between annotations, a 'Reference Model' and existing terminology repositories ('External Reference Models'). The OLiA ontologies are available from `http://purl.org/olia` under a Creative Commons Attribution license (CC-BY).

The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al., 2006), where their primary fields of application included the formalization of annotation schemes and concept-based querying over heterogeneously annotated corpora (Rehm et al., 2008; Chiarcos et al., 2008). As multiple institutions and manifold resources from several disciplines were involved, no holistic annotation standard could be developed and enforced onto our contributors. Instead, we designed a modular architecture that allows to integrate the annotation terminology of different resources in a lossless and reversible way.

In the OLiA architecture, as illustrated in Fig. 1 four different types of ontologies are distinguished (cf. Fig. 1 for an example):

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is based on existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.

[1] See `http://linguistics-ontology.org/version`, `http://languagelink.let.uu.nl/tds/ontology/LinguisticOntology.owl`

[2] See `http://www.isocat.org`

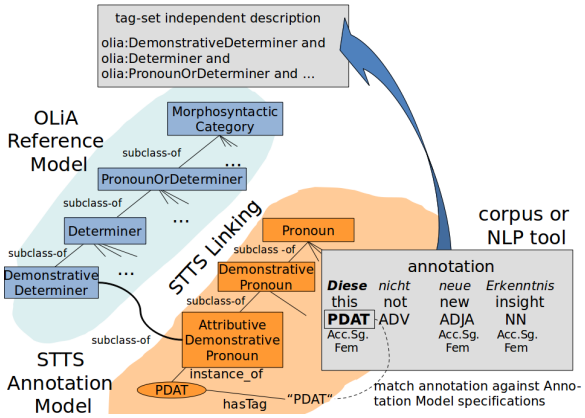


Figure 1: Interpreting annotations in terms of the OLiA Reference Model

- Multiple OLiA Annotation Models formalize annotation schemes and tagsets. Annotation Models are based on the original documentation, so that they provide an interpretation-independent representation of the annotation scheme.
- For every Annotation Model, a Linking Model defines $\sqsubseteq$ relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model.
- Existing terminology repositories can be integrated as External Reference Models, if they are represented in OWL2/DL. Then, Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g., `olia:Determiner`) and grammatical features (e.g., `olia:Accusative`), as well as properties that define relations between these (e.g., `olia:hasCase`).

Conceptually, Annotation Models differ from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model. Figure 1 gives an example for the individual `PDAT` from the STTS Annotation Model, the corresponding STTS concepts, and their linking with Reference Model concepts. Taken together, these allow to interpret the individual (and the part-of-speech tag it represents) as an `olia:Determiner`, etc.

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as extensions for discourse annotations (coreference, discourse relations, discourse structure and information structure). Annotations for lexical semantics are only covered to the extent that they are found in syntactic and morphosyntactic annotation schemes. Other aspects of lexical semantics are beyond the scope of OLiA.[3]

At the time of writing, the OLiA Reference Model distinguishes 280 `MorphosyntacticCategory` concepts (word classes), 68 `SyntacticCategory` concepts (phrase labels), 18 `MorphologicalCategory` concepts (morphemes), 7 `MorphologicalProcesses`, and 405 different values for 18 `MorphosyntacticFeatures`, 5 `SyntacticFeatures` and 6 `SemanticFeatures` (for glosses, part-of-speech annotation and for edge labels in syntax annotation).

As for morphological, morphosyntactic and syntactic annotations, the OLiA ontologies include 32 Annotation Models for about 70 different languages, including several multilingual annotation schemes, e.g., EAGLES (Chiarcos, 2008) for 11 Western European languages, and MULTEXT/East (Chiarcos and Erjavec, 2011) for 15 (mostly) Eastern European languages. As for non-(Indo-)European languages, the OLiA ontologies include morphosyntactic annotation schemes for languages of the Indian subcontinent, for Arabic, Basque, Chinese, Estonian, Finnish, Hausa, Hungarian and Turkish, as well as multilingual schemes applied to languages of Africa, the Americas, the Pacific and Australia. The OLiA ontologies also cover historical varieties, including Old High German, Old Norse, Old English and Old Tibetan. Additionally, 7 Annotation Models for different resources with discourse annotations have been developed.[4] Recent extensions include prototypical Annotation and Linking Models for the language-specific editions of the Universal Dependencies,[5] eventually extending OLiA coverage with 19 additional treebanks and 4 additional languages. Furthermore, a linking with LexInfo (Cimiano et al., 2011), a vocabulary widely used among lexical-semantic resources in the LLOD cloud, is currently under development.

External reference models currently linked to the OLiA Reference Model include GOLD (Chiarcos, 2008), the OntoTag ontologies (Buyko et al., 2008), an ontological remodeling of ISOcat (Chiarcos, 2010a), and the Typological Database System (TDS) ontologies (Saulwick et al., 2005). A prototype for the language-independent specifications of the Universal Dependencies has been developed at the EUROLAN-2015 summer school on Linguistic Linked Open Data. Its integration with the Universal Dependencies has been requested and waits for approval from the UD community.[6]

As compared to a direct linking between annotation models and these terminology repositories, the modular structure limits the number of linkings that need to be defined (if a new Annotation Model is linked to the Reference Model,

[3] Existing reference resources for lexical semantics available in RDF include WordNet, VerbNet and FrameNet which are recommended for the purpose. In the edition by (Eckle-Kohler et al., 2015), the morphosyntactic features of these resources are defined with reference to OLiA.

[4] `http://purl.org/olia/discourse`

[5] `http://universaldependencies.github.io/docs/`

[6] `https://github.com/UniversalDependencies/docs/pulls`

it inherits its linking with ISOcat, GOLD, OntoTag, TDS, etc.).

3. Design issues

We discuss selected modeling decisions specific to OLiA, i.e., (a) a reified (concept-based) representation of relational annotations, (b) modeling and interpretation of tags and instances, (c) peculiarities of `has Feature` properties, and (d) the limited use of cardinality and disjointness axioms.

3.1. Representing relational annotations

Relational annotations are of utmost importance to formalizing, for example, the syntactic structure of languages with free word order. This is not only true for classical dependency syntax, but also for languages whose adequate representation of phrase structure requires excessive use of empty elements (such as traces or zero pronouns). Accordingly, also phrase structure grammars introduced edge labels much alike the labels used in dependency annotation, e.g., for German (Brants et al., 2002) or Old English (Taylor, 2007).

Edge labels are thus equally important for both dependency and constituency annotations, and ideally, they represent the (annotation) semantics of the relationship between the elements linked. Taking the classical Stanford Dependencies (De Marneffe and Manning, 2008) as an example, all edge labels are specializations of the underspecified *dep* relation. However, not all Stanford dependency labels are truly relational, some merely represent morphosyntactic characteristics of the dependent independently from its syntactic head, e.g., *discourse* (element), *expl*(etive), *predet*(erminer), etc. Similarly, token-based annotations of syntactic relations can be found as well, e.g., Petrova et al. (2009) annotate *grammatical function* as a property of token(span)s rather than relations.

For designing OLiA, we thus cannot rely on any systematic differentiation between relational and token-/span-based annotation but must be able to accommodate resources that deviate from our a priori assumptions. Nevertheless, we aim to harmonize both ways of representing relational annotations in a model with consistent data types. In RDF, this is only possible with an instance/class-based representation of annotation concepts, as object properties (if applied to represent relational annotations) can be represented as individuals by means of RDF reification, but not the other way around.

OLiA thus models 'inherently' relational annotations by means of individuals and classes rather than properties, thereby enforcing a reified representation of syntactic relations in RDF as employed, for example, by POWLA (Chiarcos, 2012b).

While this allows us to generalize over *any* kind of edge annotation, it poses a challenge for approaches which represent syntactic relations by means of an `ObjectProperty` in a straight-forward fashion. The formally correct solution with RDF reification yields a less concise representation, so that, for example, the NIF wrapper of Stanford Core NLP[7] uses a modified OLiA Annotation Model whose classes and individuals have been converted to properties.[8] This modified Annotation Model allows NIF to continue using an OLiA Annotation Model for representing edge labels, but its linking with the OLiA Reference Model involves a type conversion, thereby breaking OWL2/DL constraints.

3.2. Modeling tags as instances

In an OLiA Annotation Model for a small-scale tagset, every tag is represented by a single individual, characterized by the Annotation Model concept(s) it is assigned to, by its string representation and, optionally, by a description.

Yet, this 'classical' approach only permits us to cover annotation schemes with up to a few hundred individual tags. For morphologically rich languages, larger-scale part-of-speech tagsets have been designed which incorporate numerous morphosyntactic features whose combinations generate tagsets with thousands of tags, e.g., MULTEXT-East.[9] Other part-of-speech tagsets have been enriched with syntactic and semantic information, leading to large-scale annotation schemes for morphologically poor languages, as well, e.g., Sampson (2002) for English.

To 'decompose' positional annotation schemes efficiently into morphosyntactic categories, morphological features, etc., we extended the original semantics of individuals to represent groups of tags which share parts of their string representation. For this purpose, the OLiA system subontology[10] provides the properties `hasTagContaining`, `hasTagStartingWith`, and `hasTagEndingWith` for matching substrings in a tag, and `hasTagMatching` for full-fledged regular expressions, in addition to `hasTag` for literal matches. Note that this not only permits mapping one individual to a number of tags, but a full $n : m$ mapping between where every tag can be assigned multiple individuals. If an actual tag matches multiple individuals, this should be interpreted such that the element the tag applied to inherits their definitions and falls in the intersection of their respective superclasses.

As OLiA applications like Apache Stanbol[11] rely on `hasTag` properties alone, it is recommended to compile the `hasTag`X properties into `hasTag` properties: Using an annotated corpus, we bootstrap an exhaustive list of tags and generate `hasTag` properties for all tags matching a particular `hasTag`X pattern.

Although individuals are thus capable to represent groups of tags, we preserved the instance-based (rather than a class-based) modeling in accordance with the strong typing in OWL2/DL.[12] Still, a class-based model would have the advantage that words or token spans can be directly assigned an annotation as their `rdf:type`. In fact, we do not

[7] `http://site.nlp2rdf.org/demo/`

[8] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/vm/dep/stanford#`

[9] `http://nl.ijs.si/ME/`

[10] `http://purl.org/olia/system.owl`

[11] `https://stanbol.apache.org/docs/trunk/components/enhancer/nlp/nlpannotations`

[12] User-defined properties such as `hasTag`X should be applied to individuals only, not OWL classes. Otherwise, OWL classes would be re-cast as individuals. While this may be tolerated by reasoners, it represents a design flaw.

exclude this possibility, as users are free to develop Annotation Models where every individual is defined an instance of a singleton class, so that this single-tag class can be assigned as type.
This discussion, touches the core semantics of OLiA Annotation Model individuals: They do not necessarily provide reference semantics for individual tags, but they act as entry points to OLiA Reference Model concepts. For part-of-speech annotation, individuals may thus represent either

1. individual tags (`hasTag`), or
2. patterns defining a mapping from tags to potentially complex type definitions (`hasTag`X).

These definitions overlap for the case of singleton classes mentioned above, but they have different implications for possible references to OLiA Annotation Models: Under the first interpretation, it is possible to refer to tags from external resources as target of a designated object properties. OLiA does not define such a property, but `nif:oliaLink`[13] has been designed for this purpose. Under the second interpretation, individuals can be used by a tool developer to aggregate the definition of all matching individuals in a conjunction ($\sqcap$), and then to assign this as a complex type to his (application-specific) unit of analysis using `rdf:type`.
As a general-purpose repository of linguistic annotation terminology, OLiA stays deliberately agnostic about these interpretations and permits both kinds of references, using `nif:oliaLink` or direct `rdf:type`. As we explicitly permit the second interpretation, it is possible to assign entities of *any* type an OLiA Annotation Model class: OLiA semantics are thus not limited to tag semantics, but cover any entity such annotations can are applied to. OLiA semantics thus refer only to linguistic characteristics of arbitrary entities, but remain *underspecified* with respect to their material manifestation.
Ît is thus equally possible to assign an OLiA class as a type to a word in a text, to an annotation attached to this word, to a lexical entry in a dictionary, to a lexeme in a language, to a term in a grammatical treatise or to a concept in a terminological resource.

3.3. `hasFeature` properties

One of the original use cases of OLiA was its application for facilitating corpus querying. As an example, we would like to support the following "naive" query for a plural noun: `Noun` $\sqcap$ `Plural`. For tagsets like that of the Penn Treebank, this can be easily achieved by defining `NNS` $\sqsubseteq$ `Accusative` (and `CommonNoun`), etc. Other annotation schemes, however, do not group grammatical features with parts of speech,[14] and for these, it is necessary to use the respective property `olia:hasNumber` for querying, as they do not refer to the same individual (i.e., annotation string): `Noun` $\sqcap$ $\exists$`hasNumber.Plural`

[13] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#oliaLink`

[14] For example, interlinear glosses combine grammatical features with English hyperlemmata `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`

While the implementation-specific mapping of multi-layer annotations to OLiA properties is beyond the scope of this paper, OLiA should allow this query to return comparable results for POS+gramm annotations as common in NLP. In order to do so, OLiA concepts for feature values comprise reflexive axioms, e.g., `Number` $\sqsubseteq$ $\exists$`hasNumber.self`. `Plural` inherits this axiom, so that the second query returns the same results as the first if run against the Penn Treebank. It is thus not necessary to specify the property explicitly in the linking.
The inherent reflexivity of subproperties of `olia:hasFeature` is a domain-specific adaptation that facilitates both intuitive linking and querying for non-ontologists.

3.4. Limits of axiomatization

Rendering annotation terminology in an OWL ontology naturally invites proposals for axiomatizing these terms, e.g., by formalizing dependencies between grammatical features and linguistic categories by domain assignments or cardinality axioms, e.g., that gender is a property of adjectives and nouns, only. Rules of this kind are provided, for example, by MULTEXT/East or the Universal Dependencies.
Yet, this is limited in the OLiA Reference Model: OLiA provides language-independent terminology, OLiA Reference Model axioms involve feature-concept assignments only if inherent to the definition of a concept, e.g., `PastParticiple` $\equiv$ `Participle` $\sqcap$ $\exists$`hasTense.Past`.
Beyond this, the OLiA Reference Model does not provide axioms regarding conventional associations between categories and features, as it will inevitably lead to inconsistencies when directly applied to *existing* corpora, dictionaries and annotations.[15]
This can be illustrated by providing counter-examples to commonly accepted assumptions that represent candidates for axiomatization:

`Adverb` $\sqsubseteq$ $\not\exists$**`hasPerson`** Adverbs do not have person agreement? Some pronominal adverbs in German do, e.g., *meinetwegen* 'because of me', *deinetwegen* 'because of you (sg.)', *seinet-/ihretwegen* 'because of him/her'.

`NonFiniteVerb` $\sqsubseteq$ $\not\exists$**`hasTense`** Non-finite verbs do not have tense? Actually, English has past and present participles, which may be modeled as having morphological tense.

`FiniteVerb` $\sqsubseteq$ $\not\exists$**`hasGender`** Finite verbs do not have gender agreement? The simple past in Russian does:

[15] A wider use of cardinality axioms is feasible only for a limited domain, where certain conventions or phenomena can be taken for granted (e.g., Eastern Europe), or within a standardization approach that aims at actively transforming existing resources towards common specifications in a labor-intense process, cf. `http://universaldependencies.org/introduction.html`. In comparison to MUILTEXT/East, OLiA is not limited to a geographic area. In comparison to Universal Dependencies, it is a light-weight approach that does not require data transformation.

čitat' 'to read', *on čita**l*** 'he read', but *ona čita**la*** 'she read'.

This list can be further extended, in particular if more exotic languages are considered. Many of these apparent exceptions arise from language-specific grammaticalizations (the German *-(t)wegen* adverbs are originally prepositional phrases with a pronominal head, and Russian past forms originate from participles). Also note that grammatical features may be interpreted differently, and as a resource-, language- and theory-independent resource, the OLiA Reference Model has to be underspecified in this regard. A property like `hasTense` may be defined either with a focus on (morpho)syntax (as a property of finite verbs), or with a focus on morphology (e.g., whether the present or the past form of an English verb stem is used to form a particular participle). OLiA permits both possibilities in order to allow a lossless and ontologically consistent representation of features specified in associated resources.

Accordingly, OLiA does neither restrict the domain of its properties nor does it provide cardinality axioms requiring or prohibiting the assignment of grammatical features to instances of a particular concept.

For similar reasons, OLiA does provide very few disjointness axioms only. In the reality of linguistic annotation, categories may overlap, so that a language-independent and clear-cut differentiation between, say, participles in attributive use and deverbal adjectives cannot be taken for granted in resources as currently provided.[16]

As its function is to mediate between annotations and full-fledged terminology repositories, OLiA is designed to stay agnostic about such axioms and expects them to be either inherited from External Reference Models (e.g., GOLD), or to be provided in language-specific or domain-specific submodels (e.g., MULTEXT/East for Eastern Europe or the Universal Dependencies for their accompanying corpora).

4. Documenting and Retrieving Language Resources

OLiA has been employed to provide reference terminology since its beginning. Accordingly, is has served a **documentation function** (Schmidt et al., 2006), originally with a focus on linguistic annotations and NLP tools, but subsequently extended to morphological dictionaries (Chiarcos and Erjavec, 2011) and lexical resources (Eckle-Kohler et al., 2015). With the ISO TC37/SC4 Data Category Registry (ISOcat) in re-orientation and its development currently stalled, this function is expected to rise in importance.

4.1. Resource documentation

From the ontologies, dynamic HTML can be generated, and tags in the annotation can be represented as hyperlinks pointing to the corresponding definition (Chiarcos et al., 2008). Figure 2 shows a screenshot of the HTML version of the OLiA Annotation Models of the MULTEXT/East morphosyntactic specifications (Chiarcos and Erjavec, 2011).

[16]This problem is also mentioned as being unresolved in the definition of VERB in the Universal Dependencies, `http://universaldependencies.org/u/pos/VERB.html`.

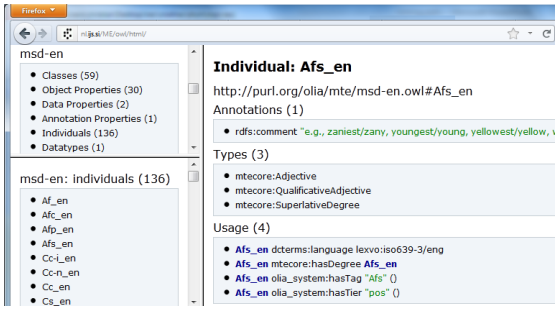


Figure 2: HTML version of the OLiA Annotation Model for the MULTEXT/East morphosyntactic specifications for English, `http://purl.org/olia/mte`.

4.2. A novel application: A Virtual Library

In the function to **retrieve language resources** by means of their metadata, we currently integrate OLiA with the *Lin|gu|is|tik web portal*,[17] a virtual library for the entire field of linguistics, largely grounded on the Bibliography of Linguistic Literature (BLL). This is subject to a recently started, DFG-funded project in collaboration with the University Library of the Goethe University Frankfurt. Established in 1971, the BLL is one of the most comprehensive linguistic bibliographies worldwide with an annual growth of about 10,000 references. BLL provides a hierarchically categorized bilingual German-English thesaurus of linguistic terms used for indexing online resources and the bibliography. It comprises 5340 subject terms and 2141 language identifiers organized in a hierarchical structure. Since July 2015, an LOD edition of the BLL Thesaurus is being produced: Following an automated conversion from OCLC PICA to SKOS, `skos:broader` properties are manually inspected and transformed into `rdfs:subClassOf` relations in an effort to create a full-fledged ontology (Pastor et al., 2009).

By Feb 18th, 2016, the SKOS edition of the BLL Thesaurus comprises 55,048 triples, the BLL Ontology focuses on morphology and syntax with a current coverage of 15% of the original subject terms (775/5340). For the BLL layers of morphology and syntax, we expect a preliminary linking with the OLiA Reference Model by mid-2016. By providing a explicit linking model with the OLiA Reference Model, we augment the BLL thesaurus with reference semantics. In addition, OLiA acts as an interface with the LLOD cloud, and BLL concepts become interoperable with OLiA, GOLD, ISOcat, TDS, etc.

On this basis, we plan to extend the *Lin|gu|is|tik web portal* with an additional search functionality to access LLOD resources. For this purpose, an **LLOD crawler** is currently being implemented. Its architecture comprises four interdependent components all using a central triple store for gathering information:

- A local, live mirror of OLiA, including Linking, Annotation Models, and BLL linking.

[17]`http://linguistik.de/`

- A local, live mirror of the LingHub RDF data[18] that is used to generate the LLOD cloud diagram.
- A crawling routine scanning the LingHub cache for `dcat:accessURL` resp. `dcat:downloadURL` links to LLOD resources. These data sets are subsequently scanned for BLL concepts, resp. concepts from OLiA, OLiA Annotation Models and terminology repositories linked with OLiA. As a result, we generate a continuously growing mapping of BLL concepts to LLOD resources.
- A new search functionality in the *Lin|gu|is|tik web portal* that allows end users to access the output data of the crawler.

In late 2016, the LLOD crawler will be published and the novel search functionality added to the *Lin|gu|is|tik web portal*. As a result, end users will be able to formulate queries that retrieve not only publications, but also corpora with appropriate annotations.

5. Modeling and Accessing Resources

In addition to its documentation function, OLiA can establish conceptual interoperability between language resources, e.g., the same queries can be applied to corpora with different annotation schemes.

5.1. Corpora and NLP tools

Figure 1 illustrates how annotations can be mapped onto Reference Model concepts for the German phrase *Diese nicht neue Erkenntnis* 'this well-known (lit. not new) insight' from the Potsdam Commentary Corpus (Stede, 2004, file 4794): Given the information that its part-of-speech annotations follow the STTS scheme (Schiller et al., 1999), we may consult the corresponding Annotation Model,[19] and find that the tag `PDAT` matches the string value of the property `hasTag` of the individual `stts:PDAT`. The associated class `stts:AttributiveDemonstrativePronoun` is a subconcept of `olia:DemonstrativeDeterminer`.[20] The word *diese* 'this' from the example can thus be described in terms of the OLiA Reference Model as `olia:DemonstrativeDeterminer`, etc.

These ontology-based descriptions are comparable across different corpora and/or NLP tools, across different languages, and even across different types of language resources. For **querying annotated corpora**, OLiA has been used for ontology-based query rewriting, e.g., in an extension of ANNIS (Chiarcos and Götze, 2007): Assume we wanted to retrieve noun phrases from German newspaper corpora; instead of querying for `cat="NX"` on TüBa-D/Z (Telljohann et al., 2003) *or* `cat="NP"` on NEGRA (Skut et al., 1998), a query for `cat in {olia:NounPhrase}` can be expanded into a disjunction of possible tags and formatted according to the query language under consideration. Ontology-based query rewriting can be applied to corpora in any format, it was implemented, for example, in a generic query framework for linguistic corpora in heterogeneous XML-formats (Rehm et al., 2008). An important drawback, however, is that corpus queries are expanded into potentially huge disjunctions which are then run against a corpus querying engine. These disjunctions may become intractable, and have thus to be heuristically simplified, e.g., by condensing alternative tags into more abstract regular expressions.

As such heuristics may, however, lead to losses and inconsistencies, we explored a different route, i.e., to use Semantic Web formalisms for an *integrated* representation of corpora and annotations. Instead of a solution relying solely on query rewriting, this allows for a more flexible combination of backward chaining (query expansion) and forward-chaining (data enrichment) techniques (Chiarcos, 2012a). With corpora becoming increasingly available as part of the Linguistic Linked Open Data cloud (McCrae et al., 2016), they can be directly linked with OLiA Annotation Models, and then queried with off-the-shelf technologies. For efficiency, however, we recommend forward-chaining within the OLiA Reference Model, and to rely on more lightweight solutions like SPARQL 1.1 property paths or RDFS reasoning for full OLiA with Annotation and Linking Models.

In a similar vein, OLiA is employed in **NLP pipeline systems** for tagset-independent, interoperable information processing (Hellmann et al., 2013; Hahm et al., 2012). In this function, OLiA is part of the NLP Interchange Format (NIF) specification[21] to formalize linguistic annotations in a conceptually interoperable way. Using OLiA, the NLP2RDF platform developed on this basis unifies various NLP result outputs and maps them into RDF. This is closely related to developing RDF-native corpus querying engines, as NIF – even though limited to single-layer annotations –, has also been applied to represent annotated corpora (Siemoneit et al., 2015), along with other, more advanced proposals based on OWL2/DL (Chiarcos, 2012a), RDFa (Rubiera et al., 2012), or Open Annotation (Verspoor et al., 2015).

5.2. Lexical-semantic resources

In addition to corpora, the OLiA ontologies have been applied to represent grammatical specifications of **machine-readable dictionaries**, that thus became interoperable with OLiA-linked corpora (McCrae et al., 2011; Eckle-Kohler et al., 2015). At the same time, the development of the LLOD community lead to the extension and the establishment of other term bases for lexical resources, most notably `lexinfo.org`, which has been employed to provide linguistic reference concepts for lexical resources in the LIDER project.[22] At the moment, the LLOD cloud lists 31 resources using/linked with LexInfo. By providing a Linking Model, LexInfo is currently being integrated into the modular OLiA architecture as a domain-specific model in the same way as an Annotation Model. As LexInfo is based on ISOcat, this linking model is partially boot-

[18] `http://linghub.lider-project.eu/linghub.nt.gz`
[19] `http://purl.org/olia/stts.owl`
[20] `http://purl.org/olia/stts-link.rdf`
[21] `http://persistence.uni-leipzig.org/nlp2rdf/`
[22] `http://lider-project.eu`

strapped from the existing OLiA-ISOcat linking (Chiarcos, 2010a).

6. Ontology-based NLP

Using Semantic Web formalisms to represent corpora and annotations provides us with the possibility to develop novel, **ontology-based NLP** algorithms.

6.1. Tagset integration and ensemble combination

One application are ensemble combination architectures, where different NLP modules (say, part-of-speech taggers) are applied in parallel, so that they produce annotations for one particular phenomenon, and that these annotations are then integrated. Using OLiA Reference Model specifications to integrate the analyses of multiple NLP tools for German, Chiarcos (2010b) showed that a simple majority-based combination increased both the robustness and the level of detail of morphosyntactic and morphological analyses: Despite imposing rigid ontological consistency constraints, abstraction from tool-specific representations and integration of different annotations on this basis resulted in an increase of recall. Similar results have been obtained with the OntoTag ontologies for Spanish (Pareja-Lora and Aguado de Cea, 2010).

We see possible applications of this technology in situations where multiple, domain-specific NLP tools are available. In a monolingual setting, this may be the case where rule-based morphologies (Zielinski and Simon, 2008) or parsers (Tapanainen and Järvinen, 1997) are to be combined with robust statistical part-of-speech taggers, whose coarse-grained tagsets cannot be trivially mapped onto the detailed annotations provided by deep, rule-based systems. Here, OLiA representations leverage tools or annotated corpora with different granularity.

As a proof of principle, Sukhareva and Chiarcos (2015) experimented with POS annotations from two English corpora, the Penn Treebank (PTB) and the Susanne corpus, whose annotations differ greatly in granularity and detail. We aimed to show that existing OLiA Annotation Models allow us to train tools on both corpora without loosing accuracy (as compared to a PTB-trained tagger) or granularity (as compared to a Susanne-trained tagger). But rather than providing an imprecise mapping from one tagset to another, we decomposed both annotations. As both corpora overlap in parts of the original Brown corpus, we observed a number of important differences, e.g., Susanne annotated adjectives in proper nouns as adjectives whereas PTB annotated these as proper nouns. With OLiA, such conceptual mismatches do not have to be resolved in favor of one or the other alternative: Tagsets require adjectives and nouns to be assigned distinct tags, but using ontologies, no implicit disjointness criterion applies (Chiarcos, 2008; Chiarcos and Erjavec, 2011).

By decomposing tagsets into informational atoms defined on grounds of one or multiple ontologies, the linking of OLiA Reference Model and OLiA Annotation Models yields a fully automated, *informationally lossless* conversion to a common representation formalism. This not only leverages granularity differences, but also incompatible definitions found even among apparently identical categories.

6.2. New experiments: Tagging with OLiA

As a proof of principle on how to train directly on ontological features rather than strings, Sukhareva and Chiarcos (2015) employed a feed-forward neural network with resilient backpropagation to this data, using pre-trained embeddings for the word under investigation, its predecessor and its successor as input features. Every node in the output layer represents an attribute-value assignment about the token under consideration (i.e., an RDF triple). The activation of this node is then interpreted as a confidence score for the respective statement, and during training, it is initialized with +1 for triples observed in the gold annotation, 0 for attribute-value combinations that were not distinguished in the gold annotation, and -1 for non-observed triples which were predictable from the gold annotation.

Sukhareva and Chiarcos (2015) employ a feed-forward neural network with resilient backpropagation:

1. input neurons initialized with the concatenated precompiled word embeddings of the word, its predecessor and successor;
2. one hidden layer with tanh activation and the number of neurons heuristically set to the average length of input and output layers, thus, a natural geometric (pyramidal) design;
3. a layer of output neurons that represent OLiA `MorphosyntacticCategory` triples; the tanh-normalized activation of these neurons represents the output vector.

The output of the neural network is decoded using different an OLiA-based *pruning* to extract ontologically consistent descriptions of maximum granularity and confidence. As OLiA does not provide disjointness axioms, different heuristics to infer consistency constraints for pruning among `MorphosyntacticCategory` concepts were tested, and remarkably, a heuristic structural pruning performed satisfactorily. In structural pruning, two concepts are consistent iff. one is a subclass of the other. Sukhareva and Chiarcos (2015) showed that the precision of predicted triples corresponds to the scores obtained for (the OLiA triple representation of) the output of conventional trained on the same data. Unlike conventional taggers, however, the neural network run over training data from both corpora, and also then, triple precision remains stable. This approach does thus not only guarantee ontologically consistent results, but it also is way more flexible than any string-based annotation and tools trained on that basis, whereas tags represent more or less opaque bundles of features.

Recently, we conducted experiments on the morphosyntactic annotation of Middle Low German (MLG), a historical low-resource language. Originating spoken in Northern Germany and the Netherlands, MLG evolved during the Middle Ages into a lingua franca around the Baltic Sea with a lasting impact on modern Scandinavian languages.

We worked with a 15th c. MLG Gospel of John, a relatively sparse training set with a total of only 19,000 tokens, unfortunately coming without annotations, but with parallel texts from other Germanic languages. We thus normalized MLG to modern German, Dutch and English, using an ensemble combination of one WBMT and two CBMT systems (Pettersson et al., 2014). The normalized text was then tagged with off-the-shelf part-of-speech taggers for German, Dutch and English, respectively, and their annotations were projected back to the MLG text.

These annotations could be represented in an interoperable way as sets of OLiA triples. We used the German normalization to assign pre-trained 100-dimensional German word embeddings[23] to MLG words and then used the experimental setup of Sukhareva and Chiarcos (2015) to train a neural classifier on this data, with the modern annotations as gold annotations. While the exact results of this experiment will be reported elsewhere, we would like to emphasize that this setting allowed us to assess the classification quality per for every type of OLiA triple (i.e., relation-object tuples, e.g., `[] rdf:type olia:Noun`) individually. By requiring a minimal f-score per tuple, we could filter out triples which were reliably predicted. This represents an interesting and instructive application of OLiA as we could now bootstrap a preliminary tagset for Middle Low German directly from the hierarchy of `MorphosyntacticCategory` concepts in the OLiA Reference Model:

```
AV   Adverb
CO   Conjunction
COc  CoordinatingConjunction
COs  SubordinatingConjunction
DT   Determiner (PronounOrDeterminer)
DTa  (definite) Article
DTd  DemonstrativeDeterminer
DTi  IndefiniteDeterminer
DTp  PossessiveDeterminer
NN   Noun
NNc  CommonNoun
NNp  ProperNoun
NU   Numeral (Quantifier)
NUc  CardinalNumber
PN   Pronoun (PronounOrDeterminer)
PN$  PossessivePronoun
PNp  PersReflPronoun
PNpp PersonalPronoun
PP   Preposition (Adposition)
PU   SentenceFinalPunctuation (Punctuation)
VE   Verb
VEf  FiniteVerb (i.e., finite main verb)
VEm  ModalVerb (AuxiliaryVerb)
VEn  NonFiniteVerb
VEnp Participle
O    other MorphosyntacticCategory
```

Even though this tag set is underspecified for OLiA concepts which could not reliably be recovered (e.g., `Adjective`), it is already more fine-grained than the tagset of the Universal Dependencies, and hence, more instructive than MLG annotations that could be achieved by state-of-the-art UD-based annotation projection (Agić et al., 2015). Our experiment thus represents early, but very promising steps towards bootstrapping both a tagset and annotation tools for Middle Low German – a language for which no annotations or tools whatsoever are available at the moment. Using an active learning approach, the underspecified tags in this tagset can be further refined with minimal manual effort. In this way, the combination of tool adaptation on parallel text, OLiA and machine learning provides a template for creating annotation schemes, tools and annotations for low-resource language varieties in general.

OLiA helps harmonizing different source annotations, it allows decomposing annotations into features that can be individually learned by a neural network, it is used to decode the activation of this neural network using structural constraints from the ontology, and it provides a structural template for the hierarchical organization of the tagset.

OLiA is a crucial element in this process of tagset bootstrapping, pointing towards the prospective role of OLiA in this line of research.

7. Summary

This paper described the Ontologies of Linguistic Annotation, their motivation and architecture, the current status of OLiA with respect to applications in documenting and retrieving language resources, modeling and accessing language resources, and developing annotation tools. Moreover, we elaborated important design principles of OLiA and sketched recent developments such as its extension to Universal Dependencies, a novel application in the context of a virtual library, and innovative strategies to develop ontology-based annotation tools which combine OLiA with neural networks.

Acknowledgments

We would like to thank two anonymous reviewers for input and feedback to this paper, as well as OLiA users, contributors and funders for feedback on and support for OLiA development. OLiA has originally been developed at the Collaborative Research Center (SFB) 441 “Linguistic Data Structures” (University of Tübingen, Germany) in the context of the project “Sustainability of Linguistic Resources” in cooperation with SFB 632 “Information Structure” (University of Potsdam, Humboldt-University Berlin, Germany) and SFB 538 “Multilingualism” (University of Hamburg, Germany) from 2006 to 2008. From 2007 to 2011, is has been maintained and further developed at SFB 632 in the context of the project “Linguistic Data Base”. In 2012, the first author continued his research on OLiA in the context of PostDoc fellowship at the Information Sciences Institute of the University of Southern California funded by the German Academic Exchange Service (DAAD). Since 2013, our work on OLiA has been partially supported by the LOEWE cluster “Digital Humanities” at the Goethe-University Frankfurt (2011-2014) and a DFG-funded project on a Virtual Library for General Linguistics conducted in cooperation with the University Library of the Goethe-University Frankfurt.

[23] `http://www.cis.upenn.edu/~ungar/eigenwords/`

References

Agić, Z., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 268–272, Beijing.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. 1st Workshop on Treebanks and Linguistic Theories (TLT-2002)*, Lisbon, Portugal.

Buyko, E., Chiarcos, C., and Pareja-Lora, A. (2008). Ontology-based interface specifications for a NLP pipeline architecture. In *Proc. 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Chiarcos, C. and Erjavec, T. (2011). OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proc. 5th Linguistic Annotation Workshop (LAW 2011)*, pages 11–20, Portland, Oregon.

Chiarcos, C. and Götze, M. (2007). A linguistic database with ontology-sensitive corpus querying. System demonstration at *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Frühjahrstagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV 2007)*, Tübingen, Germany.

Chiarcos, C., Dipper, S., and Götze, M. e. (2008). A flexible framework for integrating annotations from different tools and tag sets. *TAL (Traitement Automatique des Langues)*, 49(2):217–246.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Chiarcos, C. (2010a). Grounding an ontology of linguistic annotations in the Data Category Registry. In *Proc. Workshop on Language Resource and Language Technology Standards (LT&LTS 2010)*, pages 37–40, Valetta, Malta.

Chiarcos, C. (2010b). Towards robust multi-tool tagging. An OWL/DL-based approach. In *Proc. 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 659–670, Uppsala, Sweden.

Chiarcos, C. (2012a). Interoperability of Corpora and Annotations. In C. Chiarcos, et al., editors, *Linked Data in Linguistics*, pages 161–179, Heidelberg, Germany. Springer.

Chiarcos, C. (2012b). POWLA: Modeling linguistic corpora in OWL/DL. In *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*, pages 225–239, Heraklion, Greece. LNCS 7295, Springer, Heidelberg.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

De Marneffe, M. and Manning, C. D. (2008). The Stanford typed dependencies representation. In *Proc. COLING-2008 Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK.

Eckle-Kohler, J., McCrae, J., and Chiarcos, C. (2015). *lemonUby* - A large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal*, 6(4):371–378.

Farrar, S. and Langendoen, D. T. (2010). An OWL-DL implementation of GOLD: An ontology for the Semantic Web. In A. Witt et al., editors, *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Springer, Dordrecht, Netherlands.

Hahm, Y., Lim, K., Park, J., Yoon, Y., and Choi, K.-S. (2012). Korean NLP2RDF resources. In *Proc. 10th Workshop on Asian Language Resources (ALR 2012)*, pages 1–10, Mumbai, India.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference (ISWC 2013)*, pages 98–113, Sydney, Australia. LNCS 8219, Springer, Heidelberg.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. (2009). ISOcat: Remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with *lemon*. In *Proc. 8th Extended Semantic Web Conference (ESWC 2011)*, pages 245–259, Heraklion, Greece.

McCrae, J., Chiarcos, C., et al. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data cloud. In *Proc. 10th Language Resources and Evaluation Conference (LREC 2016)*, Portorož, Slovenia.

Pareja-Lora, A. and Aguado de Cea, G. (2010). Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In *Proc. 7th Language Resources and Evaluation Conference (LREC 2010)*, Valetta, Malta.

Pastor, J. A., Martinez, F. J., and Rodriguez, J. V. (2009). Advantages of thesaurus representation using Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, 14(4):Paper 422.

Petrova, S., Solf, M., Ritz, J., Chiarcos, C., and Zeldes, A. (2009). Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. *TAL (Traitement automatique des langues et langues anciennes)*, 2(50):47–71.

Pettersson, E., Megyesi, B., and Nivre, J. (2014). A multilingual evaluation of three spelling normalisation methods for historical text. In *Proc. 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014)*, pages 32–41, Gothenburg, Sweden.

Rehm, G., Eckart, R., Chiarcos, C., and Dellert, J. (2008). Ontology-based XQuery'ing of XML-encoded language resources on multiple annotation layers. In *Proc. 6th Language Resources and Evaluation Conference (LREC 2008)*, pages 525–532, Marrakech, Morocco.

Rubiera, E., Polo, L., Berrueta, D., and El Ghali, A. (2012).

TELIX: An RDF-based model for linguistic annotation. In *Proc. 9th Extended Semantic Web Conference (ESWC 2012)*, Heraklion, Greece.

Sampson, G. (2002). Briefly noted - English for the computer: The SUSANNE corpus and analytic scheme. *Computational Linguistics*, 28(1):102–103.

Saulwick, A., Windhouwer, M., Dimitriadis, A., and Goedemans, R. (2005). Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proc. 17th Conference on Advanced Information Systems Engineering (CAiSE 2005)*, Porto, Portugal.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universities of Stuttgart and Tübingen, Germany.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In *Proc. E-MELD workshop on Digital Language Documentation*, East Lansing, Michigan, USA.

Siemoneit, B., McCrae, J. P., and Cimiano, P. (2015). Linking four heterogeneous language resources as linked data. In *Proc. 4th Workshop on Linked Data in Linguistics: Resources and Applications (LDL 2015)*, pages 59–63, Beijing, China.

Skut, W., Brants, T., Krenn, B., and Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. In *Proc. ESSLLI-1998 Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proc. ACL-2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.

Sukhareva, M. and Chiarcos, C. (2015). An ontology-based approach to automatic part-of-speech tagging using heterogeneously annotated corpora. In *Proc. 2nd Workshop on Natural Language Processing and Linked Open Data (NLP&LOD 2015)*, Hissar, Bulgaria.

Tapanainen, P. and Järvinen, T. (1997). A nonprojective dependency parser. In *Proc. 5th Conference on Applied Natural Language Processing (ANLP 1997)*, pages 64–71, Washington, DC, USA.

Taylor, A. (2007). The York-Toronto-Helsinki parsed corpus of Old English prose. In J. C. Beal, et al., editors, *Creating and digitizing language corpora; Vol. 2, Diachronic Databases*, pages 196–227. Palgrave Macmillan, London, UK.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2003). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Verspoor, K., Kim, J.-D., and Dumontier, M. (2015). Interoperability of text corpus annotations with the semantic web. *BioMed Central Proceedings*, 9(Suppl 5).

Zielinski, A. and Simon, C. (2008). Morphisto: An open-source morphological analyzer for German. In *Proc. 11th International Conference on Finite State Methods in NLP (FSMNLP 2008)*, Ispra, Italy.

The Representation of an Old English Emotion Lexicon as Linked Open Data

Fahad Khan, Javier E. Díaz-Vera, Monica Monachini
University of Ca' Foscari Venezia, University of Castilla-La Mancha, ILC-CNR
Italy, Spain, Italy
fahad.khan@unive.it, javierenrique.diaz@uclm.es, monica.monachini@ilc.cnr.it

Abstract

We present the ongoing conversion of a lexicon of emotion terms in Old English (OE) into RDF using an extension of *lemon* called *lemonDIA* and which we briefly describe. We focus on the translation of the subset of the lexicon dealing with terms for shame and guilt and give a number of illustrative examples.

Keywords: Linguistic Linked Open Data, Old English, Lexicon

1. Introduction

The success of Linked Open Data(LOD) as a means of publishing wide coverage lexical resources, such as Wordnet (McCrae et al., 2014) and Wiktionary (Sérasset, 2015), and in making those datasets more accessible and inter-operable through the use of shared vocabularies and models, has encouraged the publication of more specialised language resources – resources which bring with them a whole host of new and diverse challenges. In particular, the core emphasis of LOD on the linking together of individual datasets makes it an inherently attractive paradigm for publishing lexical resources pertaining to fields such as historical linguistics or classical philology, since these fields also tend to bring together data from various different sources in interesting and complex ways.

However the number of such 'historical/philological' datasets in the Linguistic Linked Open Data cloud is small at present[1]. Arguably, this is in large part due to the lack of such data being publicly available and in easily proccessable formats – and also to the difficulty of translating such complicated data to linked data using currently available models, standards and best practices. However important work has been done in this area by (De Melo, 2014) and (Moran and Bruemmer, 2013) among others.

It can also be argued[2] that the relative paucity of such datasets is due to a lack of awareness within the philological/DH communities of the potential of LLOD for publishing resources in the field. This makes the development of appropriate vocabularies and models, targeted towards the kinds of datasets and resources that arise out of these disciplines (along with example datasets published using these vocabularies/models) extremely important as a means of making LLOD more accessible to these communities.

In this article we aim to contribute to ongoing discussions by describing one such vocbaulary, namely *lemonDIA*, which was developed in order to enable the publication in RDF of diachronic lexico-semantic datasets, along with an example of a philological dataset – an Old English (OE) lexicon of emotions derived from a corpus of OE texts – and its (ongoing) conversion into linked open data using *lemonDIA*.

[1]Among the explicitly historical/philological datasets that are present on the LLOD cloud there are: the historical gazateer Pleiades(`http://pleiades.stoa.org/places/1043/turtle`); the Semantic Quran(`http://datahub.io/dataset/semanticquran`); and the Linked Old Germanic Dictionaries(`http://linghub.lider-project.eu/datahub/germlex`).

[2]The following line of argumentation was suggested to us by one of our anonymous reviewers.

2. The *Sceamu* Dataset

The lexical dataset that we are working with was collected by the second author as part of a comprehensive study into the use and distribution of emotion terms in a representative corpus of OE texts (E Díaz-Vera, 2014). The study in question took both underlying historical and socio-linguistic factors along with relevant results from cognitive linguistics into consideration when determining the reasons for the use of certain terms in the labeling and description of emotions. This cross-disciplinary approach is reflected in the make-up and structure of the data which contains extensive information on the semantic shifts undergone by expressions belonging to the same lexical fields. As an initial test-case we decided to focus on that part of the dataset which was dedicated to lexical items for describing shame/embarrassment and guilt. We felt that this subset would by itself make a valuable addition to the LLOD cloud. In what follows we will refer to this dataset as the *Sceamu* dataset after the OE word which "occupies hypernomic position within the lexical domain of shame" (E Díaz-Vera, 2014).

The *Sceamu* dataset contains a total of 122 lexical entries (77 for embarrassment/shame and 45 for guilt). Each **lexical entry** contains the following: a specification of **part of speech**; information about **the lexical root** of the term and its **etymology** (this usually relates back to some form in Proto-IndoEuropean or Proto-Germanic); a list of **lemmas** (defined as the spelling and inflectional variants of the same entry) each with its own individual frequency in the corpus; **the corpus frequency** of the individual entry (before semantic disambiguation), and **the total frequency** as given in the Dictionary of Old English (Cameron et al., 2007); a number of different descriptions of the meaning – this may includes references to the arrangement of the terms in the lexicon into synsets; and finally, when relevant, **a categorisation of the semantic shift and a specification of the so-called theme**.

The following taxonomy is used to categorise the semantics of the words in the lexicon as well as specifying the different types of semantic shifts between expressions.

Literalness	Classification	Conceptualizations
+	literal	the emotion is an emotional experience
	metonymic	the emotion is a cause of the emotional experience
		the emotion is a response to the emotional experience
	synesthetic	the emotion is a sensorial experience
	metaphoric	the emotion is a living entity
		the emotion is a substance
		the emotion is an object
		the emotion is a force
-		the emotion is a place

Table 1: A Classification of the Semantics of Expressions.

For example, the entry for the verb *areodian*, which means 'to turn red, blush' and 'to turn red, blush with shame', specifies the following:

- the Lexical Root (in OE): *reod*;
- the Etymology of the Word (language and root): Proto-IndoEuropean, **reudh-* 'red';
- a list of Lemmas for the Entry and their Individual Frequencies: *areodigen[3], areodode_vbd[2], areoddian[1]*;
- Corpus Frequency (before semantic disambiguation):**6**;
- Total Frequency: **6**;
- a specification of the OE synset to which the entry belongs: the *SHAME* synset;
- a specification of the type of metonymy involved in the semantic shift: *RESULTATIVE METONYMY*.

Specific information about the time periods involved has only been included for a small number of entries, and so we haven't yet included this in the translations. We are planning to include it later however.

3. *lemon* and *lemonDIA*

As things currently stand *lemon* is the most popular model for representing lexico-semantic resources as linked; indeed it has come to be considered as a de facto standard for the representation of lexical resources as linked data. The *lemon* model has the great advantage of giving a clear, formal definition of the interface between a lexicon, viewed as a repository of specifically linguistic data, and an ontology, viewed as a knowledge base that contains information about the extensions of words (Cimiano et al., 2013). In previous work (Khan et al., 2014) we developed an extension of *lemon* called *lemonDIA* with the intention of representing semantic change over time in lexical linked data resources. The main idea was to represent word senses as processes in time, so that the change in the meaning of a term would not need to be represented in the ontology but would appear instead in the sense interface between the lexicon and the ontology.

Datasets like the *Sceamu* lexicon, which describe the changes in the lexicon of a language, or of several languages over time, and which relate to work in fields such as historical linguistics or philology, usually contain a lot of detailed data on the use of words. This might include, for instance, the various different derived and inflected forms and spelling variants of individual lexical entries and derivation relations between lexical entries – in addition to more or less detailed specifications of context and genre. The fact that there is often uncertainty as to how a word was used or why – since we cannot rely on native speaker intuition in these cases – makes it even more important that we are as faithful as possible when representing such datasets using an RDF based model. On the other hand, it is also imperative that we adhere closely to LOD best practises and conventions, such as the re-use of already existing vocabularies and datasets, when carrying out these translations: in order to take full advantage of the benefits of publishing in LOD. In the light of these and other considerations we decided to update the original *lemonDIA* to make it more amenable to representing this kind of historico-philological data[3]. We will now describe the relevant parts of this new version of *lemonDIA*.

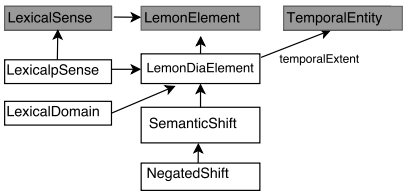


Figure 1: The Main *lemonDIA* classes.

In the diagram above we show some of the main classes in *lemonDIA* (classes from external vocabularies, in this case *lemon* and the OWL-time ontology[4], are given in grey).

- **`LemonDiaElement`** is a subclass of `LemonElement` that includes lexical elements with a history or in other words a temporal extent. For instance `LexicalpSense`, a subclass of `LexicalSense` and `LemonDiaElement`, is a `LexicalSense` viewed as a process in time namely the process or event of a given lexical entry, *l*'s, meaning a (ontological) concept *c*. We call these senses, *pSenses*, that is *perdurant senses*, to distinguish them from normal *lemon* senses.
- We have defined a new class **`LexicalDomain`** to cover those cases where we would like to describe the meaning of a lexical entry (and in particular this applies to roots, morphemes, reconstructed words) using some ontological category, but where the relationship between the meaning and the category doesn't

[3]Note that although *lemonDIA* is based on *lemon* rather than the newly finalised ontolex-lemon model, we do not foresee any difficulties in updating *lemonDIA* to meet the new specifications.

[4]`https://www.w3.org/TR/owl-time/`

quite meet the constraints placed upon *lemon* lexical senses (Cimiano et al., 2013). For instance in the following examples we will talk about the lexical domains of reconstructed roots or of expressions representing a cluster of related lexemes, as a means of describing the semantics of these types of lexical entries. We use the object property `lexicalDomain` to relate lexical entries and lexical domain objects, and `lexicalDomainConcept` to relate lexical domain objects with ontology concepts.

- The class **`SemanticShift`** tells the history of the polysemy of a word (more precisely of a word sense or a lexical domain). **`NegatedShift`** is used to cover cases where a privative affix is attached to a root lexeme thereby negating the meaning of the original lexeme, and where the succeeding compound undergoes a meaning shift[5]. We created two new object properties, `shiftSource` and `shiftTarget` to model the source and target of a semantic shift, and `shiftType` to specify the type of the shift (we give the details for `shiftSource` below):

```
:SemanticShift rdf:type owl:Class ;
        rdfs:subClassOf :LemonDiaElement .


:semanticShift rdf:type owl:ObjectProperty ;
        rdfs:domain owl:unionOf ( lemon:LexicalSense
                                        :LexicalDomain
                                  );
        rdfs:range :SemanticShift .

:shiftSource rdf:type owl:ObjectProperty ;
        rdfs:domain :semanticShift;
        rdfs:range [ rdf:type owl:Class ;
        owl:unionOf (
              lemon:LexicalSense
                    :LexicalDomain
                         )
            ] .
             .
```

- In addition we created a number of new subclasses of **`LexicalEntry`**; these include **`Expression`**, intended to cover word clusters encompassing lexical roots, morphological derivations and other variants; and **`Lemma`**, which we take to cover spelling and inflectional variants of the same lexeme. We can link an object of the class `Expression` to its `LexicalEntries` using the object property `:entlex`, and the other way round with `:lexent`.

The extra elements in *lemonDIA*, both the time-based elements and the extra lexicographically salient classes and properties, allow us to model historical datasets which deal with semantic variation over time. In the following examples we focus on these lexicographic elements, and the specification of semantic shift.

4. Modelling the *Sceamu* Dataset.

In this section we describe the modelling of the *Sceamu* dataset using the *lemonDIA* vocabulary.

[5]Note that in a previous version of *lemonDIA* the class `SemanticShift` was originally called `DiachronicShiftObject` .

As preliminaries we transformed the taxonomy in Table 1 into a SKOS taxonomy. We also created a number of classes and properties specific to the dataset amongst which are: `theme` describing the source domain of a semantic shift, as well as the data properties `corpusFreq` and `totalFreq`.

4.1. Examples

In this section we present two different examples of OE emotion expressions from the *Sceamu* dataset, *arleas* and *areodian*, and their translation into RDF. The conversion of these 122 entries has been carried out semi-automatically, using scripts written in python to carry out the majority of the conversion, the results of which were checked and corrected manually.

The OE expression *arleas* exemplifies a semantic shift of the type *causative metonymy* or more precisely, in reference to the categories in Table 1, 'the emotion is a cause of the emotional experience'. In this case, the emotion of shame is referred to by one its causes, namely, loss of honor: the root lexeme of *arleas*, the OE noun *ar(e)* refers to the concept of honour, and *-leas* is a privative suffix. We represent this in RDF in the following manner.

We define **`AR_LEAS`** to be an object of type `Expression`. `AR_LEAS` is linked to three different separate lexical entries **`:AR_LEAS_ADJ`**, **`:AR_LEASLICE_ADV`**, **`:AR_LEASNES_N`** using the `:entlex` property. Furthermore, we specify the lexical domain to which `AR_LEAS` belongs, its related root, and the language to which the expression belongs [6] as well as linking `AR_LEAS` to its suffix (using the property `privativeSuffix`):

```
:AR_LEAS a lemond:Expression ;
        lemond:entlex :AR_LEAS_ADJ,
        :AR_LEASLICE_ADV, :AR_LEASNES_N;
        lexinfo:root :AR_ROOT;
        lemond:privativeSuffix "-leas"@ang;
        lemond:lexicalDomain :arleas_domain;
        lemon:language "ang" .
```

The lexical domain **`arleas_domain`** points to the concept of `Shame` in dbpedia.

```
:arleas_shameful_domain a lemond:LexicalDomain ;
        lemond:lexicalDomainConcept dbpedia:Shame .
```

We give the entry for `AR_LEAS_ADJ` below. In the interests of space we have left out most of the associated lemmas (there are 15 in total).

```
:AR_LEAS_ADJ a lemon:LexicalEntry ;
        lemond:lexent :AR_LEAS ;
        wordnet:synset_member :OEASHAMED_ADJ ;
        lexinfo:partOfSpeech lexinfo:adjective ;
        lemond:pSense [lemon:reference dbpedia:Wickedness ;
                                :theme :dishonor ];
        lemond:lemma :arleas_adj_n, :arleas, :arleas_plus_a;
        :corpusFreq "517"^^xsd:nonNegativeInteger;
        :totalFreq "600"^^xsd:nonNegativeInteger ;
        lemon:language "ang" .
```

The lemmas associated with each lexical entry contain information about the distribution of the different variants of a lexeme. Currently each lemma instance only gives the form of the lemma and the frequency for that lemma in the

[6]We use the the ISO code for Old English, "ang", here.

corpus; we would like in future to be able to cite the texts in question using URN's[7]

Note the inclusion of data on corpus and total frequencies for each entry as well as information about the theme of the entry (where theme has the special sense given above). We have not yet added temporal information to any of the *lemonDia* elements, and so far we have only used the DBpedia ontology to provide reference for the pSenses in the dataset; both of these are only provisional. We are planning to enrich both aspects of the dataset in the near future, for example by using a more specialised ontology dealing with emotions, as well as the other relevant semantic aspects of the dataset. We have already defined the various time periods involved in the description of the dataset, e.g., the different stages in the evolution of OE.

Here is the lexical entry for the root *ar(e)* as well as its sense:

```
:AR_ROOT a lemon:LexicalEntry ;
        lemond:pSense :ar_sense;
        lemon:lexicalForm "ar(e)";
        lexinfo:etymologicalRoot :AIS ;
        lemon:language "ang" .
:ar_sense a lemond:LexicalpSense;
        lemond:senseGloss "honor, reverence, respect"@en;
        lemon:reference dbpedia:Honor .
```

The shift object `loss_res_to_shame` is a `NegatedShift`, of type metonymy (here `emotion_cause` refers to an item in the semantics taxonomy representing 'the emotion is a cause the emotional experience'), with a source `ar_sense` and target `arleas_shameful_domain`:

```
loss_res_to_shame a lemond:NegatedShift ;
        lemond:shiftSource :ar_sense;
        lemond:shiftTarget :arleas_shameful_domain ;
        lemond:shiftType :emotion_cause .
```

Our final example relates to the OE verb *areodian* (to redden) which represents a metonymic shift of the variety 'the emotion is a response to the emotional experience'. In this case the same lexical entry has two different senses, and the sense shift takes place between these two senses and is of type `emotion_response` (the node representing 'the emotion is a response to the emotional experience'). We also specify the fact that the shame sense of *areodian* is only found as a latin gloss in the corpus using the `textualDistribution` property.

```
:AREODIAN_VB a lemon:LexicalEntry ;
       lemon:pSense :areodian_sense_1, :areodian_sense_2 ;
        wordnet:synset_member :OESHAME_VB ;
        lexinfo:root :REOD ;
        lexinfo:partOfSpeech lexinfo:verb ;
        lemond:lemma :areodigen, :areodode_vbd, :areoddian ;
        :corpusFreq "6"^^xsd:nonNegativeInteger;
        :totalFreq "6"^^xsd:nonNegativeInteger ;
        lemon:language "ang" .
:REOD a lemon:LexicalEntry  ;
        lemon:pSense [lemon:reference dbpedia:Red];
        lemon:language "ang" .
:redden_to_shame a lemond:SemanticShift;
        lemond:shiftType :emotion_response ;
        lemond:shiftSource :areodian_sense_1 ;
        lemon:shiftTarget :areodian_sense_2 .
```

[7]URN stands for "Uniform Resource Name". Using URN's along with with canonical service technologies we can create permanent identifiers for texts, or parts of texts. See `http://cite-architecture.github.io/ctsurn/`

```
:areodian_sense_1 a lemond:LemonpSense ;
lemon:reference dbpedia:Blushing .
:areodian_sense_2 a lemond:LemonpSense ;
        lemon:reference dbpedia:Shame ;
        :textualDistribution :Gloss .
```

4.2. Conclusion

We have described the initial stages in the publication of an OE lexical resource using an extension of the `lemon` model called *lemonDIA*. The RDF version of the guilt/shame dataset is currently incomplete but we plan to publish it, in an enriched version, in the following couple of months.

5. Bibliographical References

Cameron, A., Amos, A. C., diPaolo Healey, A., et al. (2007). Dictionary of old english: A to g online. toronto: Doe project.

Cimiano, P., McCrae, J., Buitelaar, P., and Montiel-Ponsoda, E. (2013). On the role of senses in the ontology-lexicon. In *New Trends of Research in Ontologies and Lexical Resources*, pages 43–62. Springer.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. In *LREC*.

E Díaz-Vera, J. (2014). From cognitive linguistics to historical sociolinguistics: The evolution of old english expressions of shame and guilt. *Cognitive Linguistic Studies*, 1(1):55–83.

Khan, F., Boschetti, F., and Frontini, F. (2014). Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. Proceedings of the Workshop on Linked Data in Linguistics 2014 (LDL-2014).

McCrae, J., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Moran, S. and Bruemmer, M. (2013). Lemon-aid: using lemon to aid quantitative historical linguistic analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 28 – 33, Pisa, Italy, September. Association for Computational Linguistics.

Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires

Elena González-Blanco[1], Gimena del Río[2], Clara I. Martínez Cantón[1]
[1]Universidad Nacional de Educación a Distancia (UNED)
Bravo Murillo, 38 -- Madrid, Spain
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
C1033AAJ, Av Rivadavia 1917, Buenos Aires, Argentina
E-mail: egonzalezblanco@flog.uned.es, gdelrio@conicet.gov.ar, cimartinez@flog.uned.es

Abstract

This paper describes the creation of a poetic ontology in order to use it as a basis to link different databases and projects working on metrics and poetry. It has been built on the model of the Spanish digital repertoire ReMetCa, but its aim is to be enlarged and improved in order to fit under every poetic system. The conceptual semantic model, written in OWL, includes classes and metadata from standard ontological models related to humanities fields (such as CIDOC or Dublin Core), and adds specific elements and properties to describe poetic phenomena. Its final objective is to interconnect, reuse and locate data disseminated through poetic repertoires, in order to boost interoperability among them.

Keywords: ontology, TEI, verse, metrics, poetry, CIDOC

1. Introduction

Poetic features have been analysed and classified in the different literary traditions since the beginnings of Literary Theory studies. These features have been organized in form of poetic repertoires (first printed in paper and later web-based) that give account of metrical and rhythmical schemes of each poetical tradition or school. They gather long corpora of poems, which are defined by their main characteristics.

Performing comparative analysis of the existing digital poetic repertoires and databases poses important problems, as data sources are a rich and heterogeneous mosaic of virtual poetry collections integrated by multilingual corpora, such as French lyrical collections (Nouveau Naetebus), Italian (BedT), Hungarian (RPHA), Medieval Latin (Corpus Rhythmorum Musicum, Annalecta Hymnica Digitalia, Pedecerto), Gallego-portuguese (Oxford Cantigas de Santa María, MedDB2), Castilian (ReMetCa), Dutch (Dutch Song Database), Occitan (BedT, Poèsie Neotroubadouresque, The last song of the Troubadours), Catalan (Repertori d'obres en vers), Skaldic (Skaldic Project), or German (Lyrik des Minnesänger), among others.

Each repertoire belongs to its own poetical tradition and each tradition has developed its own analytical terminology for years in a different and independent way (González-Blanco & Sélaf, 2014). The result of this uncoordinated evolution is a bunch of varied terminologies to explain analogous metrical phenomena through the different poetic systems whose correspondences have been hardly studied.

From the philological point of view, there is no uniform academic approach to analyse, classify or study the different poetic manifestations, and the divergence of theories is even bigger when comparing poetry schools from different languages and periods. To give an example, the same quatrain of dodecasyllables can be encoded in different ways depending on the philological tradition: 12A12A12A12A, 4x(7pp+7p) or 4aaaa; or even named with different meaning: "alexandrine" means 14-syllable line in Spanish but only 12-syllables in French.

There are also important technical issues, as these repertoires were created in different periods and most of them are driven by stand-alone collected databases. The ER (Entity-Relationship) data model is the most commonly used for this purpose, together with the data model based on records for the logical implementation (Elmasri & Navathe, 2011), which is widely accepted, but the technological implementation varies from one project to another. So, there are repertoires that use SQL databases, others that are based on XML tagging or even new models based on non-structured databases.

Although the current ICT infrastructures are prepared to harvest such collections and provide access to them by a search engine, it is absolutely necessary to standardize metadata and vocabularies at philological level in order to be able to climb up the semantic layer and link data between different traditions. There are a few studies which deal obliquely with some of the above mentioned aspects (Bootz & Szoniecky, 2008; Zöllner-Weber, 2009), but there is not yet a conceptual model of ontology referred to metrics and poetry.

The closest related works to this topic are probably the conceptual model of CIDOC[1], the vocabularies of the Getty Museum[2], as they are designed to express relations and artistic manifestations in the field of humanities, the controlled vocabularies of English Broadside Ballad Project[3] and the linked data relations offered by the Library of Congress[4], which do not offer a deep information on metrics vocabulary.

2. Our Proposal

The aim of this paper is to present a model able to serve as a uniform solution for terminological issues in order to

[1] http://www.cidoc-crm.org
[2] http://www.getty.edu/research/tools/vocabularies
[3] http://ebba.english.ucsb.edu/
[4] http://id.loc.gov/

build a solid semantic structure as a basis to link the different poetic systems. This structure will enable to publish repertoires on the web in a structured format and using open standards in order to build an open-source and collaborative platform based on a poetic ontology which lets interoperability among the different European metrical repertoires. Performing comparative studies would allow researchers to move a step forward beyond the current philological state-of-the art, explaining phenomena like the origins of vernacular poetry or the evolution from accentual to syllabic rhythmical patterns.

2.1 The Basis for our Proposal

The data model proposed in this paper is based on the conceptual model designed for the Spanish Digital Repertoire of Medieval Poetry, ReMetCa[5]. ReMetCa has tested different systems (commercial, free, open-source, and proprietary). The final decision, after experimenting with Oracle Express Edition (González-Blanco & Rodríguez, 2013), has been using a relational database MySQL combined with a XML tagging using the TEI-verse module. The advantage of this hybrid model is that its relational structure provides both a uniform description of the formal characteristics of each poem, and flexibility and richness to show the complex metrics features of our texts thanks to TEI tagging.

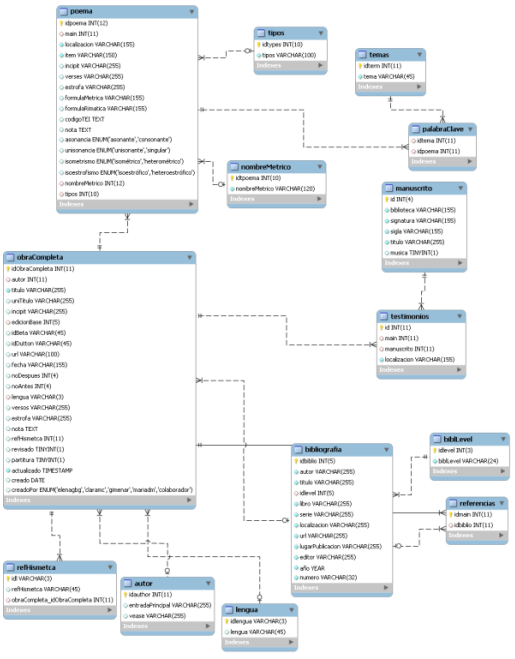


Figure 1: ReMetCa data model

2.2 The Data Model

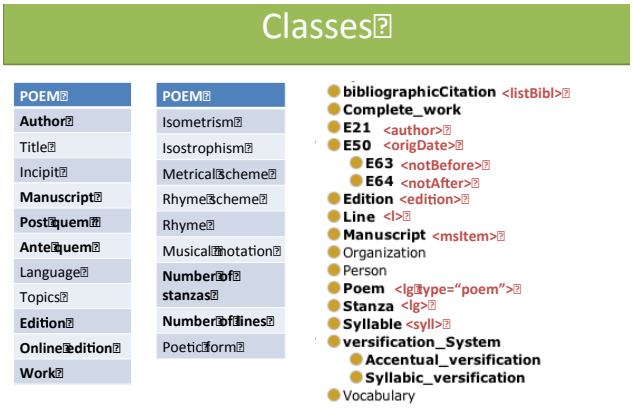
The conceptual model, designed on the basis of ReMetCa, has been transferred to the Semantic Web as Linked Open Data. The abstraction of this initial model is prepared to be amplified with the necessary fields and terms to define metrical phenomena which are not shown in the Spanish poetic system or in the other repertoires which have been taken into account to design this first version of the semantic prototype. In order to enlarge its horizons, structure, description and contents, datasets of various corpora have also been taken into account.

The implementation of the model makes use of one of the most recognized standards for the Semantic Web description: the Ontology Web Language (OWL), developed by W3C as an extension of RDFS[6]. The ontology integrates sets of predefined metadata using namespaces and it has been built using WebProtégé (Tudorache et al. 2011).

The ontology has been built based on the common categories or metadata of the existing repertoires. Some of them have been modeled as classes (poems, stanzas or lines), as they may contain individuals. Other fields reflect, however, the relationships that can be established between individuals, such as "composed by" which link poem and author. Others have been modeled as data properties, since they link entities to literals and values, line number, musical notation, or metrical scheme. Therefore, the current ontology does not collect all the fields of our database and tags we use but just the ones that it shares with other databases in order to provide interoperability between them.

The resulting first version of this ontology is hosted at various places[7].

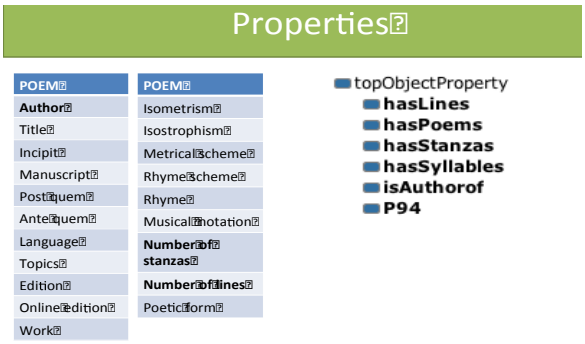
Figure 2 below shows a sample of the classes, properties and data properties of the poetry ontology related to the previous ReMetCa data model (Figure 1):



[5] www.remetca.uned.es

[6] https://www.w3.org/TR/owl-features/

[7] www.purl.org/net/ReMetCa;
http://datahub.io/dataset/ReMetCa-ontology;
http://lov.okfn.org/dataset/lov/vocabs/ReMetCa.



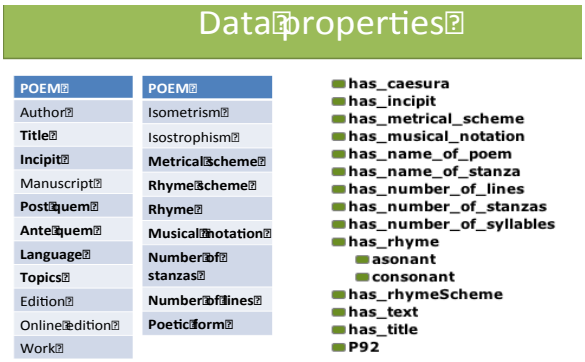


Figure 2: Sample of classes, properties and data properties related to the ReMetCa data model

This model is based on previous works that combine linked data and TEI. There are some preliminary approaches in the field of Philology developed by Christian-Emil Ore and Øyvind Eide (2007). Although the authors focus on the use of Topic Maps, they also point at the creation of a Conceptual Referential Model (CRM) model based on the TEI document and filled with all instances of mapped elements, having in mind that although TEI provides a richer vocabulary than EAD (Encoded Archival Description) or DCMI (Dublin Core Metadata Initiative), it is less abstract than RDF or METS (Metadata Encoding and Transmission Standard). Taking all these reflections into consideration, our ontology includes some elements of CIDOC into its classes and properties. For example, the entity "author" can be linked with DC: creator, FOAF: agent and CIDOC E21, as this is shown among other mappings in Figure 3 below.

As this is a relatively small ontology with interoperability purposes, there are not many elements shared with other ontologies. We have 13 entities, 5 object properties and 14 data properties or attributes. Almost all of them have a TEI equivalent or origin (especially the entities), but only a few are linked with other ontologies or vocabularies, such as CIDOC or DC. Only in these cases, our ReMetCa URIs have been substituted by their existing URIs.

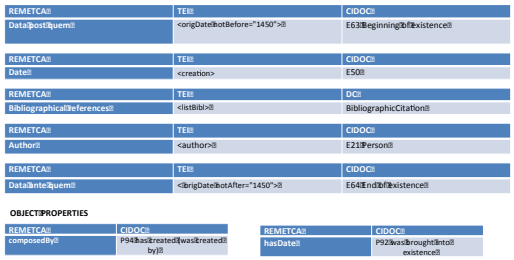


Figure 3: Mapping among ReMetCa, TEI and CIDOC models

Another issue is the number of attributes used for issues like "topics", "names of the poem types" or "functionality". All these categories are defined by the content of attributes like <poem type="">, or @subtype. The solution has been including TemaTres terminological software both to work as a lexical/content provider for TEI tags and to organize metadata. A general controlled vocabulary on Medieval Castilian Poetry at CAICYT-CONICET´s Semantic Server in order to create a more consistent categorial prototype has been set at http://vocabularios.caicyt.gov.ar/pmc/ which is one of the most useful applications of Linked Data in combination with TEI of this proposal, as it complements the XML structure with enriched content semantically organized and structured.

3. Conclusion

To sum up, this project of a poetic and metrical ontology intends to be much more than a repository of datasets, thesauri or controlled vocabularies. It aims to create a semantic standardized structure to describe, analyze and develop logical operations through the different poetic digital repertories and their related resources. Its final objective is to interconnect, reuse and locate the data disseminated through poetic databases in order to get interoperability among projects, to perform complex searches and to make the different resources "talk" to each other following a unique but adaptive model.

4. Acknowledgements

This paper has been developed thanks to the research projects funded by MINECO and led by Elena González-Blanco: Acción Europa Investiga EUIN2013-50630: Repertorio Digital de Poesía Europea (DIREPO) and FFI2014-57961-R. Laboratorio de Innovación en Humanidades Digitales: Edición Digital, Datos Enlazados y Entorno Virtual de Investigación para el trabajo en humanidades, and the Starting Grant ERC-2015-STG-679528 POSTDATA.

5. Bibliographical References

Bootz, P. & S. Szoniecky, (2008). "Towards an ontology of the field of digital poetry", paper presented at Electronic Literature in Europe, 2008. Full text available at http://elmcip.net/node/415

Burnard, L. & S. Bauman, eds., "TEI P5: Guidelines for Electronic Text Encoding and Interchange. Ver. 2.5.0." http://www.tei-c.org/release/doc/tei-p5-doc/en/html/. Accessed 30-10-2015.

Ciula, A., P. Spence & J. M. Vieira. (2008). "Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project", Literary and Linguist Computing 2008, 23 (3): 311-325.

Eide, Ø. & C.-E. Ore. (2007). "From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration between Text Collections and Other Sources of Cultural Historical Documentation", paper presented at the DH conference 2007. Abstract available at: http://www.edd.uio.no/artiklar/tekstkoding/poster_156_eide.html

Elmarsi, R. & S. B. Navathe. (2011). Fundamentos de Sistemas de Bases de Datos, Madrid, Pearson, Addison Wesley, 2011.

González-Blanco, E. & J. L. Rodríguez. (2013). "ReMetCa: a TEI based digital repertory on Medieval Spanish poetry", at The Linked TEI: Text Encoding in the Web, Book of Abstracts - electronic edition. Abstracts of the TEI Conference and Members Meeting 2013: October 2-5, Rome edited by Fabio Ciotti & Arianna Ciula, DIGILAB Sapienza University & TEI Consortium, Rome 2013, 178-185. http://digilab2.let.uniroma1.it/teiconf2013/abstracts/. Accessed 30-10-2013.

González-Blanco, E. & L. Seláf. (2014). "Megarep: A comprehensive research tool in Medieval and Renaissance poetic and metrical repertoires", Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world, edited by L. Soriano, M. Coderch, H. Rovira, G. Sabaté & X. Espluga., Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien, Peter Lang, 2014.

Tudorache, T., C. I. Nyulas, N.F. Noy & M.A. Musen. (2011). "WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web", Semantic Web Journal, IOS Press, 2011. http://www.semantic-web-journal.net/content/webprot%C3%A9g%C3%A9-distributed-ontology-editor-and-knowledge-acquisition-tool-web. Accessed: 30/10/2015.

Zöllner-Weber, A. (2009). "Ontologies and Logic Reasoning as Tools in Humanities?", DHQ 2009, 3: 4. http://www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html Accessed: 30/10/2015.

Linked Open Lexical Resources for the Greek Language

Sotiris Karampatakis, Sofia Karampataki, Charalampos Bratsas, Ioannis Antoniou
School of Mathematics
Aristotle University of Thessaloniki, Greece
Open Knowledge Greece
sokaramp@physics.auth.gr, s.karampataki@gmail.com, cbratsas@math.auth.gr, iantonio@math.auth.gr

Abstract

The continuous rise of information technology has led to a remarkable quantity of linguistic data that are accessible on the Web. Linguistic resources are more useful when linked but their distribution on the Web in various or closed formats makes it difficult to interlink with one another. So, they often end up restricted in data ”silos”. To overcome this problem Linked Data offers a way of publishing data using Web technologies in order to make feasible the connection of data coming from different sources. This paper presents how Greek WordNet was converted and published using a Resource Description Framework (RDF) model in order to expose as Linked Open Data.

Keywords: WordNet, RDF, Linked Data, Linguistic Data, NLP

1. Introduction

Language is the main means of communication among human beings and the most complicated one. It is estimated that there are more than 7.000 languages[1] in the world each one consisted by a large number of words, different grammar rules and syntax structure. All these constitute an enormous amount of resources that is the object of the scientific study of human language, or as it is called linguistics. Only a small part of the world languages data is available on the Web but they are either published in various formats either limited accessible(Chiarcos and Hellmann, 2011). Most of this data is contained in books which are not even digitalized yet. So the problem is how data from different sources can be retrieved and combined(McCrae et al., 2012). Linked Data was created to solve this problem by using the principles of the Web. Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets(Bizer et al., 2009). In order to succeed this, Linked Data use the RDF data model for publishing the data(Lassila and Swick, 1999). RDF is a method for expressing information in the form of subject-predicate-object expressions, known as triples .The subject (a resource) is related with its object(another resource) while the predicate expresses the relationship between them. Every resource and every relationship is denoted by an IRI which helps them be unambiguous in the web. This enables the interface between resources from different locations. Moreover linguistics linked data are referenced as an integral part of the Semantic Web(De Melo and Weikum, 2008). In this paper, we discuss how the principles of Linked Data can be applied to the publication of linguistic data. We present in detail the conversion of the Greek WordNet into Linked Data resources using the RDF model and give a brief dataset description.

2. WordNet

WordNet was created in the Princeton University and it is a lexical database for the English language (Fellbaum, 1998). Nouns, verbs, adjectives and adverbs are clustered into sets of synonyms called synsets, each representing a discrete concept. Semantic and lexical relations link the synonym sets and the result is a network of meaningfully related terms and concepts. This structure is a useful tool for computational linguistics and natural language processing (NLP)(Miller and Fellbaum, 2007). WordNet aims to produce a combination of dictionary and thesaurus and to support automatic text analysis and artificial intelligence applications(Morato et al., 2004). There have been created WordNets for many languages. Most of these were produced through projects like EuroWordNet (EWN)[2] (Vossen, 2002) and Balkanet[3]. EuroWordNet developed WordNets for several European languages and linked them together and Balkanet extended the number of languages that the EWN contains now counting up to six balkan languages. The Global WordNet project attempts to coordinate the production and linking of WordNets for all available languages(Fellbaum and Vossen, 2007).

2.1. BalkaNet - Greek WordNet

BalkaNet is an EC funded project (IST-2000-29388) that started in September 2001 and finished in August 2004(Stamou et al., 2004) . It is a multilingual lexical database that contains six monolingual WordNets for Balkan languages, specifically Bulgarian, Greek, Romanian, Serbian, Turkish and Czech. BalkaNet’s ambition is to correlate semantically words in each Balkan language and link them together in order to create an on line multilingual semantic network. Balkanet is constantly enriched. So far each monolingual WordNet consists of about 15K synsets. Greek WordNet is the resource out of which our work emerged and developed as part of BalkaNet project in the University of Patras from a team

[1] http://www.iana.org/assignments/language-subtag-registry

[2] http://www.illc.uva.nl/EuroWordNet/

[3] http://www.dblab.upatras.gr/balkanet/

of linguists with the attendance of University of Athens. The status of the Greek WordNet, as derived from the Balkanet project, is illustrated in Table 1. Specifically, the total number of synsets, literals and their ratio are given. Moreover, the total number of language internal relations between Greek WordNet, as well as the average ratio of links per synset. Finally, the table illustrates the number of non-lexicalized concepts, and the total number of glosses encoded for Greek WordNet synsets. Table 2 provides the number of synsets for each BCS subset along with the POS distribution of all Greek WordNet synsets. The numbers of semantic relations in the Greek WordNet, for each type of relation is given in Table 3.

Metric	Value
Total Number of Synsets	18,461
Literals	24366
Ratio Literals/Synsets	1.33
Lexico-semantic Relations	24,368
Ratio Relations/Synsets	1.32
Glosses	18,461

Table 1: Statistical data of the Greek WordNet

Part of Speech	Count
Nouns	14426
Verbs	3402
Adjectives	617
Adverbs	16

Table 2: POS Distribution for Greek WordNet

Semantic Relation	Count
HYPERNYM	18324
HOLO_MEMBER	1320
HOLO_PART	2660
HOLO_SUBSTANCE	57
HOLO_PORTION	162
VERB_GROUP	424
BE_IN_STATE	143
SUBEVENT	132
CAUSES	76
ALSO_SEE	210
SIMILAR_TO	46
DERIVED	103
NEAR_ANTONYM	689
ANTONYM	22
Total	24368

Table 3: Semantic Relations in the Greek WordNet

3. Converting WordNet to RDF

WordNet databases provide an XML dump which, in the Greek WordNet case, contained 18.461 synsets. Every synset is described by the following tags:

- **SYNSET**: contains all the data relative to Synset.
- **ID**: identifier of the ILI(InterLingual Index). The prefix ENG20 means that it had been created by the Princeton WordNet, version 2.0, while the prefix BILI means that the synset is a BalkaNet specific one.
- **POS**: part of speech. The possible values are: *n* (noun), *v* (verb), *b* (adverb) and *a* (adjective).
- **SYNONYM**: list of the literals of this synset. At least one literal is mandatory.
- **LITERAL**: wording of the literal.
- **SENSE**: number used for the sense differentiation.
- **LNOTE**: note about this literal.
- **Def**: gloss of the synset. This wording allows to describe the synset. It's not mandatory.
- **STAMP**: gives some additional information about this synset : author, date, etc.
- **USAGE**: gives an example of use of the synset.
- **BCS**: number of the base concept associated with this synset. The possible values are 1, 2 or 3.
- **ILR**: InterLingual Relation. Gives a relation between this synset and the specified ILI.
- **TYPE**: type of this relation. The possible values are : *be_in_state*, *category_domain*, *causes*, *derived*, *eng_derivative*, *holo_member*, *holo_part*, *holo_portion*, *hypernym*, *near_antonym*, *particle*, *region_domain*, *similar_to*, *subevent*, *usage_domain*, *verb_group*.

An example sysnet can be seen in Listing 1

```
<SYNSET>
  <ID>ENG20-08833936-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL> θάλασσα
      <SENSE>1</SENSE>
      <LNOTE> θάλασσα</LNOTE>
    </LITERAL>
  </SYNONYM>
  <ILR>ENG20-08651117-n
    <TYPE>hypernym</TYPE>
  </ILR>
  <ILR>ENG20-08726856-n
    <TYPE>holo_part</TYPE>
  </ILR>
  <DEF> η υδάτινη αλμυρή έκταση που καλύπτει το μεγαλύτερο μέρος της επιφάνειας της γης</DEF><BCS>2</BCS></SYNSET>
```

Listing 1: Sample xml structure for the synset "θάλασσα-noun-1"

3.1. WordNet Ontology

The English WordNet of Princeton 2.0 was converted in RDF using the RDF/OWL ontology(Van Assem et al., 2006a). The BalkaNet Project has a very similar data structure to the Princeton WordNet, since most of the semantic relations represented within BalkaNet have been carried from EWN and PWN. Therefore, we used the same ontology, as described by (van Assem et al., 2006b). The

conversion schema has three main classes: Synset, Word and WordSense. Synset and WordSense have subclasses depending on the part of speech. Synset subclasses are NounSynset, VerbSynset, AdjectiveSynset in turn subclass AdjectiveSatelliteSynset) and AdverbSynset. WordSense subclasses are NounWordSense, VerbWordSense, AdjectiveWordSense and AdverbWordSense. The Word class does not have subclasses such as VerbWord. It has only the subclass Collocation used to represent words that have hyphens or underscores in them. We extended the WordNet Full Ontology by adding some extra properties in order to include some fields that were not on Princeton WordNet and were introduced by the Balkanet Project, such as LNOTE which contains the pronunciation[4]. We also made use of the properties rdfs:label and rdf:type in all entities.

3.2. Ontology Mappings

Each synset's element from the XML file is mapped to a property, based on the element name. The mappings that were used are illustrated in Table 4. Some additional mappings were done, such as mapping the first synonym's literal data of each synset to the property rdfs:label, in order to provide a label for the synset, rdf:type to link the entities to their classes, owl:sameAs for linking to the Princeton WordNet 2.0 etc. Moreover, Wordsenses and Words were mapped to the appropriate classes, and words to the wn20s:lexicalForm property.

3.3. RDFizer

In order to convert the xml wornet file to RDF we developed an open-source *WordNet RDFizer* in the C++ programming language. [5] The RDFizer:

1. reads the WordNet XML file,
2. assigns an IRI to each entity,
3. produces the entities of WordSense and Word,
4. makes connections between the synsets and other resources (both interlink and intralink),
5. and generates a N3 formatted RDF file.

The IRI name pattern uses English WordNet model. For example, if the entity was a Synset IRI would be in the form[6]
`wordnet-gr:synset-{first_literal_of_synset}-{pos}-{sense}.`
The IRI for WordSenses: was assigned respectively:
`wordnet-gr:wordsense-{first_literal_of_wordsense}-{pos}-{sense}.`
as well as for Words:
`wordnet-gr:word-{lexical_form_of_word}.`
The RDFizer process follows the next three steps:

- In the first step the rdfizer reads each synset separately. Initially reads the first elements of synset such as ID, DEF, USAGE, BCS etc. and makes the corresponding triples, taking into account the mappings that have been defined in rdfizer's settings. In addition, the rdfizer converts the synsetID provided by the data to the corresponding ID at the form of WN2.0. For instance, in BalkaNet, the synset that belongs to BalkaNet has ID form : ENG20-00208807-v. The same synset in WordNet 2.0 has ID form: 200208807, where number 1,2,3 or 4 is added depending on whether the word is noun, verb, adjective or adverb respectively. Then, the rdfizer reads each synonym included in synset and constructs the corresponding entities WordSense and Word. Finally, reads the element ILR containing the child element type and makes two tables: the first table has three columns where the first contains the synsetID of the parent synset, the second contains the synsetID which includes the ILR and the third one contains the relation between these two synsets as it is described by the element type of ILR. The second table acts in fact as an index where in the first column there is the ID parent synset and in the second one the attributed IRI. The first step ends after every element of the synset has been read and then the process is repeated for all the synsets of the xml file.
- After the reading of the XML file ends, the rdfizer constructs the ILR connections. The procedure is done as follows: the rdfizer reads each line of the table separately, maps each synsetID with the suitable IRI and according on the ILR connection type finds the suitable property and constructs the triplet.

ELEMENT	MAPPED PROPERTY	CLASS
ID	wn20s:synsetId	Synset
DEF	wn20s:gloss	Synset
SENSE	wn20s:sense	WordSense
LITERAL	rdfs:label	Synset, WordSense, Word
LITERAL	wn20s:lexicalForm	Word
LNOTE	wngre-onto:lnote	WordSense, Word
SYNONYM	wn20s:containsWordSense	Synset
BCS	wngre-onto:BCS	Synset
n	wn20s:NounSynset	Synset
v	wn20s:VerbSynset	Synset
b	wn20s:AdverbSynset	Synset
a	wn20s:AdjectiveSynset	Synset
n	wn20s:NounWordSense	WordSense
v	wn20s:VerbWordSense	WordSense
b	wn20s:AdverbWordSense	WordSense
a	wn20s:AdjectiveWordSense	WordSense
category_domain	wn20s:classifiedByTopic	Synset
causes	wn20s:causes	Synset
derived	wn20s:derivationallyRelated	Synset
holo_member	wn20s:memberHolonymOf	Synset
holo_part	wn20s:partHolonymOf	Synset
hypernym	wn20s:hypernymOf	Synset
near_antonym	wngre-onto:nearAntonymOf	Synset
antonym	wn20s:antonymOf	Synset
region_domain	wn20s:classifiedByRegion	Synset
similar_to	wn20s:similarTo	Synset
usage_domain	wn20s:classifiedByUsage	Synset
verb_group	wn20s:sameVerbGroupAs	Synset
holo_substance	wn20s:substanceHolonymOf	Synset
also_see	wn20s:seeAlso	Synset
be_in_state	wngre-onto:beInState	Synset
eng_derivative	wngre-onto:engDerivativeOf	Synset
holo_portion	wngre-onto:holoPortionOf	Synset
particle	wngre-onto:particleOf	Synset
subevent	wngre-onto:subevent	Synset

Table 4: Mappings for the WordNet conversion to RDF

[4] `http://wordnet.okfn.gr/ontology/wngre-onto.rdfs`
[5] `https://github.com/okfngr/wordnet`
[6] wordnet-gr: `http://wordnet.okfn.gr/resource/`

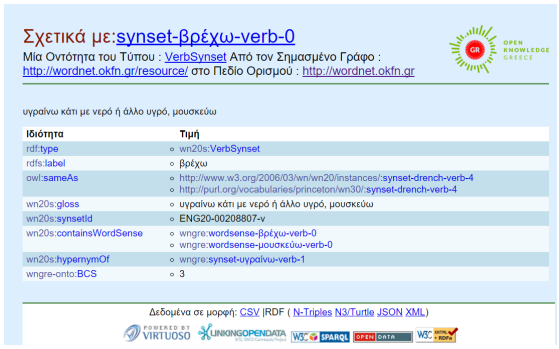


Figure 1: Example Linked Data content negotiation for a Greek-WordNet resource

- In the third step we constructed owl:sameAs type links (inter-links) of the synsets included in the Greek WordNet with the corresponding synsets of English WordNet version 2.0. The mappings between the synsetID of Greek WordNet and the synsetID of English WordNet were done one by one so there is no doubt about their correctness, exploiting the synsetID similarity as described at the first step.

With completion of this procedure we made connections using Silk Framework(Volz et al., 2009) to the according synsets of version 3.0 of PWN. We utilized the WordNet 2.0 to Wordnet 3.0 mappings provided[7]. For every synset containing a WN2.0 link we created the according link to WN3.0 based on the mentioned mappings.

4. Dataset description

The converted Greek-WordNet dataset resulted to a total of 351.913 triples, involving 298.284 properties, 69.655 intra-links and 53.629 sameAs links towards the English WordNet versions 2 and 3. There are totally 24 different predicates, 16 of them indicating intra-links. The dataset is accesible as an RDF dump[8], through a SPARQL endpoint[9] and as a Linked Data interface[10] providing also a simple free-text search application. The Linked Data interface is designed to provide proper content negotiation(Sauermann et al., 2011) and has a vital part on IRI dereferencing. An example HTML display for a resource can be seen in Figure 1. The dataset is provided under the *Creative-Commons Attribution-ShareAlike* (CC-by-SA)[11] lisence. The dataset is already part of the LOD cloud since 2013(Chiarcos et al., 2014) [12].

5. Conclusion

In this paper, we dealt with the modelling procedure which was developed in support of exposing Greek WordNet as Linked Data. Specifically we described the method developed for the conversion to RDF, based on the four principles of Linked Data. We focused on the description of an rdfizer which was developed for the conversion process of Greek WordNet. Our proposed converting method can easily be applied on every WordNet. As these conversion processes are freely available we think that they will foster users of language resources and NLP research in general. It is the first Linguistic Dataset of the Greek Language in a Linked Data format, as a demonstrator for more datasets to come. Future work could add linking into various datasets, NLP applications explotiting the enchanced information contained and tools for the expansion of the Greek and other WordNets utilizing the Linked Open Data cloud knowledge. Recently, the Greek Edition of DBpedia, el.DBpedia(Kontokostas et al., 2012), created over 6.7k links to this dataset enriching both datasets with even more useful information. The dataset has already been included in Open Multilingual Wordnet(Bond and Foster, 2013). We believe that the demonstration of capabilities using this dataset will increase the need to revive and expand the Greek WordNet and the rest of WordNets of the Balkan area. Furthermore, as Linked Data become more popular among Web processors new tools will be available to facilitate the work of NLP community.

[7] https://github.com/jrvosse/wordnet-3.0-rdf/tree/master/rdf/wn20mappings

[8] http://wordnet.okfn.gr/downloads

[9] http://wordnet.okfn.gr/sparql

[10] http://wordnet.okfn.gr

[11] http://creativecommons.org/licenses/by-sa/3.0/

[12] https://datahub.io/dataset/greek-wordnet

6. Bibliographical References

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362.

Chiarcos, C. and Hellmann, S. (2011). Working group for open data in linguistics: Status quo and perspectives. In *CEUR Workshop Proceedings*, volume 739.

Chiarcos, C., Mccrae, J., Osenova, P., and Vertan, C. (2014). Linked Data in Linguistics 2014 . Introduction and Overview. *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages vii–xv.

De Melo, G. and Weikum, G. (2008). Language as a foundation of the semantic web. In *International Semantic Web Conference (Posters & Demos)*.

Fellbaum, C. and Vossen, P. (2007). Connecting the universal to the specific: Towards the global grid. In *Intercultural Collaboration*, volume 4568, pages 1–16.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database.

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., and Metakides, G. (2012). Internationalization of Linked Data: The case of the Greek DBpedia edition. *Journal of Web Semantics*, 15:51–61.

Lassila, O. and Swick, R. R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. *World Wide Web Internet And Web Information Systems*, 2004(October):1–54.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

Miller, G. A. and Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, 41(2):209–214.

Morato, J., Marzal, M. Á., Lloréns, J., and Moreiro, J. (2004). WordNet Applications. In *Proceedings of the 2nd Global Wordnet Conference,*, pages 270–278.

Sauermann, L., Cyganiak, R., and Völkel, M. (2011). Cool uris for the semantic web.

Stamou, S., Nenadic, G., and Christodoulakis, D. (2004). Exploring balkanet shared ontology for multilingual conceptual indexing. In *LREC*.

Van Assem, M., Gangemi, A., and Schreiber, G. (2006a). Conversion of wordnet to a standard rdf/owl representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC?06), Genoa, Italy*, pages 237–242.

van Assem, M., Gangemi, A., and Schreiber, G. (2006b). Rdf/owl representation of wordnet. *W3C Public Working Draft of*, 19.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk–a link discovery framework for the web of data. In *the 2nd Linked Data on the Web*, volume Madrid, Sp, pages —-.

Vossen, P. (2002). WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée*, 7(1):27–38.