LREC 2018 Workshop

6th Workshop on Linked Data in Linguistic (LDL-2018)

PROCEEDINGS

Edited by

John P. McCrae, Christian Chiarcos, Thierry Declerck,
Jorge Gracia, Bettina Klimek

ISBN: 979-10-95546-19-1
EAN: 9791095546191

12 May 2018

Proceedings of the LREC 2018 Workshop
"6th Workshop on Linked Data in Linguistic (LDL-2018)"

12 May 2018 – Miyazaki, Japan

Edited by John P. McCrae, Christian Chiarcos, Thierry Declerck, Jorge Gracia, Bettina Klimek

http://ldl2018.linguistic-lod.org/

Acknowledgements: This work was supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015.

Organising Committee

- John P. McCrae, National University of Ireland, Galway, Ireland
- Christian Chiarcos, Goethe University Frankfurt, Germany
- Thierry Declerck, DFKI GmbH, Germany and ACDH-ÖAW, Austria
- Jorge Gracia, University of Zaragoza, Spain
- Bettina Klimek, University of Leipzig, Germany

Programme Committee

- Eneko Agirre, University of the Basque Country, Spain
- Guadalupe Aguado-de-Cea, Universidad Politécnica de Madrid, Spain
- Núria Bel, Universitat Pompeu Fabra, Spain
- Claire Bonial, University of Colorado at Boulder, USA
- Paul Buitelaar, Insight Center for Data Analytics, National University of Ireland Galway, Ireland
- Nicoletta Calzolari, ILC-CNR, Italy
- Steve Cassidy, Macquarie University, Australia
- Damir Cavar, Indiana University, USA
- Philipp Cimiano, University of Bielefeld, Germany
- Gerard de Melo, Rutgers University, USA
- Francesca Frontini, Université Paul Valéry, Montpellier, France
- Jeff Good, University at Buffalo, USA
- Dagmar Gromann, Technical University Dresden, Germany
- Yoshihiko Hayashi, Waseda University, Japan
- Fahad Khan, ILC-CNR, Italy
- Dave Lewis, ADAPT, Ireland
- Vanessa Lopez, IBM Research, Ireland
- Elena Montiel-Ponsoda, Universidad Politécnica de Madrid, Spain
- Steve Moran, Universität Zürich, Switzerland
- Roberto Navigli, “La Sapienza” Università di Roma, Italy
- Sebastian Nordhoff, Language Science Press, Berlin, Germany
- Petya Osenova, IICT-BAS, Bulgaria
- Antonio Pareja-Lora, Universidad Complutense Madrid, Spain
- Francesca Quattri, Jiangsu University, China

- Mariano Rico, Universidad Politécnica de Madrid, Spain
- Felix Sasaki, Berlin, Germany
- Andrea Schalley, Karlstad University, Sweden
- Gilles Sérasset, University Grenoble Alpes, France
- Milena Slavcheva, JRC-Brussels, Belgium
- Armando Stellato, University of Rome, Tor Vergata, Italy
- Marieke van Erp, KNAW Humanities Cluster, the Netherlands
- 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 LDL workshop series is organized by the Open Linguistics Working Group of the Open Knowledge Foundation, and has contributed greatly to the emergence and growth of the Linguistic Linked Open Data (LLOD) cloud. LDL workshops contribute to the discussion, dissemination and establishment of community standards that drive this development, most notably the Lemon/OntoLex model for lexical resources, as well as standards for other types of language resources still under development.

Building on our earlier success in creating and linking language resources, LDL-2018 will focus on Linguistic Data Science, i.e., research methodologies and applications building on Linguistic Linked Open Data and the existing technology and resource stack for linguistics, natural language processing and digital humanities.

LDL-2018 builds on the success of the workshop series, incl. two appearances at LREC (2014, 2016), where we attracted a large number of interested participants. As of 2016, LDL workshops alternate with our stand-alone conference on Language, Data and Knowledge (LDK). LDK-2017 was held in Galway, Ireland, as a 3-day event with 150 registrants and several satellite workshops. Continuing the LDL workshop series together with LDK is important in order to facilitate dissemination within and to receive input from the language resource community, and LREC is the obvious host conference for this purpose. LDL-2018 will be supported by the ELEXIS project on an European Lexicographic Infrastructure.

J. McCrae, C. Chiarcos, T. Declerck, J. Gracia, B. Klimek May 2018

Programme

Opening Session

09.15 – 09.30 Introduction by Workshop Chairs

09.30 – 10.30 Francis Bond (Nanyang Technological University, Singapore)
Teaching through Tagging — Interactive Lexical Semantics (Invited Talk)

10.30 – 11.00 Coffee Break

First Session: Language and Time

11.00 – 11.25 Alia Bahanshal, Hend Al-Khalifa and AbdulMalik Al-Salman
Modeling Semantic Change as Linked Data using Distributional Semantics: A Case on the Arabic Language

11.25 – 11.50 Frances Gillis-Webber
Managing Provenance and Versioning for an (Evolving) Dictionary in Linked Data Format

11.50 – 12.15 Yalemisew Abgaz, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt and Andy Way
A Semantic Model for Traditional Data Collection Questionnaires enabling Cultural Analysis

12.15 – 12.35 Fahad Khan
Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web

12.35 – 13.45 Lunch Break

Second Session: Lexical Data and Annotation

13.45 – 14.10 Christian Chiarcos, Ilya Khait, Émilie Pagé-Perron, Niko Schenk, Lucas Reckling and Jayanth Jayanth
Annotating Sumerian: A LLOD-enhanced Workflow for Cuneiform Corpora

14.10 – 14.35 Ranka Stankovič, Cvetana Krstev, Biljana Lazić and Mihailo Škorić
Electronic Dictionaries - from File System to lemon Based Lexical Database

14.35 – 15.00 Sabine Tittel, Helena Bermúdez-Sabel and Christian Chiarcos
Using RDFa to Link Text and Dictionary Data for Medieval French

Third Session: Representation and Formalization Issues

15.00 – 15.20 Livy Real, Alexandre Rademaker, Fabricio Chalub and Valeria de Paiva
Towards Temporal Reasoning in Portuguese

15.20 – 15.45 Christian Chiarcos, Kathrin Donandt, Hasmik Sargsian, Jesse Wichers Schreur, Maxim Ionov
Towards LLOD-based Language Contact Studies. A Case Study in Interoperability

15.45 – 16.05 Thierry Declerck
Towards a Linked Lexical Data Cloud based on OntoLex-Lemon

16.05 – 16.35 Coffee Break

16.35 – 17.35 **Panel Session**: 6 years of Linguistic Linked Open Data:
Where do we stand and where do we (want to) go?

17:35 – 18:20 **Open Session**: Join the Cloud! Add your Data!

18:20 – 18:30 **Closing Session**

Table of Contents

Modeling Semantic Change as Linked Data using Distributional Semantics: A Case on the Arabic Language

Alia O. Bahanshal
King Saud University
King Abdulaziz City for Science and Technology
Riyadh, Saudi Arabia
abahanshal@kacst.edu.sa

Hend S. Al-Khalifa
King Saud University
Riyadh, Saudi Arabia
hendk@ksu.edu.sa

AbdulMalik Al-Salman
King Saud University
Riyadh, Saudi Arabia
salman@ksu.edu.s

Abstract

Semantic change focuses on the study of word usage evolution, where the new meaning of a word is somehow different from the original usage. This paper proposes a linked data model to represent semantic change identified by a distributional semantics approach applied on the Arabic language.

Keywords: Linked Data, Semantic Change, Distributional Semantics, Arabic Language

1. Introduction

Words acquire new meanings through time, and anyone who reads old texts can notice words which have different meanings today. The study of language change is essential for anyone who uses ancient literature such as religious scholars, librarians and linguists. By language change, we mean the semantic variations in language. Semantic change, also called semantic development and semantic shift, is part of the semantics (the study of meaning in a language), and it focuses on the study of word usage evolution, where the new meaning of a word is entirely different from the original usage. For example, the Arabic word "حريم" \hareem\ (women) used to mean in the old dictionary: everything forbidden, and their new meaning in the contemporary dictionary is: women. From this example, we can see how meanings changed over time. Another case in English is the word "dog" that is a specific breed of the dog, which has become later the entire race of dogs. This kind of change is called widening or extension. Semantic change occurs for political, social, economic and historical reasons or just to name things. The ability to identify semantic change would help linguists understand the evolution of words through time and recognize cultural phenomena. Furthermore, many applications could benefit from such studies, such as Natural Language Processing (NLP), automatic interpretation and translation process. For instance, to translate a recent publication, we cannot use the words meanings from an ancient dictionary since the words meanings change over time and may not reflect the current words meanings. Language change should be known to use the accurate meaning of words for the desired publication time. Furthermore, by recognizing language change, linguists would be able to identify the most used lexicons in literature and the ones diminished by observing the frequency of the words in different time periods of a corpus as in (Michel et al., 2011), and many more applications could be built.

Different approaches were used to identify words semantic change over time. One of these methods is statistical semantics, where statistics are used to determine words meanings, as stated in (Furnas et al., 1983; Weaver, 1955) that "Statistical patterns of human word usage can be used to figure out what people mean". The statistical semantics hypothesis subsumes the distributional hypothesis, which in linguistics is based on word context, as stated in (Harris, 1954) that "Words that occur in similar contexts tend to have similar meanings", and in (Firth, 1957) that "You shall know a word by the company it keeps". This hypothesis was the motivation for researchers to use distributional semantics further to measure diachronic change through time (Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Rodda et al., 2016).

Linked Data was introduced as a method to publish interlinked data forming a single global space of data from various sources (Web of Data). Its features such as openness, linking capabilities, and graph representations witnessed great attention in the field of linguistics and proved its capability in Natural Language Processing (NLP) and in representing lexical resources as open data.

The aim of this paper is to apply distributional semantics to the Arabic language to identify semantic change, where no one (to the best of our knowledge) explored this area before, to know if the methods applied to the English language would function as well. Our aim is also to propose algorithms that utilize existing methods for identifying semantic change based on Distributional Semantics Models (DSM) Vector Space Model (VSM) and Latent Semantic Analysis (LSA), where the current frameworks that used DSM relied only on visualization to identify semantic change and do not provide a broad approach to analyze and utilize the resulted visuals and information. Furthermore, we aim to propose a new model that represents the results of the distributional methods as Linked Data, and to solve existing models' lack of fundamental information needed for the semantic change identification process.

The remainder of this paper is structured as follows. Section 2 presents background information about the

Arabic language, semantic change, distributional semantics methods, Linked Data, and the *lemon* model which is used as the base for the new proposed model to represent semantic change as Linked Data. Section 3 describes related works. Section 4 introduces the dataset used in this research and the preprocessing steps. Section 5 explains the distributional semantics algorithms to identify the semantic change. Section 6 presents the proposed Arabic Semantic Change (ASC) model. Section 7 explains case studies where distributional semantics algorithms and the ASC model are applied. Finally, Section 8 provides conclusion and future works.

2. Background

In this section, background about the Arabic language, semantic change, distributional semantics, Linked Data, and the *lemon* model is presented.

2.1 Arabic Language

Arabic is the language of (Quran), Muslims religious book, and a Semitic language spoken by nearly 500 million people around the world and one of the official UN languages. Like any language, Arabic has its grammar, spelling, and pronunciation; yet it has its own characteristics which made it distinctive. Arabic is read and written from right to left (except numbers), its alphabet consist of 29 spoken letters, and 36 written characters. Classical Arabic descends Modern Standard Arabic (MSA), which is the language used in formal writing and speech, and Colloquial Arabic, which is the language spoken every day and what children speak as their first language. Arabic is written with an orthography that includes optional diacritical marks. Diacritics are extremely useful for readability and understanding, their absence in Arabic text adds another layer of lexical and morphological ambiguity. Diacritics in Arabic are optional orthographic symbols typically representing short vowels and aid the reader to disambiguate the writing or just articulate it correctly. The Quran is fully diacritized to minimize the chances of misinterpreting it. Children's educational texts, classical poetry tends to be diacritized as well. The reader should analyze the text morphologically, syntactically and semantically before reading it, i.e., restoring the diacritics. It is very rare to use diacritics in modern Arabic text. Newspapers, books, and the Internet have Arabic content that is usually written without diacritics. The Arabic language is our focus in this paper, where the semantic change is identified for it.

2.2 Semantic Change

Semantics is defined as "the study of meaning" (Lyons, 1977), "the study of meaning in language" (Hurford, 2007), "the study of meaning communicated through language" (Saeed, 1997), and "the part of linguistics that is concerned with meaning" (Löbner, 2002). Semantic development is a branch of Semantics which focus on change in the meaning of words to help researchers understand the words evolution through time (Issa and Issa, 2008). Semantic change is defined as "the gradual change in words semantics through time, where words change their meanings from one to another as a result of several life changes" and semantic development is equal to semantic change by many linguistics opinions and does not mean a rise in meaning (Qalalah, 2017). All languages in the world face semantic change from time to time and are forced by the law of change; some languages evolve faster than others in some specific time periods (Issa and Issa, 2008). The semantic change is the fastest branch of Semantics to evolve because it is bond with the human movements and life change (Abuhadeemah, 2008).

Semantic change could occur to name things. For instance, the word "انترنت" \e`ntarrnit\ does not exist in old literature but appeared as a translation of the English word Internet. The meaning of words change over time, e.g., the meaning of word "حرامي" \ḥarāmī\ in Arabic through time became the same meaning of the word "السارق" \alsāriq\, the person who made mistakes in general and became the person who steals things particularly. In the English language, the word 'mouse' that means the small animal witnessed the change of the addition of another meaning 'the computer device'. The semantic change identification using computational approach is the main goal of this paper.

2.3 Distributional Semantics

Distributional semantics is built upon distributional hypothesis, which is in linguistics is based on word context where words with similar meanings occur in a similar context (Firth, 1957; Harris, 1954). Thus, the meaning of a word is related to the distribution of words around it. The efforts to apply this hypothesis on semantics usually led to vectors and metrics, which was the motivation to investigate further vector space model (VSM) and its relationship with words meaning (Turney and Pantel, 2010). "The representation of a set of documents as vectors in a common vector space is known as the ***vector space model***" (Chowdhury, 2010)."The idea of the VSM is to represent each document in a collection as a point in space (a vector in a vector space). Points that are close together in this space are semantically similar, and points that are far apart are semantically distant" (Turney and Pantel, 2010).

The performance of information retrieval is improved when the number of vectors' components is limited. To reduce the dimensionality of vector models, Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is used, which maps documents and terms to a common conceptual space. The statistical technique used is called Singular Value Decomposition (SVD) (Golub and Reinsch, 1970). LSA creates a semantic space of a sizeable term-document matrix where terms and documents that are closely related are placed nearby each other. With SVD, the small and unrelated data are ignored, and a new space is arranged with the most associative data. VSM and LSA are further investigated in this paper to identify the semantic change in the Arabic language.

2.4 Linked Data

In 2006, Linked Data was introduced by Tim Berner-Lee as a method to publish interlinked data forming a single

global space of data from various sources (Web of Data) (Heath and Bizer, 2011). Linked Data format is understood by machines, and thus raw data can be retrieved. It is defined as: “best practices for connecting and publishing structured data on the Web” (Bizer et al., 2009). Linked Data, same as the web of documents, connects different online resources, but it interlinks both data and documents in a predefined standard format.

Resource Description Format (RDF) (Beckett, 2004) is the standard model for the expression of data and relations in Linked Data. RDF data model consists of (subject - predicate – object) triples. The subject and object could be URIs referencing to resources. The object could be a string literal, while the predicate represents the relation between a subject and an object. Linked Data was used in several domains such as Medical and Health, Education, Government, Linguistic domain and more. Data from different data sources were converted into RDF format and interlinked forming the Linking Open Data (LOD) cloud (Figure 1). The LOD cloud is a community effort founded in 2007, and the World Wide Web Consortium (W3C) assists with its production under the Linked Open Data project coordination (Bizer, 2009).

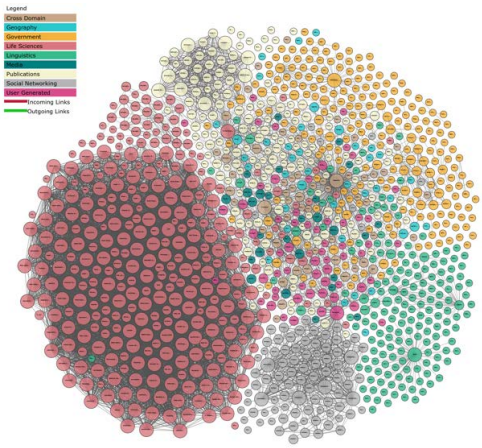

Figure 1 LOD Cloud Diagram (Abele et al., 2017)

Linked Data witnessed a considerable attention in the Linguistic domain and many tools and applications utilized Linguistic Linked Open Data (LLOD). Natural Language Processing (NLP), is one of the goals of creating a sub-cloud of datasets related to the Linguistic domain (Chiarcos et al., 2013).

In this paper, we propose an extension to LLOD that will represent the semantic change information using distributional semantics methods as Linked Data. We named the extended model "Arabic Semantic Change" (ASC).

The *lemon* model (McCrae et al., 2012a) was chosen as the base for our proposed ASC model. *lemon* model was introduced to describe vocabularies that are used to enrich ontologies vocabulary elements with information that are realized linguistically and in natural languages, and this is because ontology languages such as OWL [1] (The Web Ontology Language) and RDF lack the support of linguistic data[2]. *lemon* represents lexical entries morphological and syntactic properties and acts as a syntax-semantics interface. It was first developed within the European project “Monnet” and further developed by W3C Ontology-Lexica Community Group[2].

lemon follows “semantics by reference” principle, where the lexical meaning is stated entirely in the ontology, and the lexicon only points to the proper concept, unlike other lexical resources which include as part of the lexicon the semantic relations, such as synonymy and antonymy (McCrae et al., 2012a).

OntoLex is the new version of *lemon* core, and it was released in 2016 by W3C Ontology-Lexica Community Group[2]. Figure represents the core of *lemon* (OntoLex), which covers “the basic elements required to define lexical entries and associate them to their lexical forms as well as to concepts in the ontology representing their meaning” (McCrae et al., 2012a).

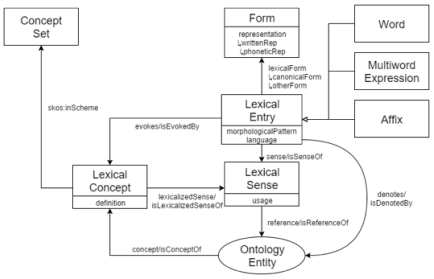

Figure 2 *lemon* OntoLex Core[2]

The main elements of the OntoLex are Lexical Entry, Forms, Semantics, Lexical Sense & Reference, and Lexical Concept. Lexical entry is the main class of the OntoLex core, and it is defined as: “A lexical entry represents a unit of analysis of the lexicon that consists of a set of forms that are grammatically related and a set of base meanings that are associated with all of these forms. Thus, a lexical entry is a word, multiword expression or affix with a single part-of-speech, morphological pattern, etymology and set of senses”[2]. There are different forms for each lexical entry from the grammatical point of view, and it is defined as: “A form represents one grammatical realization of a lexical entry”[2].

The Semantics in the model represents the meaning of a lexical entry using `denotes` property by pointing to the ontological concept following the *semantics by reference* principle. Lexical senses were introduced because the property `denotes` was not sufficient for all the linking cases of the lexical entry with the ontology. `LexicalSense` is defined as: “A lexical sense represents a reification of a pair of a uniquely determined lexical entry

[1] https://www.w3.org/OWL/

[2] https://www.w3.org/2016/05/ontolex/

and a uniquely determined ontology entity it refers to. A link between a lexical entry and an ontology entity via a Lexical Sense object implies that the lexical entry can be used to refer to the ontology entity in question."[2]. The lexical concept is defined as: "A lexical concept represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses"[2].

3. Related Work

In this section, some works related to distributional semantics methods to identify semantic change and previous models used to represent that change as Linked Data are presented.

For the distributional semantics methods, Jatowt and Duh (2014) proposed a framework for identifying semantic change at three levels: lexical, contrastive-pair and sentiment orientation levels. The framework is based on distributional semantics, where the meaning of a word is identified from contexts in which it occurs in texts. In their approach, they viewed each time period's context words as a vector and calculated the similarity between vectors. If the vectors are dissimilar, it means a semantic change has occurred. To evaluate their approach, they used visualization only and listed the top context words. In our work, we proposed an algorithm that further utilizes the top context words to identify the different meanings in the various time periods and according to these meanings the semantic change was detected. Also, we employed the method on the Arabic language to identify the semantic change.

In the area of representing semantic change as Linked Data, van Aggelen et al. (2016) proposed a model for describing semantic change and the connection with WordNet (Fellbaum, 1998) *lemon* model (McCrae et al., 2012b). The problem with this model is that it focused only on representing the similarity scores between decades and did not include any information about the meanings or the context words in different decades. Figure 3 shows the

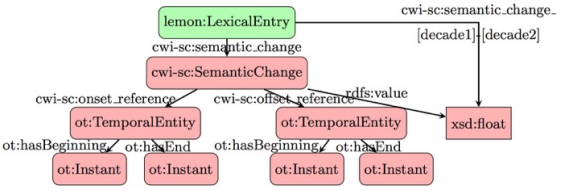

Figure 2 A model for connecting WordNet entries to cross-decade scores of lexical changes. prefix ot stands for OWL-Time and cwi-sc for the purpose-built vocabulary (van Aggelen et al., 2016)

structure of that model and the connection with *lemon*'s Lexical Entry, the old version of OntoLex[2].

The model has two time periods and their similarity scores, however, no information about the meanings are displayed. Thus, in our proposed model, this problem will be solved

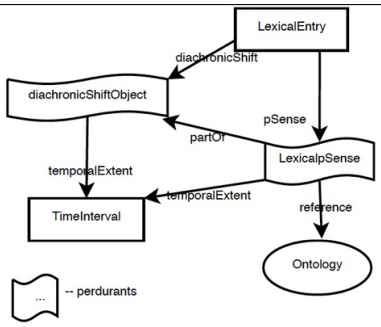

Figure 4 The lemonDia Model (Khan et al., 2014)

by including the top context words retrieved from the corpus, also we incorporated the different meanings obtained from the results of the proposed DSM algorithm.

Also, Khan et al. (2014) proposed an extension to *lemon* model, the lemonDia model to represent the diachronic change between word meanings. The problem with this model is that it focused on representing known words that changed their meanings through time and did not include information about the distributional semantics methods information such as similarity scores or context words. Figure 4 shows the structure of lemonDia model, where no information about the word's distributional semantics is contained. Therefore, in our proposed model this issue will be solved by including the similarly scores for each period along with the top context words.

4. Dataset

For the words that changed their meanings through time, N-gram data need to be collected from a corpus. This activity is required to know the most occurred terms around the words in the study and based on the distributional semantics hypothesis, the meaning of a word is known by the words around it (its context). N-gram is a sequence of n terms usually collected from text or collection of data, e.g., corpus. In this paper, the n size is five, and the 5-gram is the context words surrounding the target word, two words before and two words after, which was used by Wijaya and Yeniterzi (2011) and showed good results for semantic change identification. To collect the 5-gram, we have used KACST corpus (Al-Thubaity, 2015) a large and diverse Arabic corpus with a free access. It was developed to be used for several purpose applications, from linguistics research to developing NLP applications. The corpus size started at seven hundred million words and currently has one billion words[3] in Arabic language (Classical Arabic and MSA). KACST corpus is divided by time periods (0-600 to 2011-2020). However, the number of data in the earlier years was small in size, yet, the time period (1700-1800) is where the corpus started to have a sufficient number of data. Therefore, we chose the periods (1700-1800 to 2011-2017) in our study to ensure that a suitable amount of datasets is collected from the corpus.

After collecting the frequencies of the 5-gram words for all periods of each study word, the results were recorded in Excel sheets. Our dataset is constructed from every period,

[3] http://corpus.kacst.edu.sa

and its 5-gram unique words along with the words frequencies count. After constructing the dataset, we noticed that the collected data contained many symbols such as (“), and stop words such as the preposition word في \fī\ (in), which may affect our analysis results. Therefore, we developed a Java program to clean the data from unwanted contents (stop words, punctuations, and symbol). The program used an Arabic stop words list[4] that originally contained 750 words. This list was expanded to include 1659 items including Arabic stop words and symbols. These items were retrieved after manually cleaning the dataset and it is available for download[5].
Also, Google Books corpora[6] were used to extract 5-gram dataset for the English words to apply the proposed algorithm and model on them, and to test if they can be generalized to all languages. The Google Books corpora have 155 billion words, and the extracted 5-gram words should appear at least 40 times in the corpus, the dataset is divided by decades from the year (1810) to year (2000).

5. Using Distributional Semantics to Identify Semantic Change

Vector Space Model (VSM) is a widely used approach in Information Retrieval (IR) and NLP, and in our case, it is used to identify the semantic change. Figure 5 shows the steps needed to construct VSM to identify the semantic change.

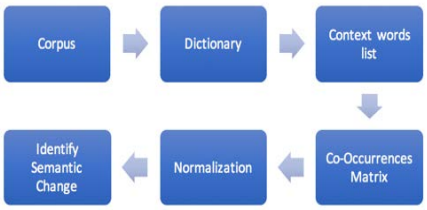

Figure 5 The steps needed to construct the VSM to identify semantic change

In Figure 5, after collecting the 5-gram (context words) from KACST corpus for all periods, they are listed as an Excel sheet. Next, we developed Arabic Semantic Change tool (Figure 6) using Java to construct a co-occurrences $M \times N$ matrix from the dataset. The matrix rows are the M context words in all positions around the target words. The matrix columns are the N time periods of the diachronic KACST corpus. In our case, the time is partitioned from the period 1700-1800 to 2011-2017.
At this point, we can view each time period in the matrix as a vector, which depends on the number of occurrences or the frequency of a context word in that period. After constructing the vectors for each time period, where the vector of a word w in a time period i is represented as $t_i[w]$, the tool calculates the weight of each vector’s context word. The weight is the count of a context word w in t_i (or the term frequency (Tf) of w in t_i) divided by the total number of context words in that time period (or the

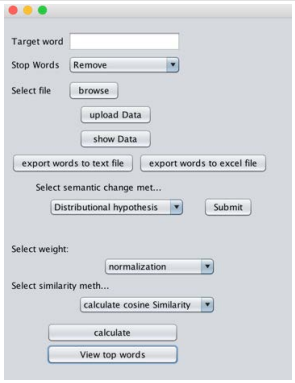

Figure 6 Arabic Semantic Change Tool

normalized length of the vector). Euclidean length is used to compute the vectors’ lengths.
Afterward, the semantic change is measured by calculating the similarity between the same word vector in time period i and time period j. We denote that by $\mathrm{sim}(t_i[w], t_j[w])$. Semantic change is likely to occur if $[\mathrm{sim}(t_i[w], t_j[w]) \rightarrow 0]$. To measure the semantic change, we followed (Jatowt and Duh, 2014) approach to compute the similarity of word w vector in last time period vector similarity with all previous time periods vectors. Similarity is calculated using well-known similarity measures Cosine similarity (Huang, 2008). Next, to identify the semantic change of a word, we plot the similarity values and observe the plotted line. If the line has a steep increase or fall, it means semantic change likely occurred, and if it is almost a straight line, it indicates the word has a stable meaning through time. This result is converted into a numerical value, called the (*VSM Semantic Change Plotting Result*) and it ranges from 0 to 1 according to the similarity curve. Additionally, the top context words are retrieved to know if the meanings were changed over time. To extract the top context words, the approach by (Jatowt and Duh, 2014) was used. From each time period, the context words with less than 1% frequency are removed. Then, the frequency of each of the remaining context words a in time period t_i are then compared to the frequency of the same word in time period t_{i-1}.

$$S(a, t_i) = {f(a, t_i)}/{f(a, t_{i-1})}$$

To identify semantic change, the meaning of a word should be known, and the retrieved context word can be used to determine that meaning. This assumption was made after presenting the collected list of top context words to three different linguists to identify the semantic change. They gave different opinions about the rising and the reduction of the word meaning but they all agreed on that meaning could be comprehended from the list of context words in each time period. Thus, an algorithm (Algorithm 1) that utilizes this agreement was proposed to identify the semantic change.
From Algorithm 1, the $vx1$ matrix is constructed, where

[4] https://github.com/mohataher/arabic-stop-words
[5] https://github.com/abahanshal/arabic-stop-words-list
[6] https://googlebooks.byu.edu

Algorithm 1: Identify semantic change using VSM

For $i = 1$ *to* k *, where* k *is the number of time periods* d *selected from corpus* c
Do
  Retrieve v *context* words from c for each d_i
  Construct k occurrences $v \times 1$ matrix, *where rows are* v *context words and column is time period* d_i
End
For $i = 1$ *to* k
Do
Remove all context words with less than 1% frequency from d_i occurrences matrix
Compute top context words T of d_i by dividing each word a *frequency* in d_i with same word a *frequency* in d_{i-1}
$$T(a, d_i) = f(d_i, a)/f(d_{i-1}, a)$$
Identify word meaning m_{d_i} from T
End
For i*=1 to* k
Do For $j = 1$ *to* k
  Do *For each meaning* m_{d_j} *in* d_j
    Do
      Compare meanings m_{d_i} of d_i with other time periods meanings m_{d_j} of d_j
      If Similarity $(m_{d_i}, m_{d_j}) > 0.5$
      Then
      w has same meaning in d_i and d_j and No semantic change occurs between d_i and d_j
      Else
      Semantic change occurs between d_i and d_j *and* w has new meaning in d_i

rows are v context words of word w and column is the time period d_i, and the values are the context words' frequencies in the corpus, the top context words are retrieved from corpus c.

During this step, the word meanings are obtained. Then, the senses are compared to identify the semantic change between the time periods. If new meanings appear through years, then a semantic change has occurred. The result of the semantic change from the proposed algorithm is converted to numerical value, called the (*VSM Semantic Change Algorithm Result*), and it ranges from 0 to 1 according to the meaning comparison results.

The overall value of semantic change is calculated by the averaging the two numerical values as in the following formula:

Semantic Change value= Average (VSM Semantic Change Plotting Result+ VSM Semantic Change Algorithm Result)

Afterward, the collected information from the VSM and the calculated semantic change value are represented as Linked Data using the proposed Arabic Semantic Change Model explained in the next section.

6. Arabic Semantic Change Model

The Arabic Semantic Change (ASC) model is proposed to represent the resulting data of the distributional semantics (VSM) method to identify semantic change as Linked Data. In the ASC model, it was focused to include all the information needed to identify the semantic change. We viewed the model as a vector representation of time periods that includes the distributional semantics information. Each period has a start and end year, a set of top context words, a set of different meanings and the similarity scores between the last period and the referenced period. All the information needed to apply our proposed semantic change identification method and algorithm are represented in ASC model as shown in Figure 7.

The model used existing vocabularies, and new ones were introduced, the new proposed vocabularies and properties are recognized by prefix asc which is the name space or URI used for the model.

In Figure 7, the ontolex:LexicalEntry is connected to blank node asc:SemanticCahngePeriod that represents the semantic change period which is connected to four other nodes. The first node is the time period which is modelled using OWL-Time ("Time Ontology in OWL," 2017), and it has a start and an end time to represent the start and end years of a period of the set of time periods extracted from a corpus (e.g., 2011-2017). In the other cases where the period could be a decade instead of an interval time, the start and end times will have the same value. The second node connected to asc:SemanticChangePeriod is asc:ContextWord that represents the extracted top words in each period, and it is connected to the external dataset (DBpedia ("DBpedia," 2018)) using lemon ontolex:reference property. From the set of the context words, the meanings of a word in each period could be identified. The third node is asc:Meaning that represents the identified meanings from context words in each period, and it is connected to the external dataset (DBpedia) using lemon ontolex:reference property. The VSM algorithm is applied to the meanings to identify the semantic change. The forth node has a value xsd:float an XML Schema Datatype (Carroll and Pan, 2006) with a float value that represents the similarity between the last period vector and the referenced time period vector. These similarity scores can be used to plot the similarity curve to identify the semantic change. Furthermore, in the model the word or the LexicalEntry is connected to an xsd:float that has a range of float values that represents the amount of semantic

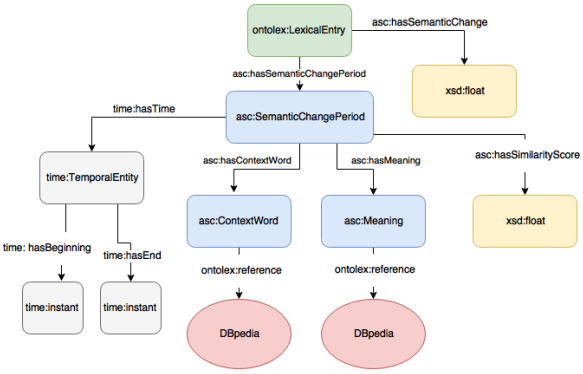

Figure 7 Arabic Semantic Change Model

Table 1: Arabic Semantic Change (ASC) Model

ASC Element	Definition
SemanticChangePeriod	A resource that represents the time period vector and is connected to four other resources
ContextWord	A resource that represents a word's context words retrieved from the corpus
Meaning	A resource that represents the word's different meanings in a Semantic Change Period.
hasSemanticChangePeriod	A property that relates a Lexical Entry with a Semantic Change Period
hasContextWord	A property that relates a Semantic Change Period with the word's Context Word.
hasMeaning	A property that relates a Semantic Change Period with the word's Meaning.
hasSimilarityScore	A property that relates a Semantic Change Period with the Similarity Score float value.
hasSemanticChange	A property that relates a Lexical Entry with the Semantic Change float value.

change occurred. If the value is closer to 1, then a semantic change has likely happened, and if closer to zero, then no semantic change exists, according to meanings and semantic change identification algorithm. Table 1 lists ASC model elements and their definitions.

7. Case Studies

In this section, the Arabic word بئر \be`r\ (well) and the English word (gay) will be used to identify the semantic change occurred to them using VSM algorithm, and to represent their semantic change identification information as Linked Data using Arabic Semantic Change (ASC) Model.

7.1 Arabic Word بئر \be`r\ (well)

The VSM algorithm was used to identify the semantic change of the word بئر \be`r\ (well). First, the 5-gram context words were collected from KACST corpus. Then, the top context words were retrieved, and the similarity scores are computed and plotted using the Arabic Semantic Change tool. Figure 8 shows the plotted similarity curve for the word بئر \be`r\ (well).

From Figure 8, the cosine similarity curve has steep

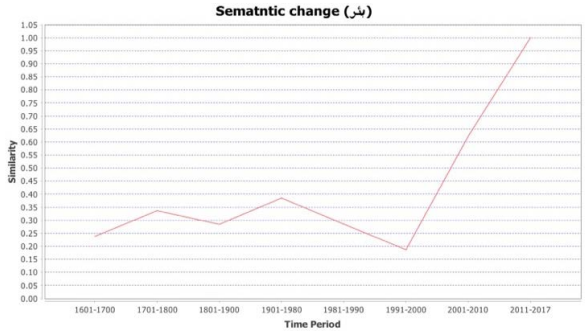

Figure 8 Similarity curve for word بئر \be`r\ (well).

increase and fall in all periods which indicates that the vectors are not similar, and the word has witnessed change through time. Thus, the VSM semantic change plotting result is equal to one. Additionally, the similarity value in the period (2011-2017) is equal to one because we are comparing the period vector with itself.

Table 2 Word بئر \be`r\ (well) top context words and obtained meanings in different time periods

	2011-2017	2001-2010	1991-2000	1981-1990	1901-1980	1801-1900	1701-1800
	غنم	معونة	بضاعة	معونة	معونة	بضاعة	بضاعة
	نفط	رومة	معونة	بضاعة	بضاعة	معونة	مدينة
	مياه	عميقة	رومة	رومة	رومة	رومة	معونة
	ماء	ماء	أريس	زمزم	حفر	زمزم	رومة
	عميقة	عبد	ماء		زمزم		أريس
	حسن	زمزم	جمل				زمزم
	عبد		زمزم				
	عاتر						
	نفطية						
	سلم						
	زمزم						
Identified Meaning	بئر نفط بئر ماء حفرة عميقة	بئر ماء حفرة عميقة	بئر ماء	بئر ماء	بئر ماء	بئر ماء	بئر ماء

Next, the top context words were retrieved using the Arabic Semantic Change tool, and the meanings of the word in different time periods were manually obtained from the list of top context words. Table 2 shows the retrieved top context words and the obtained meanings of the word بئر \be`r\ (well).

Then, by following Algorithm 1, the meanings in all periods were compared and the results were recorded.

Table 3 presents the algorithm comparison results. From observing the senses, the word بئر \be`r\ (well) had as stable meanings of بئر ماء \be`r māa`\ (water well) and حفرة\ḥofrah\ (a dig) from the period (1700-1800) to the period (2011-2017). However, a new meaning arises in the latest period (2011-2017) بئر نفط \be`r nift\ (petroleum well).

Table 3 Word بئر \be`r\ (well) VSM algorithm results

Meaning	2011-2017	2001-2010	1991-2000	1981-1990	1901-1980	1801-1900	1701-1800
بئر ماء \be`r māa`\ (water well)	1	1	1	1	1	1	1
بئر نفط \be`r nift\ (petroleum well)	1	0	0	0	0	0	0
حفرة\ḥofah\ (a dig)	1	1	1	1	1	1	1

This is an indication that the word has witnessed semantic change through time. Thus, the VSM semantic change algorithm result is equal to one. Also, the overall semantic change value is computed by the addition of the two results, the plotting and the algorithm results, and is equal to one.

Furthermore, the Arabic Semantic Change model was used to represent the semantic change identification information and the semantic change value as Linked Data. Figure 9 shows the word بئر \be`r\ (well) Linked Data representation for the time period (2011-2017).

In Figure 9, the similarity score is equal to 1.0 because we are comparing the similarity between the last time period vector with itself. Also, the semantic change calculated using VSM is equal to one. The context words and meanings are presented and linked to the external DBpedia dataset.

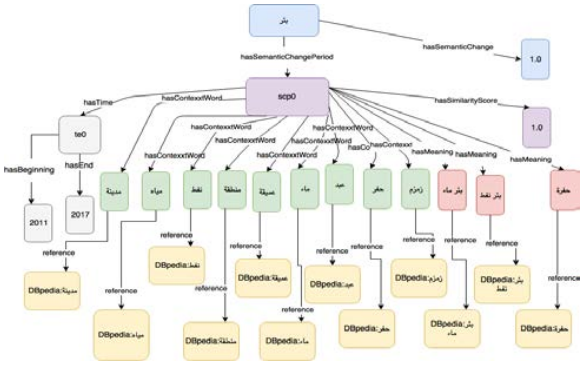

Figure 9 Word بئر \be`r\ (well) Linked Data representation using ASC model

7.2 English Word (gay)

The VSM semantic change method and algorithm were applied to the English word (gay) to identify the semantic change. First, the 5-gram were collected from Google Books corpora. Then, the top context words were retrieved and the similarity curve was plotted using Arabic Semantic Change tool. Figure 10 shows that the similarity curve has a steep increase in the latter decades, which indicates that vectors are not similar and a semantic change has occurred to the word (gay). Thus, the VSM semantic change plotting result is equal to one.

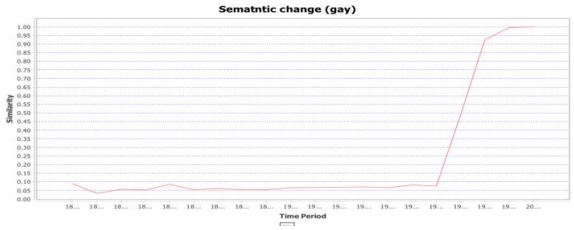

Figure 10 Word (gay) similarity curve

Next, the top context words were retrieved using the Arabic Semantic Change tool and the meanings of the word in different time periods were manually obtained from the list of top context words. Table 4 shows the retrieved top context words and the obtained meanings (last row of Table 3) of the word (gay).

Furthermore, the obtained meanings were compared, and new meanings appeared in the last two decades (lesbian and bisexual) which indicates that a semantic change has occurred to the word (gay). Thus, the VSM semantic change algorithm result is equal to one, and the overall semantic change value is equal to one.

Afterward, the Arabic Semantic Change model was used to represent the semantic change identification information as Linked Data. Figure 11 shows the word (gay) Linked Data representation using ASC model. The similarity score is equal to one because we are comparing the last decade vector with itself in other decades this value varies.

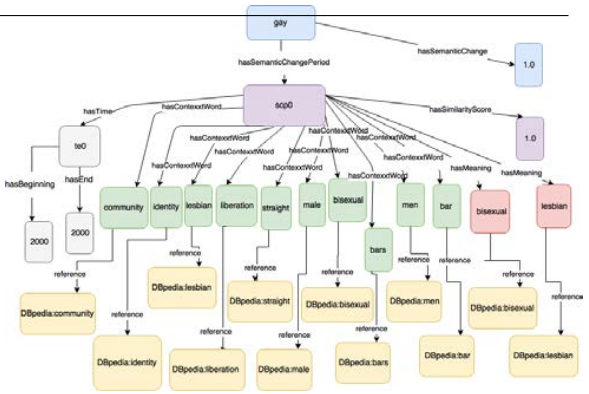

Figure 11 Word (gay) Linked Data representation using ASC model

8. Conclusion

In this paper, the modified Vector Space Model (VSM) algorithm to identify semantic change in the Arabic language was presented. Also, The Arabic Semantic Change (ASC) Model to represent as Linked Data the semantic change identification information using VSM was proposed. The algorithm and model were evaluated using Arabic and English words. Therefore, the proposed ASC model could be used with any language as shown in this paper. The words represented using ASC could be expanded and published as a Linked Dataset.

9. Acknowledgments

This project is granted by King Abdulaziz City for Science and Technology (The Graduate Studies Grant Number 37-1233). We appreciate Dr. Abdulmohsen Al-Thubaity exceptional support and insights.

Table 4 Word (gay) top context words and manually obtained meanings in different decades

2000	1990	1980	1970	1960	1950	1940	1930	1920	1910	1900	1890	1880	1870	1860	1850	1840	1830	1820	1810
liberation	liberation	liberation	lively	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant	gallant
male	male	movement	happy	noisy	blades	nineties	charming	flowers	flowers	flowers	plumage	plumage	fantastic	fantastic	fantastic	fantastic	fantastic	rose's	rose's
bars	bars	community	flowers	blades	spirited	charming	flowers	throng	throng	animated	flowers	flowers	flowers	flowers	flowers	flowers	flowers	dissolute	dissolute
identity	couples	male	colors	nineties	nineties	flowers	laughter	laughter	laughter	throng	throng	throng	throng	animated	animated	animated	animated	fantastic	sprightly
couples	identity	bars	men	charming	charming	hearted	hearted	careless	apparel	laughter	apparel	careless	careless	throng	sprightly	sprightly	sprightly	flowers	witty
bar	bar	rights		flowers	flowers	laughter	careless	laugh	careless	hearted	careless	laugh	laugh	sprightly	laugh	laugh	careless	animated	dissipated
straight	straight	identity		laughter	laughter	colorful	laugh	colors	laugh	apparel	laugh	joyous	joyous	laugh	joyous	joyous	dissipated	sprightly	men
men	men	bar		colorful	hearted	witty	colors	men	joyous	careless	joyous	colors	colors	careless	dissipated	dissipated	lively	careless	lively
lesbian	lesbian	straight		witty	colorful	laugh	men	lively	colors	laugh	colors	lively	graceful	joyous	colors	colors	happy	vermeil	thoughtless
movement	movement	males		laugh	witty	men	lively	happy	men	joyous	lively	happy	lively	colors	graceful	lively	gaudy	dissipated	sportive
community	community	men		men	laugh	colors	happy	cheerful	lively	colors	happy	thoughtless	happy	graceful	lively	happy	thoughtless	lively	cockade
bisexual	bisexual			colors	colors	lively	cheerful	colored	happy	lively	thoughtless	cheerful	thoughtless	lively	happy	thoughtless	glittering	happy	dresses
rights	rights			lively	men	friendly	colored	coloured	cheerful	happy	cheerful	fashionable	cheerful	happy	thoughtless	glittering	cheerful	thoughtless	attire
				friendly	friendly	happy	coloured	festive	colored	cheerful	fashionable	colord	fashionable	thoughtless	glittering	cheerful	fashionable	glittering	courtier
				happy	lively	cheerful	festive	ribbons	coloured	colored	colored	gorgeous	colored	cheerful	cheerful	fashionable	attire	cheerful	giddy
				cheerful	happy	colored	flags	flags	festive	festive	festive	festive	gorgeous	colored	fashionable	gorgeous	brilliant	fashionable	licentious
				colored	cheerful	festive	brilliant	brilliant	ribbons	flags	flags	flags	festive	fashionable	gorgeous	dresses	colours	attire	motes
				festive	colored	carefree	colours	colours	flags	attire	dresses	dresses	flags	gorgeous	dresses	attire	licentious	amusing	fantastic
				carefree	carefree	flags			attire	brilliant	attire	attire	dresses	dresses	attire	brilliant		colours	flowers
				flags	festive	brilliant			brilliant		brilliant	brilliant	attire	attire	brilliant	colours		licentious	enamell
				brilliant	flags	amusing			colours				brilliant	brilliant					animated
				amusing	amusing								merry						careless
				colours	brilliant														vermeil
					lighthearted														joyous
																			happy
																			glittering
																			cheerful
																			fashionable
																			amusing
																			colours
Lesbian bisexual	Lesbian bisexual		Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful	Happy cheerful

REFERENCES

Abele, A., McCrae, J. P., Buitelaar, P., Jentzsch, A., & Cyganiak, R. (2017). Linking Open Data Cloud Diagram. [cited at: 20-9-2017}, Retrieved from http://lod-cloud.net/

Abuhadeemah, T. (2008). *Studies in Arabic Dictionaries and Semantics* Riyadh, Saudi Arabia: Dar Almarefah for Human Development

Al-Thubaity, A. O. (2015). A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction. *Language Resources and Evaluation, 49*(3), 721-751.

Beckett, D. (2004). RDF/XML Syntax Specification [cited at: September 20, 2017}, Retrieved from http://www.w3.org/TR/REC-rdf-syntax/

Bizer, C. (2009). The Emerging Web of Linked Data. *Intelligent Systems, IEEE, 24*(5), 87-92. doi:10.1109/mis.2009.102

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*. doi:10.4018/jswis.2009081901

Carroll, J. J., & Pan, J. Z. (2006). XML Schema Datatypes in RDF and OWL. [cited at: March 1, 2018},

Chiarcos, C., McCrae, J., Cimiano, P., & Fellbaum, C. (2013). Towards Open Data for Linguistics: Linguistic Linked Data *New Trends of Research in Ontologies and Lexical Resources* (pp. 7-25): Springer.

Chowdhury, G. G. (2010). *Introduction to Modern Information Retrieval*: Facet publishing.

DBpedia. (2018). [cited at: March 1, 2018},

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American society for information science, 41*(6), 391.

Fellbaum, C. (1998). *WordNet*: Wiley Online Library.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. *Studies in linguistic analysis*.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Human Factors and Behavioral Science: Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems. *The Bell System Technical Journal, 62*(6), 1753-1806.

Golub, G. H., & Reinsch, C. (1970). Singular Value Decomposition and Least Squares Solutions. *Numerische mathematik, 14*(5), 403-420.

Gulordava, K., & Baroni, M. (2011). *A Distributional Similarity Approach to The Detection of Semantic Change in the Google Books Ngram Corpus.* Paper presented at the Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics.

Harris, Z. S. (1954). Distributional Structure. *Word, 10*(2-3), 146-162.

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1* (1st edition ed.): Morgan & Claypool.

Huang, A. (2008). *Similarity Measures for Text Document Clustering.* Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

Hurford, J. R. (2007). *Semantics: A Coursebook*: Cambridge University Press.

Issa, F., & Issa, R. (2008). *Semantics the Theory and the Application*. Alexandria, Egypt: Dar Almarefah Algameiah.

Jatowt, A., & Duh, K. (2014). *A Framework for Analyzing Semantic Change of Words Across Time.* Paper presented at the Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries.

Khan, F., Boschetti, F., & Frontini, F. (2014). *Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources.* Paper presented at the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing.

Löbner, S. (2002). *Understanding Semantics*: Taylor and Francis Group.

Lyons, J. (1977). *Semantics* (Vol. 53): Cambridge University Press.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Río, J. G. d., Hollink, L., Montiel-Ponsoda, E., & Spohrx, D. (2012a). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation, 46*(4), 701-719.

McCrae, J., Montiel-Ponsoda, E., & Cimiano, P. (2012b). Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*, 25-34.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., & Orwant, J. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *science, 331*(6014), 176-182.

Qalalah, A. (2017). *Semantic Development Signs and Issues: A Study in Language Measures for Ibn Fares*. Irbid, Jordan: Alam Alkutob AlHadeeth.

Rodda, M. A., Senaldi, M. S., & Lenci, A. (2016). Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *CLiC it*, 258.

Saeed, J. I. (1997). Semantics: Oxford: Blackwell Publishing.

Time Ontology in OWL. (2017). [cited at: November 7, 2017}, Retrieved from https://www.w3.org/TR/owl-time/ -time:TemporalEntity

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research, 37*, 141-188.

van Aggelen, A., Hollink, L., & van Ossenbruggen, J.

(2016). *Combining Distributional Semantics and Structured Data to Study Lexical Change.* Paper presented at the European Knowledge Acquisition Workshop.

Weaver, W. (1955). Translation. *Machine translation of languages, 14*, 15-23.

Wijaya, D. T., & Yeniterzi, R. (2011). *Understanding Semantic Change of Words Over Centuries.* Paper presented at the Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web.

Managing Provenance and Versioning for an (Evolving) Dictionary in Linked Data Format

Frances Gillis-Webber[1]
[1]University of Cape Town, Woolsack Drive, Rondebosch, 7701, South Africa
fran@fynbosch.com

Abstract

The *English-Xhosa Dictionary for Nurses* is a unidirectional dictionary with English and isiXhosa as the language pair, published in 1935 and recently converted to Linguistic Linked Data. Using the Ontolex-Lemon model, an ontological framework was created, where the purpose was to present each lexical entry as "historically dynamic" instead of "ontologically static" (Veltman, 2006:6, cited in Rafferty, 2016:5), therefore the provenance information and generation of linked data for an ontological framework with instances constantly evolving was given particular attention. The output is a framework which provides guidelines for similar applications regarding URI patterns, provenance, versioning, and the generation of RDF data.

Keywords: provenance, versioning, multilingualism, lexicography, linked data, ontolex-lemon

1. Introduction

The *English-Xhosa Dictionary for Nurses* (EXDN) is a bilingual dictionary of medical terms, authored by Neil MacVicar, a medical doctor, in collaboration with isiXhosa-speaking nurses. It was the second edition published by Lovedale Press, a South African publisher, in 1935, and as a literary work published in South Africa, it falls under the jurisdiction of the Copyright Act of South Africa, and is now in the public domain, free from any restriction. EXDN is unidirectional; the language pair is English and isiXhosa, with English as the source and isiXhosa the target (Gouws & Prinsloo, 2005; Zgusta, 1971). IsiXhosa (referred here by its endonym) is an indigenous Bantu language from the Nguni language group (S40 in Guthrie's classification) and is an official language of South Africa (Doke, 1954; "Subfamily: Nguni (S.40)", n.d.). Despite it being spoken in South Africa by a large percentage of the population (16.0% counted in the 2011 Census speak it as their L1), it has minority status only (Statistics South Africa, 2012).

Other official South African languages in the Bantu language family (referred hereon as African languages) are: isiNdebele, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, SiSwati, Tshivenḓa, and Xitsonga. In comparison to English, there are limited language resources (LRs) available for these languages, and this, combined with the socio-economic constraints of the speakers, renders these languages under-resourced (Pretorius, 2014; "What is a …", n.d.). Despite English being an ex-colonial language in South Africa, with L1 speakers numbering 9.6%, it is a lingua França with high status, associated with both economic and political power in the country (Ngcobo, 2010; Statistics South Africa, 2012). The African languages listed above, although spoken by the majority, are minority languages and through language shift and death, are at risk of becoming endangered (Pretorius, 2014; Ngcobo, 2010).

In 2009, the first Human Language Technology Audit was conducted (the second audit is currently underway at time of writing (Wilken, personal communication 2017, Dec 12)), with Grover, van Huyssteen and Pretorius identifying the following as issues:

> *"the lack of language resources, limited availability of and access to existing LRs, [and] quality of LRs"*

which hamper the development of new LRs for under-resourced languages (2011, cited in Pretorius, 2014). Although EXDN was published more than seventy-five years ago, as a LR for an under-resourced language, its content is still valuable. Linked Data is a simple data model with an interoperable format and by publishing lexicographic resources, particularly lesser-known resources such as EXDN, in Linked Data, it enables the "aggregation and integration of linguistic resources", which can serve as an aid for the future development of new and existing LRs (Gracia, 2017).

Using the Ontolex-Lemon model, an ontological framework was created, where the purpose was to present each lexical entry as "historically dynamic" instead of "ontologically static" (Veltman, 2006:6, cited in Rafferty, 2016:5), therefore the provenance information and generation of Linked Data for an ontological framework with instances constantly evolving was given particular attention.

The rest of the paper is organised as follows: in Section 2, the structure of the dictionary is briefly described; in Section 3, the URI strategy is discussed; in Sections 4 and 5, the description of resources, provenance for lexical entries and the lexicons are considered, and the versioning and generation of Linked Data is presented. The conclusions of the paper are presented in Section 6.

2. The Structure of the Dictionary

The frame structure of a dictionary is typically composed of the central list, with front and back matter texts (Gouws & Prinsloo, 2005); however, the frame structure of EXDN consists of a central list, with front matter texts only. The central list of EXDN is represented by the Roman alphabet, with each letter acting as a guiding element for a series of article stretches (Gouws & Prinsloo, 2005).

EXDN can also be described according to its macrostructure and microstructure. EXDN's macrostructure comprises a lemmatised list in the source language only: English - ordered alphabetically with a

singular and plural lemmatisation of nouns (Gouws & Prinsloo, 2005). A dictionary's microstructure pertains to the structure of each article (lexical entry), with the lemma serving as a guiding element for each (Gouws & Prinsloo, 2005). In the case of EXDN, each article comprises one of the following (or a combination thereof): lexicographic definition, a translation, or a cross-reference entry. If the article has a single target language item, shown by a single word, then it is presumed that the article is a translation equivalent, with full equivalence (Gouws & Prinsloo, 2005). However, if the article has a lexicographic definition in the target language, then zero equivalence is presumed (Gouws & Prinsloo, 2005).

3. The URI Strategy

Archer, Goedertier and Loutas have defined a URI as "a compact sequence of characters that identifies an abstract or physical resource" and it "can be further classified as a locator, a name, or both" (2012).

A key set of principles have been identified for URIs:

- URIs should be:
  - short,
  - stable,
  - persistent, and
  - human-friendly (Archer, Goedertier & Loutas, 2012; Hogan et al., 2012; Wood et al., 2014).
- URIs should be HTTP(S) URIs (Berners-Lee, 2006; Hogan et al., 2012).
- The identifier portion of a URI should be:
  - unique,
  - and unambiguous (Simons & Richardson, 2013; Keller et al., 2011).
- URIs should be dereferenceable, with a representation returned when a human or software agent navigates to the URI (Heath & Bizer, 2011; Hyvönen, 2012).
- URIs should differentiate between the resource, and the document which describes a resource (Van Hooland & Verborgh, 2014; Heath & Bizer, 2011).

In the sub-sections that follow, fragment identifiers, URI patterns, and resource identifiers are discussed in more detail.

3.1 Fragment Identifiers

Fragment identifiers are an optional part of the URI, positioned at the end, and are of the pattern "#example". Although the usage of fragment identifiers have been cautioned against by Wood et al., primarily because web servers do not process the fragment, they are widely used in vocabularies, where "the vocabulary is often served as a document and the fragment is used to address a particular term within that document" (2014). Within the context of identifying sub-resources in relation to the parent resource, fragment identifiers can be useful, as they can clearly show a hierarchical relationship with the parent resource (however, deeper levels cannot be indicated).

According to Sachs and Finin, the URI should resolve "not to the address, but to all known information about the resource" (2010); from this one can infer that when information for a sub-resource is returned, then information for the parent resource should also be returned. Conversely, when information for a parent resource is returned, information of any sub-resources should also be returned. By doing this, the need to have a separate document to describe the parent resource and each of the sub-resources is not necessary, as one document can be used to describe the parent resource and any sub-resources.

Additionally, when publishing Linked Data and versioning is employed, by using fragment identifiers to identify sub-resources within the same document, redundancy can be reduced.

3.2 The URI Pattern

When working with EXDN data, the following use cases were determined (Gillis-Webber, 2018):

U1: *A URI which identifies a resource*

U2: *A URI which identifies a sub-resource in relation to the parent resource*

U3: *A URI which identifies a version of the resource*

U4: *A URI which identifies a version combined with a sub-resource*

U5: *A URI which identifies a document describing the resource in **U1***

U6: *A URI which identifies a document describing the resource in **U3***

A pattern for a URI has been recommended by Archer et al. (2012):

```
http://{domain}/{type}/{concept}/{reference}
```

Where:

- {domain} is the host,
- {type} is the resource (for eg. *id*) being identified,
- {concept} refers to a real world object or a collection, and
- {reference} is the local reference for the resource being identified.

When using the *lemon* model (a previous iteration of the Ontolex-Lemon model), Gracia and Vila-Suero developed a set of guidelines for publishing Linked Data for bilingual dictionaries, and they too proposed the same pattern as Archer et al. (2015). As an example, for the lexical entry "bench", the URI is as follows:

E1: `http://linguistic.linkeddata.es/id/apertium/lexiconEN/bench-n-en`

Where:

- *linguistic.linkeddata.es* is the host,
- *id* is the resource,
- *apertium* is the collection,
- *lexiconEN* is the source lexicon,
- *bench-n-en* is the reference.

When considered from a user perspective, the human-friendliness of **E1** can be evaluated accordingly:

- *id* is not particularly informative and could be deemed redundant;
- although specifying the collection (*apertium*) is useful, should a dataset from another collection be merged with the existing dataset, if there are shared lexical entries between both collections, this will result in URIs which are incongruently defined;
- both the lexicon and the reference are identifiable as English, thus *lexiconEN* could also be deemed redundant.

Ontolex-Lemon (and *lemon* as well) requires the lexical entries in a lexicon to be the same language. If modelling two languages, then the lexical entries of each language would be contained within their own lexicon, with translation relations explicitly defined between the corresponding lexical entries or their senses, using the *vartrans* module ("Final model specification", n.d.). BabelNet was also modelled on *lemon*, and by 2015 it had 271 lexicons, one for each of the languages it supported; Flavi et al. remarked on this saying *lemon* requires "us to work on a language-by-language basis, whereas in BabelNet this distinction does not need to be made explicit".

Continuing with the example lexical entry "bench", in BabelNet, the URI is as follows:

E2: `http://babelnet.org/rdf/bench_n_EN`

There should be a separation between the URIs and the model used to describe the lexical data. If the model should change, the persistence and longevity of the URIs should not be impacted, and as a result, a "URI should be agnostic of the selected model" (Gillis-Webber, 2018). For **E1** and **E2**, both the references (*bench-n-en* and *bench_n_EN* respectively) have been encoded with additional information by appending the lemma with the language shortcode and an abbreviated form of part-of-speech (POS), and by doing this, the URIs for the two examples are identifiable to be of the English language with POS noun.

E1 could therefore be revised to:

`http://linguistic.linkeddata.es/entry/bench-n-en`

And for a lexicon:

`http://linguistic.linkeddata.es/lexicon/en`

For each of the six use cases identified for EXDN at the beginning of Section 3.2, the application of this simplified pattern has continued, and below, the pattern of each use case is provided, followed by a short description thereof, as well as an associated example from Londisizwe.org, the multilingual online dictionary derived from the EXDN dataset.

A URI which identifies a resource has the form (Gillis-Webber, 2018):

U1: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}*

Where:

- {http(s):} is the http: or https: scheme
- {Base URI} is the host
- {Resource Path}, for example, *entry* for a lexical entry, and *lexicon* for a lexicon
- {Resource ID}, for example, *en-n-abdomen*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen`

A URI which identifies a sub-resource in relation to the parent resource has the form (Gillis-Webber, 2018):

U2: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}#{Fragment ID}*

Where:

- {Fragment ID} is the fragment identifier, for example, *sense1*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen#sense1`

The resource identifier, described in **U1**, will be unique relative to the resource path. The fragment identifier will be unique relative to the resource identifier.

A URI which identifies a version of the resource has the form (Gillis-Webber, 2018):

U3: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}/{Version ID}*

Where:

- {Version ID} is the version identifier, for example, *2017-09-19*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen/2017-09-19`

As the sub-resource is identified in relation to the parent resource, any change to the sub-resource would result in a change to the URI of the parent resource.

Therefore, a URI identifying a sub-resource when employing the use of versioning has the form (Gillis-Webber, 2018):

U4: *{http(s):}//{Base URI}/{Resource Path}/{Resource ID}/{Version ID}#{Fragment ID}*

An example URI is:

`https://londisizwe.org/entry/en-n-abdomen/2017-09-19#sense1`

For a resource, each version should be dereferenceable, and should remain so even as newer versions of the same resource are published. Like that of the fragment identifier, the version identifier is unique to the resource

identifier. The use case **U1** will resolve to the latest version available for that resource (Archer et al., 2012).

A URI which identifies a document describing the resource in **U1** has the form (Gillis-Webber, 2018):

U5: *{http(s):}//{Base URI}/ {Document}/{Resource Path}/{Resource ID}*

Where:

- Using content negotiation, {Document} refers to the HTML page, for example, *page*, or to the RDF representation, for example, *rdf*, using any form of serialisation.

Corresponding examples are:

`https://londisizwe.org/page/entry/en-n-abdomen`

`https://londisizwe.org/rdf/entry/en-n-abdomen`

A URI which identifies a document describing the resource in **U3** has the form (Gillis-Webber, 2018):

U6: *{http(s):}//{Base URI}/ {Document}/{Resource Path}/{Resource ID}/{Version ID}*

An example URI is:

`https://londisizwe.org/rdf/entry/en-n-abdomen/2017-09-19`

In the context of EXDN, a document which describes **U2** (or **U4**) is not necessary, and instead it resolves to **U5** (or **U6**).

3.3 Resource Identifiers

The human-friendliness of URIs has been suggested in the literature, with frequent references thereto: such as URIs should be "user-friendly" (Archer, Goedertier & Loutas, 2012), "human readable" (Hogan et al., 2012), "meaningful" (Villazón-Terrazas et al., 2012), and "natural keys" should be used (Wood et al., 2014; Heath & Bizer, 2011). Defined by Labra Gayo, Kontokostas and Auer (n.d.) as "descriptive URIs", and as "meaningful URIs" by Vila-Suero et al. (2014), this type of URI is generally used "with terms in English or in other Latin-based languages" (Labra Gayo, Kontokostas & Auer, n.d.).

Labra Gayo et al. defines "opaque URIs" as "resource identifiers which are not intended to represent terms in a natural language", with it suggested by both Labra Gayo, Kontokostas and Auer (n.d.), and Vila-Suero et al. (2014) that in a multilingual context, using opaque URIs is preferable so as to avoid language bias. By doing so within the context of the Semantic Web, Vila-Suero et al. argue that this is acceptable, as "resource identifiers are intended for machine consumption so that there is no need for them to be human readable" (2014).

Within the larger context of the Semantic Web, this view may be accurate as data models are mostly language-agnostic (Ehrmann, 2014), however in the context of Linked Data, it is in opposition to a fundamental principle thereof: a URI should be dereferenceable, to be looked up by either a web browser for human consumption or a software agent (Hyvönen, 2012).

Due to the localisation of this study within South Africa and its languages being Latin-based, a pragmatic approach was taken with regards to the URIs: descriptive URIs were used, using English, however in a similar approach to Babelnet, opaque URIs were used when modelling the lexical concepts (Flati et al., 2015).

For lexical entries, a similar approach as that used in **E1** was taken for the resource identifiers, however the elements were reordered to aid programmatic extraction (should it be required):

`{Language Code}-{POS}-{Lemma}`

Where:

- {Language Code} is the lowercase form of the language shortcode, using ISO 639-1, and if none available, then ISO 639-2 (or ISO 639-3) will be used
- {POS} is an abbreviated form of POS, described in English
- {Lemma} is the lowercase form of the lemma, with underscores replacing any hyphens or spaces and any diacritics are removed

For a lexical entry, a constraint of the Ontolex-Lemon model is that it can be associated with exactly one POS and exactly one language ("Final model specification", n.d.). For lexical entries which may share the same lemma, such as:

isiXhosa: *isibindi*
isiZulu: *isibindi*

to avoid potential collision, it was considered best for the EXDN dataset to include the language shortcode and the abbreviated POS in the identifier as well, thus allowing for the easy extensibility of the existing dataset to additional languages. Thus for the two lexical entries above, their identifiers would be as follows:

isiXhosa: `xh-n-isibindi`
isiZulu: `zu-n-isibindi`

For a lexicon, the resource identifier takes the form (shown here including the resource path):

`{Resource Path}/{Language Code}`

In combination with the resource path, the resource identifier should adequately identify the lexical entry (or lexicon), thus allowing for any language to be represented (with the exception of the written form of sign languages, which can conceivably be any language) (Gillis-Webber, 2018).

4. The Description of Resources

As previously mentioned, when returning information for a resource and any of its sub-resources, the information returned should not be limited to describing these resources, the inclusion of the following additional information could be considered as well (Gillis-Webber, 2018):

- *A description of related resources;*

- *A description of the metadata of the resource (for example, provenance and version);*
- *A description of the dataset which contains the resource (Heath & Bizer, 2011:45).*

In the case of EXDN, when publishing the information for a lexicon which resolves, for example, to the URI `https://londisizwe.org/lexicon/en`, it was not considered practical to include information of the related resources, particularly for each lexical entry. However, when publishing the information for a lexical entry which resolves, for example, to the URI `https://londisizwe.org/entry/en-n-abdomen`, it was considered necessary, and the following additional information is thus included (Gillis-Webber, 2018):

- *Description of the document which describes the lexical entry,*
- *Metadata of the lexical entry,*
- *Provenance information of the lexical entry,*
- *Identification of the lexicon to which the entry belongs,*
- *Brief description of other lexical entries, resources and ontology entities related to the lexical entry.*

5. Modelling Provenance & Versioning

According to Di Maio (2015), knowledge is "partial/incomplete/imperfect, with very few exceptions". Linked Data is about relationships, and when considered within the context of Linguistics, datasets of different lexicons can be interlinked, thus allowing for the extension of an existing lexicon; for under-resourced languages, this can be a powerful notion (Berners-Lee, 2009; McCrae et al., 2012). According to Bouda and Cysouw (2012), when retrodigitising language resources, the encoding thereof is not the challenge, but rather "the continuing update, refinement, and interpretation" of the dataset, and with each change, providing for traceability. Like RDF datasets, ontologies and vocabularies are not static, and they too evolve over time (Hyvönen, 2012). This change can be attributed to factors such as error correction, the addition of concepts and properties to the underlying model, as well as change out in the world, and our understanding thereof (Hyvönen, 2012).

As mentioned in Section 2, within the context of EXDN, until established otherwise, then full equivalence is presumed if the article has a single target language item, and if anything more than a single target language, then it is presumed the article is a lexicographic definition and there is zero equivalence (Gouws & Prinsloo, 2005).

Google's Cloud Translation API[1] was used to translate the isiXhosa texts, with English selected as the target language. There are two models available: Phrase-Based Machine Translation model (PBMT) and Neural Machine Translation model (NMT), and using each model, an article was translated ("Translating text", n.d.). As an example, the article *stomach*, which has the isiXhosa text of "Uluusu lomntu.", when translated on 2017-09-17 20:00:31 GMT+2, yielded the following:

PBMT: A person's skin.
NMT: Homosexuality.

[1] https://cloud.google.com/translate/

There are several possibilities for this: (1) the source data contains errors, (2) the source data is so outdated that it is not possible to translate this accurately, or (3) there are not enough existing language pairs within the Cloud Translation API to accurately translate the text ("Cloud translation API", n.d.). According to Google's website, the Cloud Translation API undergoes continuous updates ("Cloud translation API", n.d.) so although it is intended to periodically repeat the translation process for the EXDN dataset, for now, the translated texts are not used for disambiguation purposes.

Continuing with the article *stomach*, when the lexical entry with the identifier `en-n-stomach` was first published in 2017, its only sense (`en-n-stomach#sense1`) was linked to a lexical concept (`https://londisizwe.org/concept/000000007`) which had a language-tagged lexicographic definition `"Uluusu lomntu."@xh`, and it was set as a concept of the DBpedia resource: `http://dbpedia.org/resource/Stomach`. However, after consultation with isiXhosa mother tongue speakers in early 2018, the following was determined:

- "uluusu" was incorrectly spelt in EXDN (it should have been "ulusu"),
- the equivalent of *stomach* is also *isisu*,
- the meaning (gloss) of "ulusu lomntu" is "a person's stomach", however it was difficult to determine if the text should remain a lexicographic definition or if it should become a lexical entry with "ulusu lomntu" as the lemma.

As a result of this new information, the following changes were implemented:

- For the lexical entry `en-n-stomach`, the spelling mistake was corrected in the lexical concept.
- The lexical entry `xh-n-isisu` already existed, however another sense was added (`xh-n-isisu#sense2`), and it was linked to the same lexical concept.

Because there is a shared conceptualisation between `https://londisizwe.org/entry/en-n-stomach#sense1` and `https://londisizwe.org/entry/xh-n-isisu#sense2`, they are deemed to be equivalent.

As the purpose of digitising EXDN and converting its dataset to Linked Data is to enable its reuse by external resources, it is important that any changes are accurately recorded, by way of versioning, with provenance information included as well. The lexical entry `xh-n-isisu` had a change to one of its senses (a sub-resource), and the lexical concept `000000007` changed as well, so there is now a new version for each. As there were not any insertions or deletions for the English and isiXhosa lexicons, these remained unchanged. In the event the lexical entries had to be reviewed again, it is expected they would be subject to further refinement.

As an aside, the Cloud Translation API was used again (2018-03-02 20:36:48 GMT+2), this time with the corrected text "Ulusu lomntu.". PBMT remained unchanged, however NMT returned the following translation: "Human skin.". It was also repeated for the original source text, and those translations remained unchanged from 2017-09-17.

5.1 Versioning

Versioning is used by Babelnet, although it is applied globally for their BabelNet-lemon schema description, with Flati et al. acknowledging that "maybe a more sophisticated infrastructure would be needed in order to express more complex versioning description needs" (2015). When the generation and publication of RDF data for the Apertium Bilingual Dictionaries was detailed by Gracia et al., versioning was not included in the discussion (n.d.). Although briefly mentioned by McCrae et al. (2012), Gracia et al. (n.d.), Eckart et al. (2012), van Erp (2012), and De Rooij et al. (2016), it does not appear that versioning has been discussed further within the domain of Linguistic Linked Data, and in the context of vocabularies used by Babelnet, Flati et al. commented that changes are unaccounted for "and this aspect might thus be investigated in more detail in the [near] future by the whole community" (2015).

When describing the generation of RDF for the Apertium Bilingual Dictionaries, Gracia et al. talked of three RDF files: one per lexicon, and the third for the translations (n.d.). From this, the author inferred that if versioning was implemented, it would be done at file-level, in a similar approach to that taken by BabelNet. However, in the context of EXDN, it was felt that publishing only at the lexicon-level could become unmanageable over time, particularly on a 24-hour publishing schedule, and instead it would be more practical to implement versioning at the lexical entry-level as well. Versioning at the lexicon-level is also done, but a file only includes the changes from the previously published version, and any additional information of the lexical entries, beyond the resource identifier, is excluded. For each version of a lexical entry, the file contains: all information of the lexical entry, its senses, and translation relations for which any of its senses is the source.

Thus, the following components have been identified for the versioning of EXDN (Gillis-Webber, 2018):

- *Versioned URIs for lexicons, lexical entries, and senses*
- *Provenance metadata to describe the versions, with the latest version mapping to previous versions (Van Erp, 2012), and*
- *The generation of files, one for each version of the lexical entries and lexicons.*

Within the context of EXDN, lexical concepts are modelled as a shared conceptualisation between senses, and they can be thought of as similar to that of a WordNet *synset*, however, where WordNet models sets of similar terms, lexical concepts model sets of equivalent senses across languages (Bosque-Gil et al, 2015). Although the lexical concepts are hosted on the same domain, they are stored within a sense inventory called *Londisizwe Concepts for Senses*[2] – this is considered to be a standalone inventory, and as a result, it is not described further here, although the same principles for versioning do apply (Gillis-Webber, 2018).

Section 3.2 introduced versioned URIs, with the use cases: **U3** and **U4**. Modelling provenance, and the generation and publication of Linked Data are discussed in the sections that follow.

5.2 Modelling Provenance for a Lexical Entry, its Senses, and Translation Relations

The W3C Provenance Working Group defines provenance ("PROV-O", 2013):

as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.

A factor contributing to the reuse of a RDF dataset, either by linking or by using the downloaded data, is trust – trust in the repository supplying the data, and trust in the data itself (Faniel & Yakel, 2017). By documenting the provenance of data using a systematic schema, provenance provides a trust marker (essential in an open environment like the web); and within the context of EXDN, provenance information is documented using the PROV Ontology, DCMI Metadata terms, and versioned URIs (Faniel & Yakel, 2017; "PROV-O", 2013; Tennis, 2007; Flati et al., 2015).

The metadata used to describe the EXDN dataset is as follows:

- Each lexical entry, sense, and translation relation is identified as a `prov:Entity`.
- The `prov:generatedAtTime` property is recorded for each.
- The date a lexical entry, sense or translation relation is changed is recorded using `dct:modified`.
- The person or organisation responsible for creating the lexical entry or sense is identified using `dct:creator`.
- The source from which a lexical entry is primarily derived is identified using the `prov:hadPrimarySource` property.
- The other sources from which a lexical entry, sense or translation relation is derived, is identified using the `dc:source` property.
- One or more contributors (a person, an organisation or a service) for a lexical entry, sense or translation relation is identified using `dct:contributor`.
- The licensing agreement for a lexical entry is identified using `dct:license`, and Creative Commons is used for the licensing.
- For a lexical entry, `dct:isPartOf` is used to denote inclusion of a lexical entry in a lexicon, and inclusion of a sense in a lexical entry.
- For a translation relation, `dct:hasPart` is used to identify both the source and target language.
- For a lexical entry, `owl:sameAs` is used to indicate that **U1** is the same as the latest version of **U3**.
- For a sense or translation relation, `owl:sameAs` is used to indicate that **U2** is the same as the latest version of **U4**.
- For a lexical entry, sense or translation relation, the version is indicated using `owl:versionInfo`.
- For a lexical entry, sense or translation relation, `dct:hasVersion` is used to show the previously generated versions, using the versioned URIs (**U3** for lexical entries and **U4** for senses and translation relations).

[2] https://londisizwe.org/concept

The generated RDF for version two of the lexical entry `xh-n-isisu` follows below. The lexical concept for `000000001` is also shown for reference purposes.

```
LINE
   1  @prefix :      <https://londisizwe.org/> .
   2  @prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
   3  @prefix dbr:   <http://dbpedia.org/resource/> .
   4  @prefix dct:   <http://purl.org/dc/terms/> .
   5  @prefix foaf:  <http://xmlns.com/foaf/0.1/> .
   6  @prefix lcnaf: <http://id.loc.gov/authorities/names/> .
   7  @prefix lcsh:  <http://id.loc.gov/authorities/subjects/> .
   8  @prefix lexinfo:
   *              <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
   9  @prefix mesh:  <http://id.nlm.nih.gov/mesh/> .
  10  @prefix owl:   <http://www.w3.org/2002/07/owl#> .
  11  @prefix prov:  <http://www.w3.org/ns/prov#> .
  12  @prefix pwn:   <http://wordnet-rdf.princeton.edu/rdf/id/> .
  13  @prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
  14  @prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
  15  @prefix void:  <http://rdfs.org/ns/void#> .
  16  @prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
  17
  18  <https://londisizwe.org/rdf/entry/xh-n-isisu>
  19      rdfs:label   "RDF document for the lexical entry:
   *                  isisu, n (isiXhosa)"@en ;
  20      rdf:type     foaf:Document ;
  21      foaf:primaryTopic :entry/xh-n-isisu .
  22
  23  :entry/xh-n-isisu
  24      a ontolex:LexicalEntry , ontolex:Word , prov:Entity ;
  25      lexinfo:partOfSpeech  lexinfo:Noun ;
  26      dct:language <http://id.loc.gov/vocabulary/iso639-2/xho> ,
   *                  <http://lexvo.org/id/iso639-1/xh> ;
  27      dct:identifier :entry/xh-n-isisu ;
  28      rdfs:label    "isisu"@xh ;
  29      ontolex:canonicalForm :entry/xh-n-isisu#lemma ;
  30      ontolex:sense :entry/xh-n-isisu#sense1 ,
  31                    :entry/xh-n-isisu#sense2 ;
  32      dct:subject   mesh:D000005 ;
  33      ontolex:denotes dbr:Abdomen , dbr:Stomach ;
  34      ontolex:evokes :concept/000000001 ;
  35      dct:isPartOf  :lexicon/xh ;
  36      dct:license
   *      <http://creativecommons.org/publicdomain/mark/1.0/> ;
  37      prov:hadPrimarySource "The English-Xhosa Dictionary for
   *                  Nurses"@en ;
  38      dct:creator   <https://londisizwe.org> ;
  39      prov:generatedAtTime "2018-01-
   *                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
  40      dct:modified  "2018-01-10"^^xsd:date ;
  41      owl:versionInfo "2018-01-10"^^xsd:string ;
  42      owl:sameAs     :entry/xh-n-isisu/2018-01-10 ;
  43      owl:hasVersion :entry/xh-n-isisu/2017-09-19 ,
  44                     :entry/xh-n-isisu/2018-01-10 .
  45
  46  :entry/xh-n-isisu#lemma
  47      a             ontolex:Form ;
  48      ontolex:writtenRep  "isisu"@xh .
  49
  50  :entry/xh-n-isisu#sense1
  51      a             ontolex:LexicalSense , prov:Entity ;
  52      ontolex:isLexicalizedSenseOf :concept/000000001 ;
  53      dct:identifier :entry/xh-n-isisu#sense1 ;
  54      dct:isPartOf  :entry/xh-n-isisu ;
  55      dct:creator   <https://londisizwe.org> ;
  56      prov:generatedAtTime "2018-01-
   *                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
  57      dct:modified  "2018-01-10"^^xsd:date ;
  58      owl:versionInfo "2018-01-10"^^xsd:string ;
  59      owl:sameAs    :entry/xh-n-isisu/2018-01-10#sense1 ;
  60      owl:hasVersion  :entry/xh-n-isisu/2017-09-19#sense1 ,
  61                      :entry/xh-n-isisu/2018-01-10#sense1 .
  62
  63  :entry/xh-n-isisu#sense2
  64      a             ontolex:LexicalSense , prov:Entity ;
  65      ontolex:isLexicalizedSenseOf :concept/000000007 ;
  66      dct:identifier :entry/xh-n-isisu#sense2 ;
  67      dct:isPartOf  :entry/xh-n-isisu ;
  68      dct:creator   <https://londisizwe.org> ;
  69      prov:generatedAtTime "2018-01-
   *                  10T05:00:00Z|+02:00"^^xsd:dateTime ;
  70      owl:versionInfo "2018-01-10"^^xsd:string ;
  71      owl:sameAs    :entry/xh-n-isisu/2018-01-10#sense2 ;
  72      owl:hasVersion  :entry/xh-n-isisu/2018-01-10#sense2 .
  73
  74  :concept/000000001
  75      a             skos:Concept , ontolex:LexicalConcept ;
  76      ontolex:lexicalizedSense :entry/en-n-abdomen#sense1 ;
  77      ontolex:lexicalizedSense :entry/xh-n-isisu#sense1 ;
  78      owl:sameAs     pwn:05564576-n ;
  79      owl:sameAs     mesh:M000005 ;
  80      dct:subject    mesh:D000005 ;
  81      ontolex:isConceptOf  dbr:Abdomen .
```

Figure 1: Modelling of a lexical entry

5.3 Modelling Provenance for a Lexicon

Using the same principles from the previous sections, as well as the *lime* module from Ontolex-Lemon, the metadata of EXDN's isiXhosa lexicon is described below in RDF. The metadata only serves to describe the lexicon, and when a lexical entry is inserted or removed from a lexicon is not described. However, PROV-Dictionary[3], published by the W3C Provenance Working Group in 2013 as an extension to PROV, "introduces a specific type of collection, consisting of key-entity pairs", thus allowing for the change of lexical entries in a lexicon, as members of a collection, to be expressed as well ("PROV-Dictionary: Modeling provenance …", 2013).

The generated RDF for version three of the lexicon `xh` follows below:

```
LINE
   1  @prefix :      <https://londisizwe.org/> .
   2  @prefix lime:  <http://www.w3.org/ns/lemon/lime#> .
   3  @prefix dct:   <http://purl.org/dc/terms/> .
   4  @prefix foaf:  <http://xmlns.com/foaf/0.1/> .
   5  @prefix owl:   <http://www.w3.org/2002/07/owl#> .
   6  @prefix prov:  <http://www.w3.org/ns/prov#> .
   7  @prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
   8  @prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
   9  @prefix void:  <http://rdfs.org/ns/void#> .
  10  @prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
  11
  12  <https://londisizwe.org/rdf/lexicon/xh>
  13      rdfs:label   "RDF document for the lexicon: isiXhosa"@en ;
  14      rdf:type     foaf:Document ;
  15      foaf:primaryTopic :lexicon/xh .
  16
  17  :lexicon/xh
  18      a lime:Lexicon , void:Dataset ,
   *        prov:Dictionary , prov:Collection , prov:Entity ;
  19      lime:language  "xh" ;
  20      dct:language <http://id.loc.gov/vocabulary/iso639-2/xho> ,
   *                  <http://lexvo.org/id/iso639-1/xh> ;
  21      dct:identifier :lexicon/xh ;
  22      lime:lexicalEntries "1"^^xsd:integer ;
  23      lime:linguisticCatalog
   *             <http://www.lexinfo.net/ontologies/2.0/lexinfo> ;
  24      dct:description "Londisizwe.org - isiXhosa lexicon"@en ;
  25      dct:creator    <https://londisizwe.org> ;
  26      prov:generatedAtTime "2018-01-
   *                  15T06:00:00Z|+02:00"^^xsd:dateTime ;
  27      dct:modified   "2018-01-15"^^xsd:date ;
  28      owl:versionInfo "2018-01-15"^^xsd:string ;
  29      owl:sameAs     :lexicon/xh/2018-01-15 ;
  30      owl:hasVersion :lexicon/xh/2017-09-19 ,
  31                     :lexicon/xh/2018-01-12 ,
  32                     :lexicon/xh/2018-01-15 ;
  33      dct:references :lexicon/en ;
  34      void:dataDump  <https://londisizwe.org/data/xh-
   *                  lexicon/2018-01-15> .
  35
  36  :lexicon/xh/2018-01-12
  37      a              prov:Dictionary .
  38
  39  :lexicon/xh/2018-01-15
  40      a              prov:Dictionary ;
  41      prov:derivedByRemovalFrom :lexicon/xh/2018-01-12 ;
  42      prov:qualifiedRemoval [
  43          a          prov:Removal ;
  44          prov:dictionary :lexicon/xh/2018-01-12 ;
  45          prov:removedKey "xh-n-ulusu_lomntu"^^xsd:string ;
  46      ] ;
  47  .
```

Figure 2: Modelling of a lexicon

Where:

- Lines 36 – 37: the previous version is identified as a dictionary. There were two dictionary entries, although those entries are not listed here, instead they would have been listed in the file of the previously published URI: `https://londisizwe.org/lexicon/xh/2018-01-12`

[3] https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/

- Lines 39 – 40: the current version is identified as a dictionary.
- Line 41: states that the current version was derived from the previous version.
- Lines 42 – 46: indicates the key that was removed. There is now only one lexical entry, `xh-n-isisu`, in the isiXhosa lexicon.

The class `prov:Dictionary` is defined as "an entity that provides a structure to some constituents, which are themselves entities. These constituents are said to be members of the dictionary", and the concept of 'dictionary' can be extended to include "a wide variety of concrete data structures, such as maps or associative arrays" ("PROV-Dictionary: Modeling provenance …", 2013). Within the context of EXDN, while `prov:Dictionary` has only been applied to lexicons, it could conceivably also be applied to lexical entries and lexical concepts – both of which are containers, with each having senses as its members. While this has not yet been explored for the EXDN dataset, it is work that will be considered in the future.

5.4 Generation and Publication of Linked Data

In a similar vein to versioning, the generation and publication of RDF data is only briefly mentioned in the literature (Vila-Suero et al., 2014; Ehrmann et al., 2014; Gracia et al., n.d.), although for BabelNet, Ehrmann et al. did talk of RDF dump files (which no longer seem to be available for download). For the RDF files discussed in Section 5.1 for the Apertium Bilingual Dictionaries, Gracia et al. (n.d.) talked of loading them into a Virtuoso[4] triple store, with a SPARQL endpoint to access the RDF data, as well as the development of a Linked Data interface using Pubby[5]. The topic was explored further by Gracia in a presentation in 2017, recommending the use of a SPARQL store, with "a mechanism to make [our] URIs dereferenceable: through a common web server (as files)", or by making use of a Linked Data interface. According to Heath and Bizer (2011), storing static RDF files on a web server is "the simplest way to publish Linked Data", and within the context of EXDN, this was the selected route. A Dictionary Writing System was custom-developed for the purpose of maintaining the EXDN dataset, with automated processes implemented for file generation.

Because of the versioning requirements listed in the previous section, the following approach to publication is taken:

- When a lexical entry (or its senses or translation relations of which one of the senses of the lexical entry is the source) changes, a new file in the various formats required is generated. **U1** always point to the latest version of the lexical entry. This is an automated task, scheduled to run daily at 5AM.
- Lexical entries are members of the lexicon collection, and if there are any changes to the members (insertions or deletions), then a new version of the lexicon file is generated, using the same principle as that described for lexical entries. This process is repeated per lexicon.
- The files representing the latest version of the lexicon and its lexical entries are manually merged and compressed to create a data dump. It is planned to automate this process in the future. A SPARQL endpoint is currently not available, although it is planned to trial Dydra[6], a cloud-hosted RDF platform ("Dydra", 2011).

6. Conclusion

Although EXDN was published in 1935, once the dataset is fully converted to Linked Data, it will continue to evolve: with the identification of additional resources to link to; by merging with other LRs; as well as the planned implementation of a crowdsourcing approach to correct, change, and add lexicographic definitions, cross-reference entries, translations, senses, and annotations to lexical entries in multiple African languages. Within the context of EXDN, provenance and versioning has thus been identified as essential components whilst converting the dictionary to Linguistic Linked Data, as well as for its on-going improvements thereafter.

Furthermore, the lemmatisation approach for African languages, as well as annotations within a multilingual environment were modelling challenges identified by the author whilst working with the EXDN dataset. Likewise, the representation of hierarchy in RDF, be it in the form of sub-senses, or inflection, with multiple affixes attached to a word stem, has been identified as a modelling challenge by Gracia, Kernerman and Bosque-Gil (2017). Both a lexicography module and a morphology module for the Ontolex-Lemon model is in progress with the Ontology-Lexica Community Group, and when implemented, it is expected that the modelling of EXDN's lexical entries and senses may change (Bosque-Gil, 2017; McCrae & Gracia, 2017). Although the Ontolex-Lemon model takes a modular approach, as its range extends, provenance and versioning will be of importance so that any change to the RDF representation of data is accurately recorded.

7. Acknowledgements

Thank you to the reviewers for the kind advice and suggestions on improving this paper.

8. Bibliographical References

Archer, P., Goedertier, S. & Loutas, N. (2012). *D7.1.3 – Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC.* Available: https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf [2017, December 26].

Berners-Lee, T. (2006). *Linked data*. Available: https://www.w3.org/DesignIssues/LinkedData.html [2017, December 25].

Berners-Lee, T. (2009). *Tim Berners-Lee: the next web* [Video file]. Available: http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html [2017, April 15].

[4] https://virtuoso.openlinksw.com/

[5] http://wifo5-03.informatik.uni-mannheim.de/pubby/

[6] https://dydra.com

Bosque-Gil, J. (2017). Linked data and dictionaries [Seminar]. 2nd Summer Datathon on Linguistic Linked Open Data. 27 June.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. & Montiel-Ponsoda, E. (2015). Applying the OntoLex model to a multilingual terminological resource. In *The semantic web: ESWC 2015 satellite events*. F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker & A. Zimmerman, Eds. 283-294.

Bouda, P. & Cysouw, M. (2012). Treating dictionaries as a linked-data corpus. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 15-24.

Cloud translation API: dynamically translate between thousands of available language pairs. n.d. Available: https://cloud.google.com/translate/ [2018, February 28].

De Rooij, S., Beek, W., Bloem, P., van Harmelen, F. & Schlobach, S. (2016). Are names meaningful? Quantifying social meaning on the semantic web. In *The Semantic Web: ISWC 2016.* P. Groth, E. Simperl, A. Gray, M. Sabou, M. Krötzsch, F. Lecue, F. Flöck & Y. Gil, Eds. 184-199.

Di Maio, P. (2015). Linked data beyond libraries. In *Linked data and user interaction.* H.F. Cervone, L.G. Svensson, Eds. Berlin: Walter de Gruyter GmbH. 3-18.

Doke, C.M. 1954. *The Southern Bantu languages.* London: International African Institute.

Dydra. (2011). Available: https://www.w3.org/2001/sw/wiki/Dydra [2018, January 3].

Eckart, K., Riester, A. & Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 65-76.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P. & Navigli, R. (2014). *Representing multilingual data as linked data: the case of Babelnet 2.0.* Available: http://wwwusers.di.uniroma1.it/~navigli/pubs/LREC_2014_Ehrmannetal.pdf [2017, December 27].

Faniel, I.M. & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. In *Curating research data: Practical strategies for your digital repository*. L.R. Johnston, Ed. Chicago, Association of College and Research Libraries. 103-126.

Final model specification. n.d. Available: https://www.w3.org/community/ontolex/wiki/Final_Model_Specification [2017, December 27].

Flati, T., Moro, A., Matteis, L., Navigli, R. & Velardi, P. (2015). *Guidelines for linguistic linked data generation: multilingual dictionaries (Babelnet).* Available: https://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/ [2017, December 27].

Gillis-Webber, F. (2018). The construction of an ontological framework for bilingual lexicographic resources: applying linguistic linked data principles. M.Phil. dissertation. University of Cape Town.

Gouws, R.H. & Prinsloo, D.J. (2005). *Principles and practice of South African lexicography.* Stellenbosch: SUN MeDIA.

Gracia, J. (2017). Introduction to linked data for language resources [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data. 26 June.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). *Toward linked data-native dictionaries.* Available: https://elex.link/elex2017/wp-content/uploads/2017/09/paper33.pdf [2018, January 17].

Gracia, J. & Vila-Suero, D. (2015). *Guidelines for linguistic linked data generation: bilingual dictionaries.* Available: https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/ [2017, December 25].

Gracia, J., Villega, M., Gómez-Pérez, A. & Bel, N. n.d. *The Apertium bilingual dictionaries on the web of data.* Available: http://www.semantic-web-journal.net/system/files/swj1419.pdf [2017, December 31].

Heath, T. & Bizer, C. (2011). *Linked data: evolving the web into a global data space.* Morgan & Claypool Publishers.

Herbert, R.K. & Bailey, R. (2002). The Bantu languages: sociohistorical perspectives. In *Language in South Africa.* R. Mesthrie, Ed. Cambridge: Cambridge University Press. 50-78.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A. & Decker, S. (2012). An empirical survey of linked data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web.* 14:14-44.

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the semantic web.* Morgan & Claypool Publishers. DOI: 10.2200/S00452ED1V01Y201210WBE003

Keller, M.A., Persons, J., Glaser, H. & Calter, M. (2011). *Report on the Stanford Linked Data Workshop.* Available: https://www.clir.org/wp-content/uploads/sites/6/LinkedDataWorkshop.pdf [2017, December 26].

Labra Gayo, J.E., Kontokostas, D. & Auer, S. n.d. *Multilingual linked data patterns.* Available: http://www.semantic-web-journal.net/system/files/swj495.pdf [2017, December 27].

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

McCrae, J.P. & Gracia, J. (2017). Introduction to the Ontolex-Lemon Model [Practical session]. 2nd Summer Datathon on Linguistic Linked Open Data. 26 June.

Ngcobo, M.N. Department of Linguistics and Modern Languages. (2010). *Only study guide for LIN3704: Language planning and language description.* Pretoria: University of South Africa.

Pretorius, L. (2014). The multilingual semantic web as virtual knowledge commons: the case of the under-resourced South African languages. In *Towards the multilingual semantic web.* P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 49-66.

PROV-Dictionary: Modeling provenance for dictionary data structures. (2013). Available: https://www.w3.org/TR/2013/NOTE-prov-dictionary-20130430/ [2018, January 1].

PROV-O: The PROV ontology. (2013). Available: https://www.w3.org/TR/prov-o/ [2018, January 1].

Rafferty, P. (2016). Managing, searching and finding digital cultural objects: putting it in context. In *Managing digital cultural objects: analysis, discovery and retrieval*. A. Foster & P. Rafferty, Eds. London: Facet Publishing. 3-24.

Sachs, J. & Finin, T. (2010). *What does it mean for a URI to resolve?*. Available: http://ebiquity.umbc.edu/_file_directory_/papers/495.pdf [2017, December 26].

Simons, N. & Richardson, J. (2013). *New content in digital repositories: the changing research landscape*. Oxford: Chandos Publishing.

Statistics South Africa. (2012). *Census 2011 Census in brief*. Pretoria: Statistics South Africa. Available: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf [2017, November 5].

Subfamily: Nguni (S.40). n.d. Available: http://glottolog.org/resource/languoid/id/ngun1276 [2018, February 11].

Tennis, J.T. (2007). Scheme versioning in the semantic web. In *Knitting the semantic web*. J. Greenberg & E. Méndez, Eds. 85-104.

Translating text. n.d. Available: https://cloud.google.com/translate/docs/translating-text [2018, January 5].

Van Erp, M. (2012). Reusing linguistic resources: Tasks and goals for a linked data approach. In *Linked Data in Linguistics*. C. Chiarcos, S. Nordhoff & S. Hellman, Eds. Heidelberg: Springer. 57-64.

Van Hooland, S. & Verborgh, R. (2014). *Linked data for libraries, archives and museums*. London: Facet Publishing.

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. & Aguado-de-Cea, G. (2014). Publishing linked data on the web: the multilingual dimension. In *Towards the Multilingual Semantic Web*. P. Buitelaar & P. Cimiano, Eds. Berlin: Springer-Verlag. 101-117.

Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O. & Gómez-Pérez, A. (2012). *Methodological guidelines for publishing government linked data*. Available: https://www.lri.fr/~hamdi/datalift/tuto_inspire_2012/Suggestedreadings/egovld.pdf [2017, December 25].

What is a language resource?. n.d. Available: http://www.elra.info/en/about/what-language-resource/ [2017, November 1].

Wood, D., Zaidman, M., Ruth, L. & Hausenblas, M. (2014). *Linked data: structured data on the web*. New York: Manning Publications Co.

Zgusta, L. (1971). *Manual of lexicography*. Prague: Academia, Publishing House of the Czechslovak Academy of Sciences.

A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis

Yalemisew Abgaz[1], Amelie Dorn[2], Barbara Piringer[2], Eveline Wandl-Vogt[2], Andy Way[1]
[1] Dublin City University, [2] Austrian Academy of Sciences
Dublin Ireland, Vienna, Austria
{Yalemisew.Abgaz, Andy.Way}@adaptcentre.ie, {Amelie.Dorn, Barbara.Piringer, Eveline.Wandl-Vogt}@oeaw.ac.at

Abstract

Around the world, there is a wide range of traditional data manually collected for different scientific purposes. A small portion of this data has been digitised, but much of it remains less usable due to a lack of rich semantic models to enable humans and machines to understand, interpret and use these data. This paper presents ongoing work to build a semantic model to enrich and publish traditional data collection questionnaires in particular, and the historical data collection of the Bavarian Dialects in Austria in general. The use of cultural and linguistic concepts identified in the questionnaire questions allow for cultural exploration of the non-standard data (answers) of the collection. The approach focuses on capturing the semantics of the questionnaires dataset using domain analysis and schema analysis. This involves analysing the overall data collection process (domain analysis) and analysing the various schema used at different stages (schema analysis). By starting with modelling the data collection method, the focus is placed on the questionnaires as a gateway to understanding, interlinking and publishing the datasets. A model that describes the semantic structure of the main entities such as questionnaires, questions, answers and their relationships is presented.

Keywords: Ontology, E-lexicography, Semantic uplift

1. Introduction

There is a substantial amount of traditional data available on the internet and intranets of organisations. Traditional data, in this paper, refers to historical, socio-cultural, political, lexicographic and lexical data sets that are collected over an extended period. Public organisations such as museums, national bibliographic centres and libraries are increasingly opening their doors to facilitate access to such data to support research and development beyond their organisational boundaries (Doerr, 2009). This trend enables researchers to access a significant amount of useful primary data of historical, temporal and societal importance (Kansa et al., 2010; Beretta et al., 2014; Meroño-Peñuela et al., 2015). Making these data available, both for humans and machines, however, comes with several shortcomings.

First, in the majority of cases, these traditional data are initially available in bulk of archival formats, providing only a general description of the content of the data. However, they fail to provide detailed information about why, how, when and who collected the data and how the data can be interpreted and used. Often, consumers of such data require additional contextual information to understand and interpret the information contained in the datasets correctly. This is undoubtedly undesirable as it requires a considerable effort to understand and utilise the dataset.

Second, no matter how big and valuable a released dataset is, it is virtually impossible for machines to use the data without proper semantics for interpreting its content. As machines are becoming ever more typical consumers of such datasets, it has become crucial to include standardised machine-readable semantics in addition to the data itself. The limited availability of semantics to describe the data is, therefore, one of the leading obstacles for machines discovering and interpreting legacy data.

Third, interlinking of the data with other available datasets becomes difficult. The lack of semantics, the use of non-standard vocabulary or the absence of schema mapping (Bizer, Heath, & Berners-Lee, 2009) are some of the causes. Traditional data that includes a schema definition or a data dictionary provides useful information to aid the process of speedy utilisation, but often lacks the information about the means of interlinking the data with existing datasets especially with those available on the linked open data (LOD) platform. The interlinking of the data using a data dictionary further requires a mapping from the data dictionary to a standard vocabulary. This not only requires domain knowledge, but also a detailed knowledge of the internal structure of the data.

In this paper, we focus on a historical data collection of the Bavarian Dialects covering almost a century old data (1911-1998) from the present-day Austria. For effective opening up and utilisation of the collection, we present our approach to facilitating the semantic modelling, enrichment and publishing of traditional data, taking the data collection questionnaires and their individual questions as the starting point. The questionnaires and questions are essential parts of the entire collection as they serve as an entry point to access the answers, where typically neither the headword nor the definition are noted as standard terms. The use of linguistic and cultural concepts in the model thus allows for the exploration and exploitation of cultural links, which is one of the main aims of the exploreAT! project. The questionnaires of the "Datenbank der bairischen Mundarten in Österreich (DBÖ/dbo@ema)" within the project exploreAT! (Wandl-Vogt, 2012) is used as a case study to demonstrate the process. The approach is composed of major steps such as domain analysis, schema analysis, semantic model and semantic up-lift. Domain analysis includes the understanding of the rationale of the data collection, the method of data collection, the original documents used, primary agents that produced the data collection methods and those agents who collected the data. By employing this step, it is possible to collect significant semantics that describes the collection. Schema analysis of the dataset at various stages is also a crucial step, which includes a closer inquiry of the structure of the data, the relationship between entities and their attributes and investigation of any inconsistencies and anomalies. The semantic modelling

step focuses on representing the structure and the semantics of the entities in the datasets using a well-defined semantic model. It is another essential step especially for domains that lack a suitable vocabulary to describe entities fully. In the absence of such vocabulary, it becomes crucial to build a semantic model of the domain from scratch. Finally, the semantic model is used to up-lift, interlink and integrate the data with other related datasets. It will serve as a means to open up valuable traditional data to support further research and possibly answer various questions involving the evolution of conceptualisations of societies in the past and the present.

This approach enables organisations to make their datasets not only digitally available but also semantically enrich the dataset to facilitate a common understanding, interpretation and consumption by both machines and humans. The focus of this paper is, thus, to present our approach and the resulting semantic model. Even if the overall semantic model covers various aspects of the data, at this stage, it will focus only on modelling the questionnaires and questions, which provides users with a unique perspective of accessing the data, looking at it from the original questions and navigating to the corresponding answers, collectors or entities of interest. The model will further facilitate conceptual interoperability (Chiarcos et al., 2013) with other LOD repositories.

This paper is structured in the following way: Section 2 sheds light on the domain and describes the nature of the datasets in use. Section 3 presents the approach including domain and schema analysis and Section 4 discusses the core semantic model using the exploreAT! case study of Bavarian Dialects. In Section 5, we present ongoing work to utilise the semantic model towards the publishing of the datasets using LOD principles. Finally, the conclusion and future work are discussed in Section 6.

2. Background

2.1 Database of Bavarian Dialects (DBÖ)

The database of Bavarian Dialects (*Datenbank der bairischen Mundarten in Österreich -DBÖ) [Database of Bavarian Dialects in Austria]* (Wandl-Vogt, 2008) is a historical non-standard language resource. It was originally collected in the Habsburg monarchy with the aim of

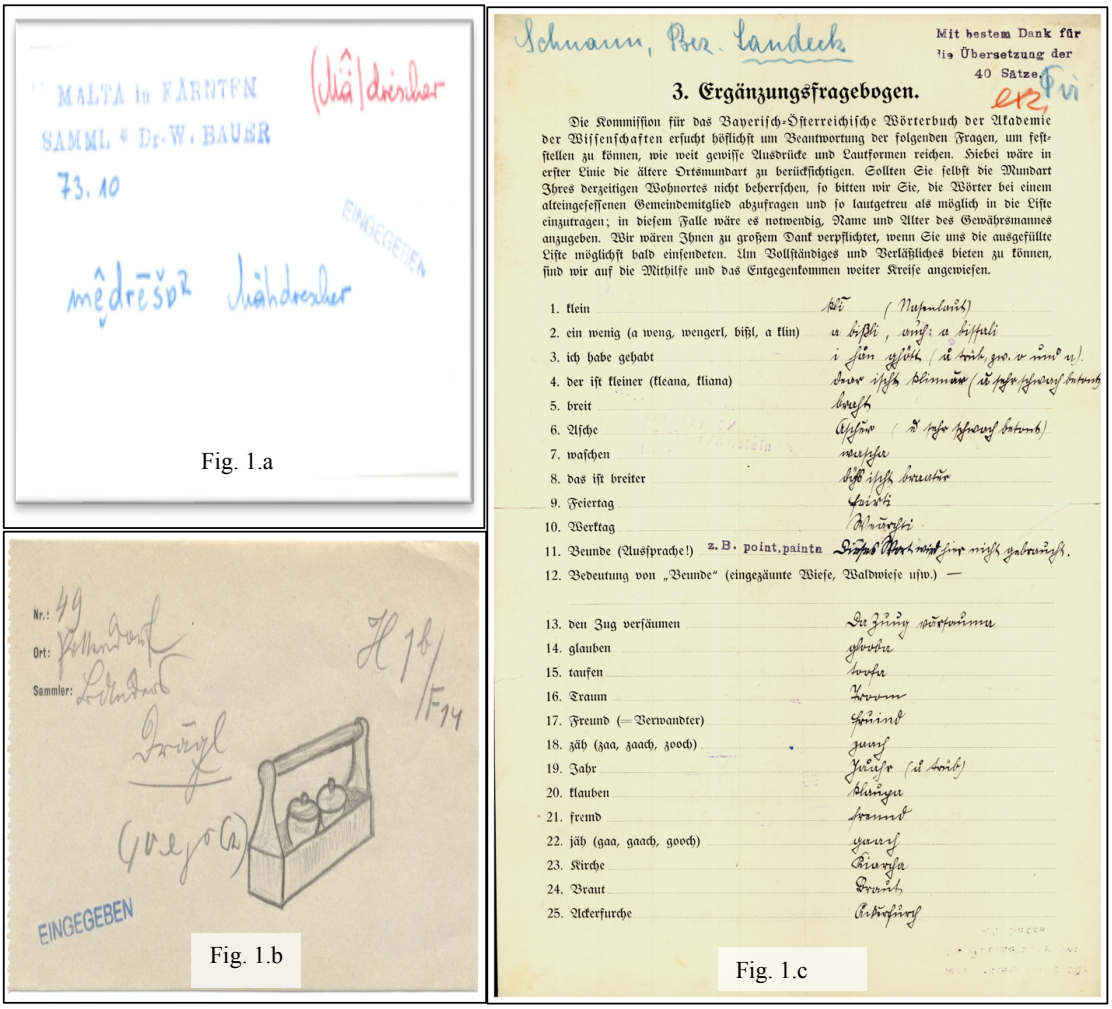

Fig. 1.a

Fig. 1.b

Fig. 1.c

Figure 1. Sample paper slips (a, b) and sample filled questionnaire((Ergänzungsfragebogen) (c).

documenting the German language and rural life in Austria from the beginnings of the Bavarian dialect to the current day. The inception of the data collection went back to 1913 and continued until 1998 in present-day Austria, Czech Republic, Slovakia, Hungary and northern Italy, leaving a century-old historical, socio-cultural and lexical data resource. Even if the original aim of the collection was to compile a dictionary and a linguistic atlas of Bavarian dialects (Arbeitsplan, 1912) spoken by the locals, the data includes various socio-cultural aspects of the day-to-day life of the inhabitants, such as traditional customs and beliefs, religious festivities, professions, food and beverages, traditional medicine, and many more (Wandl-Vogt, 2008)

The data was collected using 109 main questionnaires, nine additional questionnaires (*Ergänzungsfragebögen*) and two *Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen questionnaires* and other additional freestyle questionnaires and text excerpts from various sources such as vernacular dictionaries and literature. In total, there are 24,382 individual questions corresponding to the available questionnaires in the collection. In response to the questionnaires over the span of the project, several million (~ 3.6 million) of individual answers noted on paper slips (Fig. 1.a, b) were collected. The answers to the questions include single words, pronunciations, illustrations and explanations of cultural activities on topics such as traditional celebrations, games, plays, dances, food and other topics.

In addition to the primary data, the entire collection also includes biographies of individual collectors and contributors of various roles. 11,157 individuals who had various functions in the project had participated in the data collection process as authors of the questionnaires, data collectors, editors or coordinators, with some having several of these functions at once. Detailed information about the personal background of individual contributors which was also noted in the course of data collection and during the digitisation process in later years is stored in a specific database (*Personendatenbank [person database]*). Persons and their background are thus other important features of the data that offer additional points for the exploration and the systematic opening of the collection.

The data set further contains additional information about the geographic locations and names of places including cities, districts and regions related to the places where the questionnaires were distributed. In rare cases, the paper slips include information about the time of the data collection.

The collected data has been used to produce a dictionary, *Wörterbuch der bairischen Mundarten in Österreich [Dictionary of Bavarian Dialects in Austria]* (WBÖ); up to now five volumes (A–E, P and T) have been published. Today, about three-quarters of the collected paper slips are available in a digital format following several stages of digitisation. The available formats corresponding to the stages include scanned copies of the paper slips, a textual representation of the paper slips in TUSTEP, MySQL (Barabas et al., 2010) and TEI/XML (Schopper, 2015). This is an ongoing effort to make the data accessible and analyse them, including the use of semantic web technologies to make the data suitable for semantic publishing in the LOD platform.

3. Approach

There is an increasing focus on semantic publishing of traditional data using LOD platforms. To support this, different approaches are used to enrich and expose the data stored in legacy databases semantically. One such approach, direct conversion, converts structured databases (usually relational databases and XML files) directly to RDF triples (Berners-Lee, 1998). This approach mainly uses the schema of the legacy system to transform the data. The transformed data, usually in a triple format (subject, predicate, object), is published as a separate service to the legacy data or as a new layer on top of the legacy database. This approach allows a mass conversion of legacy data without the need for analysis beyond the available schema. However, one of the drawbacks of this approach is that it is restricted to the semantics available within the data and adds little semantics other than the one contained in the schema (Simpson & Brown, 2013). This approach is mainly applicable for general collections but requires a detailed analysis when the domain of interest becomes specialised.

The alternative to this approach focuses on the analysis of the domain of interest and generate/select one or more ontologies that describe the semantics of entities and their relationships. This approach is more rigorous in that experts define the semantics of each entity and its properties. Besides, it facilitates inclusion of the domain knowledge of the experts and opens up a way of accommodating entities that are relevant to the domain but not included in the dataset. The downside of this approach is that it requires a certain level of domain-expert involvement and may require more effort and expert agreement. However, this approach provides a robust semantics and significantly contributes to interoperability.

In our work, we merge the two approaches and use schema analysis to identify entities, attributes and their relationships and domain analysis to analyse and describe the domain and to understand the rationale of the data collection method.

3.1. Schema Analysis

The availability of the dataset in various formats motivates us to look into schema analysis. The questionnaires are available as analogue paper copies, flat text files, in TEI/XML format and a relational table format (dbo@ema). The schema analysis of the available datasets provides us with valuable information to build our semantic model. Research (Ferdinand, C. Zirpins, & D. Trastour, 2004; Deursen et al., 2008; Battle, 2006) has shown that schema analysis provides significant information. The quality of the resulting semantic data, however, depends on the completeness and expressiveness of the available schema and does not reflect the meanings of the entities. In many cases, even if the structural information is available, accurate interpretation of the meaning conveyed by a given schema and its mapping to a standard vocabulary is difficult to achieve. For example, a relational schema

which stores the year as "Year" requires accurate interpretation of whether the attribute "Year" refers to the year of publication of the questionnaire or the year it is distributed to data collectors or any other interpretation. Additionally, it requires an accurate description to resolve if "year" can be considered the same as "dcterms:date". Despite these drawbacks, schema analysis plays a significant role in identifying entities, attributes and their relationships.

```
*LT1* h-at~-in<<e [pl] *ANMO* Pl. ?
*LT2* h-at~-in [m,sg]
*BD/LT1* Abteilung für Heu
**********
*A* HK 157, d157^#3.1 = T1570317.sch^#3, korr. W.B.
*HL* (Ge—sott)tin:1
*QU* Matrei OTir. ob. lselt., Aufn. Gabriel
*QDB* {1B.0f01} oblselt.:lselt.:lselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* ks-O((ut~-in
*BD/LT1* Abteilung für Gesott
**********
*A* HK 157, d157^#4.1 = T1570317.sch^#5, korr. W.B.
*HL* Stadel:1
*QU* Matrei OTir. ob.lselt.
*QDB* {1B.0f01} oblselt.:lselt.:lselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* >st-..odl
**********
```

Figure 2. TUSTEP format

Schema description: through the life of the dataset, various software tools have been used to store and process the data. Currently, the software includes TUSTEP (Fig. 2), XML/TEI (Fig. 3) and MySQL (Fig. 4). Each of these tools keeps some schema of their own to describe the contents of the files. Having studied all these formats to understand the schema, we used the relational database schema as our

```
<entry xml:id="d157_qdb-d1e2" xml:lang="bar">
   <form type="hauptlemma">
      <orth>Tin</orth>
   </form>
   <gramGrp>
      <pos>Subst</pos>
   </gramGrp>
   <form type="lautung" n="1">
      <pron notation="tustep">|A t~-in</pron>
      <pron notation="ipa" resp="#JB" change="01">|A t~-in</pron>
      <gramGrp>
         <gram> [m,sg+U]</gram>
      </gramGrp>
   </form>
   <form type="lautung" n="2">
      <pron notation="tustep">t~-in;e</pron>
      <pron notation="ipa" resp="#JB" change="01">t~-in;e</pron>
   </form>
   <sense corresp="this:LT1">
      <def xml:lang="de">Abteilung für Heu</def>
   </sense>
   <cit type="kontext" n="1">
      <quote>in den t~-;in [m,sg4] hinein</quote>
      <quote resp="#JB" change="01">in den t~-;in hinein</quote>
   </cit>
```

Figure 3. XML/TEI format

main source containing 88 relational tables. In this paper, our focus is on the schema which is directly related to questionnaires (4), questions (2), authors (n=7) and answers (n=7).

From the schema analysis, entities such as questionnaire (Fragebogen), types of questionnaires, questions (Frage), answers and authors are identified. Attributes of these entities and their data types are also identified.

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increment
nummer	varchar(28)	NO	UNI	NULL	
titel	varchar(512)	NO		NULL	
schlagwoerter	varchar(1024)	YES		NULL	
erscheinungsjahr	int(11)	YES		NULL	
person_id	int(11)	YES		NULL	
originaldaten	text	YES		NULL	
anmerkung	text	YES		NULL	
freigabe	tinyint(1)	NO		NULL	
checked	tinyint(1)	NO		NULL	
wordleiste	tinyint(1)	NO		NULL	
druck	tinyint(1)	NO		NULL	
online	tinyint(1)	NO		NULL	
publiziert	tinyint(1)	NO		NULL	
fragebogen_typ_id	int(11)	YES	MUL	NULL	

Figure 4. MySQL format

Each attribute of the entities is examined for the relevance of the conveyed information in addition to the availability of usable data. There are attributes that contain null values for all records and columns with redundant information. For example, the attribute "wordleiste" ("MS Word Bar") in Fig. 4 contains empty values across all the records in the table. Such attributes are identified and presented to the domain experts for further analysis. There are also attributes that contain null values for some of the records and are left as they are, as there are possibilities to populate them from other sources. Expert evaluation categorised these attributes as "relevant", "needs further investigation" and "not relevant". We included the first two categories but discarded the "not relevant" ones. Finally, the entities and attributes are used as an input for preparing the semantic model.

3.2. Domain Analysis

Domain analysis serves as another step for understanding the rationale of the data collection and the data collection process itself. It provides a solid foundation about why, how, when and by whom the data was collected, stored and processed. It further provides a solid base for understanding the core entities of the datasets, the relationship among the entities and across other entities of similar purpose. Our approach starts with the study of primary sources of information, investigating and examining original materials, interviewing users and maintainers of the dataset. It also includes secondary sources to complement and clarify the domain knowledge.

Following the approach used by Boyce & Pahl (2007), the domain analysis stage seeks information related to 1) Purpose - the rationale of the data collection, 2) Source - the data collection method used, 3) Domain - the nature of the collected data, and 4) Scope - what are the core entities of interest.

Purpose: The purpose of the data collection is to document the wealth of diversity of rural life and unite it under a Pan-European umbrella with a special focus on German language and diverse nationalities in the late Austro-Hungarian Monarchy (Gura, Piringer, & Wandl-Vogt,

forthcoming). The rationale of the data collection serves as a guidance for tuning our objectives and achieving the results. Thus, accordingly, our long-term interest is to capture the lexical data, represent it using standard vocabularies and interlink it with other collections.

Source: The primary data is collected using questionnaires with one or more questions. Questionnaires were distributed to the collectors, and the collectors filled the questionnaires by asking individuals and groups. In some cases, the collectors filled out the questionnaires themselves after observing teams of respondents. Then, collectors sent out the completed questionnaires to the centre where the data was further processed. The questions could be completed by one respondent or a group of respondents. In other cases, questions were filled by the data collectors themselves. Paper slips containing answers arrived at the centre even after several years and are stored in drawers alphabetically.

An interesting aspect of the domain analysis is the identification of the different question types which are not mentioned in any of the available schemas. A closer look at the questions resulted in the identification of patterns of questions used. The data collection is systematic in that it associates certain abbreviations to the questions that have asked similar types of questions. For example, phonological questions have abbreviations such as "*Aussprache, Ausspr.* or *Ltg.*" morphological questions have "*Komp.*" and synonym questions have "*Syn.* or *Synonym*" patterns. However, not all the questions have such abbreviations. The question types and their definitions are represented in detail in the next section. As the questions are linked to the answers, it is also possible to identify the different types of answers provided for a given question. The identification of question types by the domain experts will play a significant role for question-answering systems by exploiting these categories. However, modelling the answers is beyond the scope of this paper.

Domain: The primary data collected is lexical data in direct response to the questions of the questionnaire. It covers various aspects such as names, definitions, pronunciations, illustrations and other categories targeting a linguistic atlas and dictionary compilation (Arbeitsplan, 1912). However, there are other data generated during the process, including details of data collectors, the time and place of the data collection. Regarding the domain, the main interest is the linguistic data of historical and cultural importance.

Scope: From the above steps, we already identified the core entities contained in the datasets. These entities are defined and described by experts. The focus of this exercise is to use the questionnaires as the main entry point to semantically explore the data. Questionnaires contain individual questions of a particular topic which are linked to individual answers. However, in this paper, we will mainly focus on modelling questionnaires and the questions and explore obvious links to answers, authors, collectors and geographic locations. By doing so, we provide additional information which is relevant to answer research questions regarding gender-symmetry or spatiotemporal distributions. However, modelling the answers is complex and will not be discussed in detail in this paper. A pilot for modelling geographic locations is developed and treated separately (Scholz et al., 2016; Scholz, Hrastnig, & Wandl-Vogt, 2018).

4. Semantic Modelling

As a means of semantically enriching the datasets to publish it as a LOD, a semantic model was developed that incorporated the questionnaire model (Fig. 5) and the question model (Fig. 6) with a link to the associated entities. Both models are ontological models built using the Web Ontology Language (OWL2)[1] specification following ontological principles (Noy & Mcguinness, 2001; Edgar & Alexei, 2014). These models provide:

- A succinct definition of the entities and their relations,
- Interoperability with existing semantic resources to support LOD, and
- Extensibility to introduce new classes and relations.

There are many ontologies available to describe data of interest. These ontologies range from general purpose upper ontologies to lower, domain-specific ontologies to describe fine-grained knowledge for describing historical and cultural domains. After deciding the domain and the scope, the next step in the modelling stage is to consider reusing existing ontologies as this is preferable to developing an in-house ontology. However, for domain-specific description of datasets, it is difficult to find a suitable ontology and thus requires preparation either from scratch or extending existing ones.

We searched existing ontologies that can describe our domain of interest. The main repositories searched include LOV[2] ontology repository, Schema.org[3] and other specialised search tools such as Watson semantic web search engine.[4] We found terminologies related to questions, answers and questionnaires, but they do not fit our requirements, and such ontologies are not available yet. However, we will exploit some of the concepts defined in the Ontolex-Lemon model (McCrae et al., 2017) to describe the lexical data in the collection. We will further reuse vocabularies such as FOAF, SKOS and Dublin Core to describe authors, editors, collectors, places and publication. In addition to describing the entities, generic ontological constructs are used to create an interlinking with concepts from other repositories, and to compare our data with other similar data sets using meaningful interoperability.

A combination of top-down and bottom-up approaches as proposed by Uschold & Gruninger, (1996) is used to develop the model. The approach integrated domain analysis as a top-down approach and schema analysis as a bottom-up approach to build the ontology, in order to support our domain-specific requirements. We also used existing standardised vocabularies for entities that already have compatible representations. We developed our

[1] https://www.w3.org/TR/owl2-primer/
[2] http://lov.okfn.org
[3] http://schema.org
[4] http://watson.kmi.open.ac.uk/WatsonWUI/

ontology to represent both the structure and the meaning of the entities of interest.

4.1. Questionnaire Model

The questionnaire model is built based on the detailed analysis of the original and physically compiled book of sets of questionnaires and its electronic version (dbo@ema). Up to now, we have identified three questionnaire types. Each type has its characteristics and differs from the others in its purpose, the type of information it seeks and its format, including its physical appearance. Treating the different sets of questionnaires independently is crucial to preserve the historical importance and the structural and semantic relation each questionnaire set has with the collected data. The questionnaire types are discussed below:

1. Systematic: [*Systematischer Fragebogen*] is a questionnaire that is used to collect the original data. This type of questionnaire is used from the beginning of the data collection process.
2. Additional: [*Ergänzungsfragebogen*] is a questionnaire that is used as a supplementary questionnaire to the systematic questionnaire.
3. Dialectographic *[Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen] is a* questionnaire of the Munich and Vienna Dictionary Commissions.

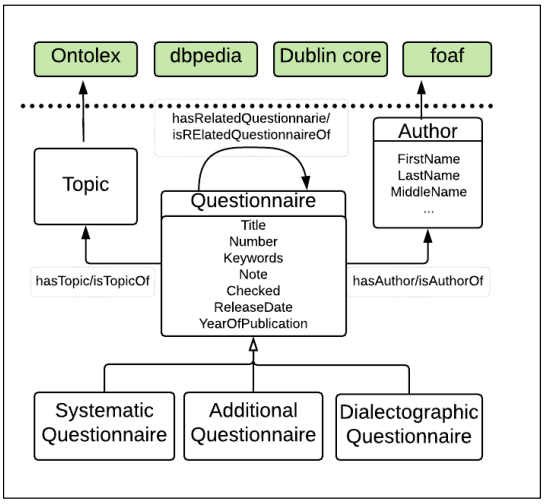

Figure 5. A Semantic model of questionnaire

A questionnaire may have one or more related questionnaires that deal with the same topic. We observed that questionnaires refer to other questionnaires. Such relationships are captured by an object property "hasRelatedQuesitonnaire" with an inverse property "isRelatedQuestionnaireOf". A questionnaire has at least one topic, and this relationship is captured by "hasTopic" with "isTopicOf" inverse object property. Furthermore, a questionnaire has at least one Author, and this relationship is captured by "hasAuthor" object property with "authorOf" inverse object property.

Topics (Questionnaire Topics). A topic is the main subject of the questionnaire or a given question. A questionnaire may focus on a general topic such as "Food" and a question may cover subtopics such as "Traditional Food". This information will be treated as a topic following a proper disambiguation technique and then relate to ontolex:lexicalConcept.

Author/collectors. Authors are defined in FOAF and Dublin Core. We will reuse the definition provided in FOAF Agent/Author classes.

4.2. Question Model

A question is a linguistic expression used to request information, or the request made using such an expression. The information requested is provided in the form of answer.[5] In this ontology, we categorise the questions mainly based on the content, the forms and the expected answers from the respondents. An analysis carried out by the experts, users and ontology engineers identified 12 different types of questions and added two more questions to accommodate future processing of additional questionnaire sets. It is important to note that these question types are not mutually exclusive to one another and there are instances of questions that belong to more than one type of questions, e.g. the question "*Kopf: Kopf/Haupt (in urspr. Bed.) in Vergl./Ra. (Kopf stehn, der Kopf mI¦chte einem zerspringen)*" is both semasiological and syntactic. The semantics of the question types are given below:

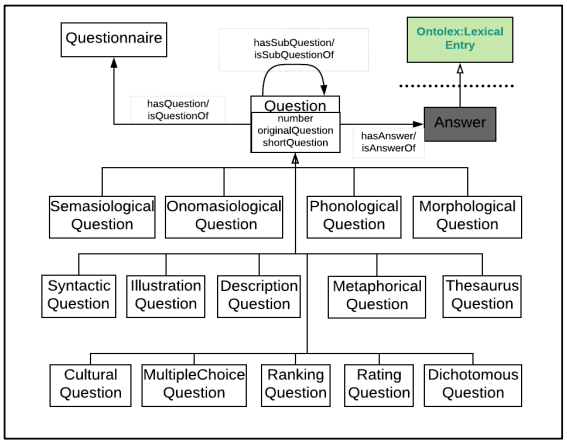

Figure 6. A Semantic model of question

1. *Onomasiological:* asks for the name of a given entity, e.g. "how do you call x?" where x represents an entity.
2. *Semasiological:* asks for the meaning of a given entity, e.g. "what does x mean?".
3. *Dichotomous:* asks for a selection of answers from a binary option. It includes yes/no or agree/disagree types of answers to stated questions.
4. *Description:* asks for a written representation of a given entity, e.g. "What would be the function of x?".

[5] https://en.wikipedia.org/wiki/Question

5. *Illustration:* asks for a pictorial or diagrammatic representation of a given entity, e.g. "What does x look like?".
6. *Morphological:* asks about the structure and the formation of words and parts of words. Based on the structure, morphological questions can take various forms.
7. *Phonological:* asks for the pronunciation, or phonetic representation of words.
8. *Syntactic:* asks for construction of phrases or sentences using a given word or a given idiom, e.g. "Provide a phrase/sentence for/using a word/idiom x".
9. *Metaphorical:* asks for some conveyed meanings given a word or an expression. Metaphorical questions are related to semasiological questions, but they ask for an additional interpretation of the expression beyond its obvious meaning.
10. *Thesaurus:* asks for a list of words or expressions that are used as synonyms (sometimes, antonyms) or contrasts of a given entity.
11. *Cultural:* asks for a belief of societies, procedures on how to make or prepare things and how to play games, contents of cultural songs, poems used for celebrations. Analysis of the existing questions shows that the cultural question type has its subtypes and has instances that significantly overlap with the other question types.
12. M*ultiple Choice:* asks for a selection of one item from a list of three or more potential answers.
13. *Rating:* asks the respondent to assign a rate (degree of excellence) to a given entity based on a predefined range
14. *Ranking Question:* asks the respondent to compare entities and rank them in a certain order.

It is commonly observed that a question may ask several other sub-questions, and this is captured by the "hasSubQuestion" object property. Thus, the object property "hasSubQuestion" relates one question with its subquestions. Each question is linked to its associated answer. A question may have several answers collected from different sources. This is captured by the "hasAnswer" object property with its inverse "isAnswerOf". Finally, a question is related to a questionnaire with the "isQuestionOf" object property where a single question is contained only in one questionnaire.

Answer: An answer is a written, spoken or illustrated response to a question. The different types of questions have answers either in a written, spoken or illustration format. In the case of questions that involve lexical data collection, the answer could be associated with some lexical category. For each types of questions, there are different types of answers including sentences, individual words, multiword expressions, affix, diagrams, etc. Modelling the answers is under investigation. However we will treat answers with single word, multiword expression or affixes as ontolex:lexicalEntries. For example, the answer to a thesaurus question is expected to be a word, or multiword expression in the OntoLex model.

Finally, an initial version of the ontology- Ontology for Lexical Data Collection and ANalysis (OLDCAN)[6] is developed following the approach discussed above. Since the project is at its development stage, a permanent URL has not been yet assigned to either the ontology or to the data. However, the ongoing results are available under a Creative Commons Licensing.[7]

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia: <http://dbpedia.org/ontology/> .
@prefix oldcan: <http://localhost/oldcan/OLDCAN#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<#QuestionnaireTriplesMap>
rr:logicalTable [ rr:sqlQuery """

  SELECT Fragebogen_V1.*, (CASE QUESTIONNAIRE_TYP_ID
    WHEN '1' THEN 'SystematicQuestionnaire'
    WHEN '2' THEN 'AdditionalQuestionnaire'
    WHEN '3' THEN 'DialectographicQuestionnaire'
  END) QUESTIONNAIRETYPE FROM Fragebogen_V1
  """ ];
    rr:subjectMap [
        rr:template "http://localhost/dboe/Questionnaire/{ID}" ;
        rr:class oldcan:Questionnaire ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate rdf:type ;
        rr:objectMap [ rr:template "http://localhost/oldcan/OLDCA#
        {QUESTIONNAIRETYPE}";
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:title ;
        rr:objectMap [ rr:column "TITLE" ;rr:language "de";] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:publicationYear ;
        rr:objectMap [ rr:column "YEAR_OF_PUBLICATION" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    rr:predicateObjectMap [
        rr:predicate oldcan:note ;
        rr:objectMap [ rr:column "NOTE" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    ];
    .
```

Figure 7. R2RML mapping excerpts

5. Semantic Up-lift

This stage focuses on the use of the semantic model and selected vocabularies to semantically enrich the data. It is used to annotate every data element with semantic information that states what it is, how it should be interpreted and how it is related to other elements within the datasets or across other datasets. There are various methods and tools used to transform relational databases to semantically compatible formats including direct mapping (Berners-Lee, 1998) and domain semantics-driven mapping (Michel, Montagnat, & Faron, 2013). We followed R2RML[8] to annotate our datasets due to its customisability for mapping relational databases into triples. Unlike direct mapping that depends on the database's structure, it is possible to use an ontology of the domain. Since R2RML is a vocabulary by itself, it stores

[6] http://exploreat.adaptcentre.ie/#Semantics
[7] https://creativecommons.org/licenses/by/3.0/at/deed.en
[8] https://www.w3.org/TR/r2rml/

the mappings from a relational database to RDF as RDF files and allows inclusion of provenance information. This facilitates knowledge discovery and reuse of mappings. However, it requires more effort compared to direct mapping. R2RML is used to map the relational data into a LOD. This phase includes the following steps:

1. Converting the major tables into classes,
2. Mapping object property relationships,
3. Mapping data property relationships,
4. Enriching the data with additional semantics.

To demonstrate the envisioned mapping, excerpts of the mapping file for both questionnaire and questions are generated. In the mapping (Fig. 8), each questionnaire is associated to oldcan:Questionnaire class using "a" ("rdf:type") property. The template defines the URL of the specific location of the questionnaire. The selected attributes are mapped to data properties, e.g. title is mapped to oldcan:title and the language of the title is included using a language tag "de".

The mapping of the questions is done similarly. Here the object property isQuestionOf is used to link the question with its questionnaire. In the ontology, the hasQuestion object property is defined as an inverse of isQuestionOf to achieve both brevity and searchability in the generated data. The different types of the questionnaires and the questions are captured. An excerpt of the resulting triple[9] is presented in Fig. 8.

```
<http://localhost/dboe/Questionnaire/1>
    a    <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
         <http://localhost/oldcan/OLDCAN#Questionnaire> ;
    <http://localhost/oldcan/OLDCAN#note>
         "resfb1" ;
    <http://localhost/oldcan/OLDCAN#publicationYear>
         "1920" ;
    <http://localhost/oldcan/OLDCAN#title>
         "Kopf (1)"@de .

<http://localhost/dboe/Questionnaire/2>
    a    <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
         <http://localhost/oldcan/OLDCAN#Questionnaire> ;
    <http://localhost/oldcan/OLDCAN#note>
         "bafb2" ;
    <http://localhost/oldcan/OLDCAN#publicationYear>
         "1920" ;
    <http://localhost/oldcan/OLDCAN#title>
         "Die Osterwoche (1)"@de .
.....
<http://localhost/dboe/Question/1-A11>
    a    <http://localhost/oldcan/OLDCAN#Question> ;
    <http://localhost/oldcan/OLDCAN#isQuestionOf>
         <http://localhost/dboe/Questionnaire/1> ;
    <http://localhost/oldcan/OLDCAN#number>
         "A11" ;
    <http://localhost/oldcan/OLDCAN#originalQuestion>
         "Kopf: breiter Kopf"@de ;
    <http://localhost/oldcan/OLDCAN#shortQuestion>
         "breiter Kopf"@de .

<http://localhost/dboe/Question/111-2>
    a    <http://localhost/oldcan/OLDCAN#Question> ;
    <http://localhost/oldcan/OLDCAN#isQuestionOf>
         <http://localhost/dboe/Questionnaire/111> ;
    <http://localhost/oldcan/OLDCAN#number>
         "2" ;
    <http://localhost/oldcan/OLDCAN#originalQuestion>
         "Altane im 1. Stock (Şller, Schrot, Laube, Br_ckel)"@de ;
    <http://localhost/oldcan/OLDCAN#shortQuestion>
         "Altane im 1.Stock (Şller, Schrot, Laube,*)"@de .
```

Figure 8. Questionnaire and question triples

6. Conclusion and Future Work

The effort to open up legacy databases to make them accessible, usable and researchable has increased with the development of LOD platforms. Such platforms facilitate publishing legacy data of a wide range of contents and formats. As the content becomes specialised, the need for finding and developing semantic models that describe the domain of interest become crucial. This paper has presented an approach which is currently used for building a semantic model for enriching and publishing traditional data of historical, cultural and lexical importance. It is argued that the use of such an approach for building semantic models to assist with semantic publishing of traditional data on the LOD platform is vital to the exploitation of data of historical importance. It further paves the way for researchers to understand and compare conceptualisation of entities at different times and their evolution through time. As the paper presents work in progress, our immediate focus is the enrichment of the semantic model by in-depth examination of the entities including answers to the questions to enable a strong semantic interlinking that will facilitate efficient question answering and comparison of the different types of questions. Furthermore, additional enrichment to interlink the data with other similar datasets and the visualisation of the dataset will be the next area to tackle.

7. Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! project and the ADAPT Centre for Digital Content Technology at Dublin City University which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

[9] http://exploreat.adaptcentre.ie/#APIs

Bibliographical References

Arbeitsplan (1912). *Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912.* Karton 1. Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch. Archive of the Austrian Academy of Sciences. Wien.

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. & Schwaiger, S. (2010). Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In H. Bergmann, M. M. Glauninger, E. Wandl-Vogt, & S. Winterstein (Eds.), *Fokus Dialekt. Analysieren – Dokumebattentieren – Kommunizieren. Festschrift für Ingeborg Geyer zum 60. Geburtstag.* (Germanistische Linguistik 199–201). Hildesheim, Zürich, New York: Olms, (pp. 47–64).

Battle, S. (2006). Gloze: XML to RDF and back again. *First Jena User Conference.* Bristol, UK.

Beretta, F., Ferhod, D., Gedzelman, S. & Vernus, P. (2014). The SyMoGIH project : publishing and sharing historical

data on the semantic web. *Digital Humanities 2014, July 2014*, Lausanne, Switzerland. (pp. 469–470).

Berners-Lee, T. (1998). *Relational Databases on the Semantic Web.* In *Design Issues for the World Wide Web.* Retrieved January 10, 2018, from https://www.w3.org/DesignIssues/RDB-RDF.html

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web Information Systems, 5*(3), 1–22.

Boyce, S. & Pahl, C. (2007). Developing Domain Ontologies for Course Content. *Educational Technology & Society, 10*, 275–288.

Chiarcos, C., Cimiano, P., Declerck, T. & McCrae, J. P. (2013). Linguistic Linked Open Data (LLOD) – Introduction and Overview. In C. Chiarcos, P. Cimiano, T. Declerck, & J. P. McCrae (Eds.), *2nd Workshop on Linked Data in Linguistics.* Pisa, (pp. i–xi).

Doerr, M. (2009). Ontologies for Cultural Heritage. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies. International Handbooks on Information Systems.* Berlin, Heidelberg: Springer.

Ferdinand, M., Zirpins, C. & Trastour, D. (2004). Lifting XML Schema to OWL. In N. Koch, P. Fraternali, & M. Wirsing (Eds.), *Web Engineering. 4th International Conference, ICWE 2004. Munich, Germany, July 26-30, 2004. Proceedings.* (Lecture Notes in Computer Sciences 3140). Berlin, Heidelberg: Springer, (pp. 354–358).

Gura, C., Piringer, B. & Wandl-Vogt, E. (forthcoming). Nation Building durch Großlandschaftswörterbücher. Das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) als identitätsstiftender Faktor des österreichischen Bewusstseins.

Kansa, E. C., Kansa, S. W., Burton, M. M., & Stankowski, C. (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies, 6*(2), 301–326.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Leiden: Lexical Computing CZ, (pp. 587–597).

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. & Van Harmelen, F. (2015). Semantic Technologies for Historical Research: A Survey. *Semantic Web, 6*(6), 539–564.

Michel, F., Montagnat, J., & Faron Zucker, C. (2013). *A survey of RDB to RDF translation approaches and tools.* Retrieved January 16, 2018, from https://hal.archives-ouvertes.fr/hal-00903568v1

Noy, N. F., & McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology.* Technical, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Scholz, J., Hrastnig, E., & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In P. Fogliaroni, A. Ballatore & E. Clementini (Eds.), *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017).* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 275–282).

Scholz, J., Lampoltshammer, T. J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Indeterminate and Crisp Boundaries. In G. Gartner, M. Jobst, & H. Huang (Eds.), *Progress in Cartography.* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 133–151).

Schopper, D., Bowers, J. & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a data-base of dialects into a rich LOD resource. Retrieved January 17, 2018 from *Text Encoding Initiative Conference and members' meeting 2015, October 28-31, Lyon, France. Papers.*

Serna Montoya, E., & Serna Arenas, A. (2014). Ontology for knowledge management in software maintenance. *International Journal of Information Management, 34*(5), 704–710.

Simpson, J. & Brown, S. (2013). From XML to RDF in the Orlando Project. In *Proceedings. International Conference on Culture and Computing. Culture and Computing 2013. 16-18 September 2013.* Kyoto: IEEE Xplore Digital Library, (pp. 194–195).

Uschold, M. & Gruninger, M. (1996). Ontologies: Principles, methods, and applications. *Knowledge Engineering Review, 11*(2), 93–155.

Van Deursen, D., Poppe, C., Martens, G., Mannens, E. & Van de Walle, R. (2008). XML to RDF Conversion: A Generic Approach. In P. Nesi, K. Ng & J. Delgado (Eds.), *Proceedings. Fourth International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution. Florence, Italy. 17 – 19 November 2008.* IEEE Xplore Digital Library, (pp. 138–144).

Wandl-Vogt, E. (2008). …wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen). In P. Ernst (Ed.), *Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. – 23. September 2006.* Wien: Praesens, (pp. 93–112).

Wandl-Vogt, E. (2012). *Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema).* Retrieved January 17, 2018 from https://wboe.oeaw.ac.at/projekt/beschreibung/

WBÖ (1970–2015). *Wörterbuch der bairischen Mundarten in Österreich. Bayerisches Wörterbuch: I. Österreich.* 5 vols. Ed. by Österreichische Akademie der Wissenschaften. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Language Resource References

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria] (DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010) (Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema). Wien. [Processing status: 2018.01.]

A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis

Yalemisew Abgaz[1], Amelie Dorn[2], Barbara Piringer[2], Eveline Wandl-Vogt[2], Andy Way[1]
[1] Dublin City University, [2] Austrian Academy of Sciences
Dublin Ireland, Vienna, Austria
{Yalemisew.Abgaz, Andy.Way}@adaptcentre.ie, {Amelie.Dorn, Barbara.Piringer, Eveline.Wandl-Vogt}@oeaw.ac.at

Abstract

Around the world, there is a wide range of traditional data manually collected for different scientific purposes. A small portion of this data has been digitised, but much of it remains less usable due to a lack of rich semantic models to enable humans and machines to understand, interpret and use these data. This paper presents ongoing work to build a semantic model to enrich and publish traditional data collection questionnaires in particular, and the historical data collection of the Bavarian Dialects in Austria in general. The use of cultural and linguistic concepts identified in the questionnaire questions allow for cultural exploration of the non-standard data (answers) of the collection. The approach focuses on capturing the semantics of the questionnaires dataset using domain analysis and schema analysis. This involves analysing the overall data collection process (domain analysis) and analysing the various schema used at different stages (schema analysis). By starting with modelling the data collection method, the focus is placed on the questionnaires as a gateway to understanding, interlinking and publishing the datasets. A model that describes the semantic structure of the main entities such as questionnaires, questions, answers and their relationships is presented.

Keywords: Ontology, E-lexicography, Semantic uplift

1. Introduction

There is a substantial amount of traditional data available on the internet and intranets of organisations. Traditional data, in this paper, refers to historical, socio-cultural, political, lexicographic and lexical data sets that are collected over an extended period. Public organisations such as museums, national bibliographic centres and libraries are increasingly opening their doors to facilitate access to such data to support research and development beyond their organisational boundaries (Doerr, 2009). This trend enables researchers to access a significant amount of useful primary data of historical, temporal and societal importance (Kansa et al., 2010; Beretta et al., 2014; Meroño-Peñuela et al., 2015). Making these data available, both for humans and machines, however, comes with several shortcomings.

First, in the majority of cases, these traditional data are initially available in bulk of archival formats, providing only a general description of the content of the data. However, they fail to provide detailed information about why, how, when and who collected the data and how the data can be interpreted and used. Often, consumers of such data require additional contextual information to understand and interpret the information contained in the datasets correctly. This is undoubtedly undesirable as it requires a considerable effort to understand and utilise the dataset.

Second, no matter how big and valuable a released dataset is, it is virtually impossible for machines to use the data without proper semantics for interpreting its content. As machines are becoming ever more typical consumers of such datasets, it has become crucial to include standardised machine-readable semantics in addition to the data itself. The limited availability of semantics to describe the data is, therefore, one of the leading obstacles for machines discovering and interpreting legacy data.

Third, interlinking of the data with other available datasets becomes difficult. The lack of semantics, the use of non-standard vocabulary or the absence of schema mapping (Bizer, Heath, & Berners-Lee, 2009) are some of the causes. Traditional data that includes a schema definition or a data dictionary provides useful information to aid the process of speedy utilisation, but often lacks the information about the means of interlinking the data with existing datasets especially with those available on the linked open data (LOD) platform. The interlinking of the data using a data dictionary further requires a mapping from the data dictionary to a standard vocabulary. This not only requires domain knowledge, but also a detailed knowledge of the internal structure of the data.

In this paper, we focus on a historical data collection of the Bavarian Dialects covering almost a century old data (1911-1998) from the present-day Austria. For effective opening up and utilisation of the collection, we present our approach to facilitating the semantic modelling, enrichment and publishing of traditional data, taking the data collection questionnaires and their individual questions as the starting point. The questionnaires and questions are essential parts of the entire collection as they serve as an entry point to access the answers, where typically neither the headword nor the definition are noted as standard terms. The use of linguistic and cultural concepts in the model thus allows for the exploration and exploitation of cultural links, which is one of the main aims of the exploreAT! project. The questionnaires of the "Datenbank der bairischen Mundarten in Österreich (DBÖ/dbo@ema)" within the project exploreAT! (Wandl-Vogt, 2012) is used as a case study to demonstrate the process. The approach is composed of major steps such as domain analysis, schema analysis, semantic model and semantic up-lift. Domain analysis includes the understanding of the rationale of the data collection, the method of data collection, the original documents used, primary agents that produced the data collection methods and those agents who collected the data. By employing this step, it is possible to collect significant semantics that describes the collection. Schema analysis of the dataset at various stages is also a crucial step, which includes a closer inquiry of the structure of the data, the relationship between entities and their attributes and investigation of any inconsistencies and anomalies. The semantic modelling

step focuses on representing the structure and the semantics of the entities in the datasets using a well-defined semantic model. It is another essential step especially for domains that lack a suitable vocabulary to describe entities fully. In the absence of such vocabulary, it becomes crucial to build a semantic model of the domain from scratch. Finally, the semantic model is used to up-lift, interlink and integrate the data with other related datasets. It will serve as a means to open up valuable traditional data to support further research and possibly answer various questions involving the evolution of conceptualisations of societies in the past and the present.

This approach enables organisations to make their datasets not only digitally available but also semantically enrich the dataset to facilitate a common understanding, interpretation and consumption by both machines and humans. The focus of this paper is, thus, to present our approach and the resulting semantic model. Even if the overall semantic model covers various aspects of the data, at this stage, it will focus only on modelling the questionnaires and questions, which provides users with a unique perspective of accessing the data, looking at it from the original questions and navigating to the corresponding answers, collectors or entities of interest. The model will further facilitate conceptual interoperability (Chiarcos et al., 2013) with other LOD repositories.

This paper is structured in the following way: Section 2 sheds light on the domain and describes the nature of the datasets in use. Section 3 presents the approach including domain and schema analysis and Section 4 discusses the core semantic model using the exploreAT! case study of Bavarian Dialects. In Section 5, we present ongoing work to utilise the semantic model towards the publishing of the datasets using LOD principles. Finally, the conclusion and future work are discussed in Section 6.

2. Background

2.1 Database of Bavarian Dialects (DBÖ)

The database of Bavarian Dialects (*Datenbank der bairischen Mundarten in Österreich -DBÖ) [Database of Bavarian Dialects in Austria]* (Wandl-Vogt, 2008) is a historical non-standard language resource. It was originally collected in the Habsburg monarchy with the aim of

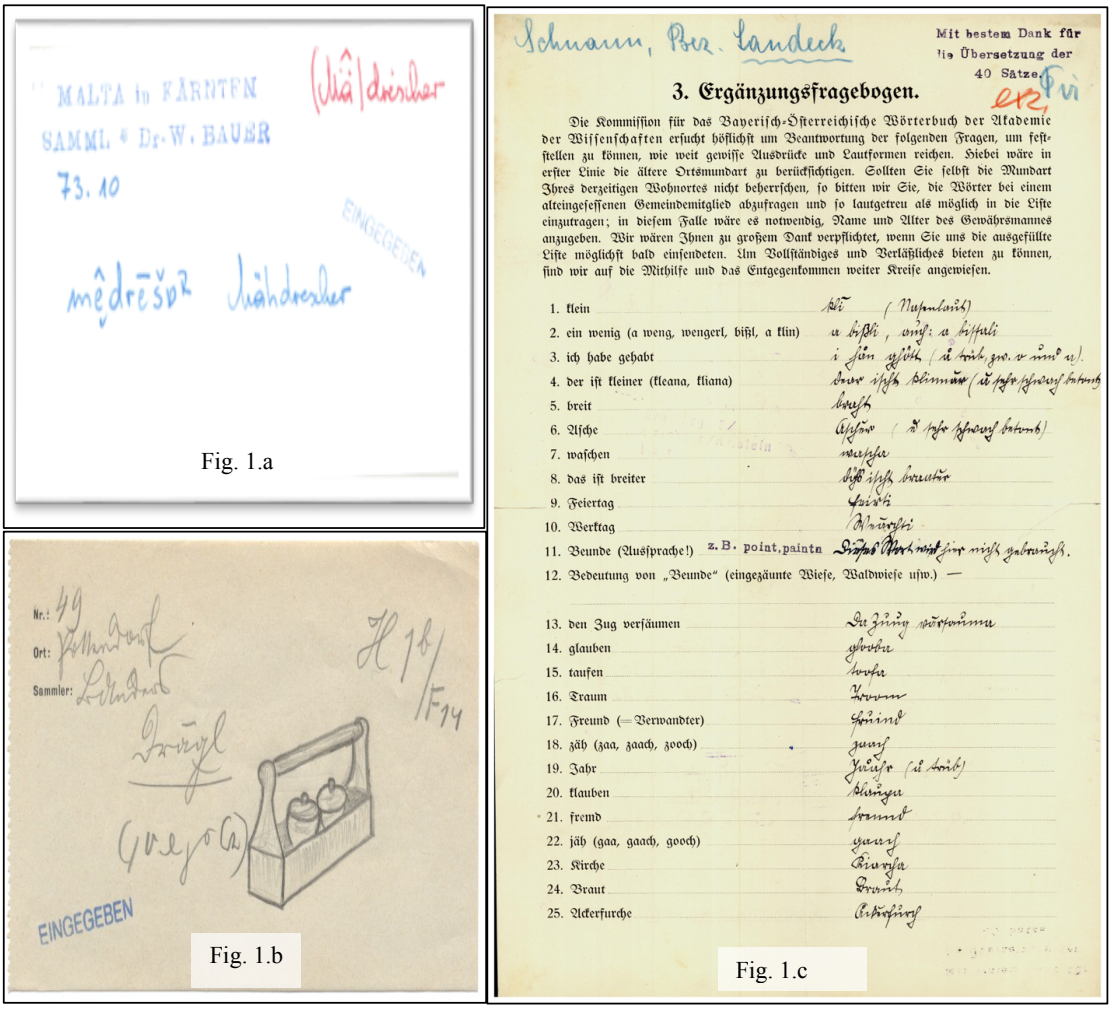

Figure 1. Sample paper slips (a, b) and sample filled questionnaire((Ergänzungsfragebogen) (c).

documenting the German language and rural life in Austria from the beginnings of the Bavarian dialect to the current day. The inception of the data collection went back to 1913 and continued until 1998 in present-day Austria, Czech Republic, Slovakia, Hungary and northern Italy, leaving a century-old historical, socio-cultural and lexical data resource. Even if the original aim of the collection was to compile a dictionary and a linguistic atlas of Bavarian dialects (Arbeitsplan, 1912) spoken by the locals, the data includes various socio-cultural aspects of the day-to-day life of the inhabitants, such as traditional customs and beliefs, religious festivities, professions, food and beverages, traditional medicine, and many more (Wandl-Vogt, 2008)

The data was collected using 109 main questionnaires, nine additional questionnaires (*Ergänzungsfragebögen*) and two *Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen questionnaires* and other additional freestyle questionnaires and text excerpts from various sources such as vernacular dictionaries and literature. In total, there are 24,382 individual questions corresponding to the available questionnaires in the collection. In response to the questionnaires over the span of the project, several million (~ 3.6 million) of individual answers noted on paper slips (Fig. 1.a, b) were collected. The answers to the questions include single words, pronunciations, illustrations and explanations of cultural activities on topics such as traditional celebrations, games, plays, dances, food and other topics.

In addition to the primary data, the entire collection also includes biographies of individual collectors and contributors of various roles. 11,157 individuals who had various functions in the project had participated in the data collection process as authors of the questionnaires, data collectors, editors or coordinators, with some having several of these functions at once. Detailed information about the personal background of individual contributors which was also noted in the course of data collection and during the digitisation process in later years is stored in a specific database (*Personendatenbank [person database]*). Persons and their background are thus other important features of the data that offer additional points for the exploration and the systematic opening of the collection.

The data set further contains additional information about the geographic locations and names of places including cities, districts and regions related to the places where the questionnaires were distributed. In rare cases, the paper slips include information about the time of the data collection.

The collected data has been used to produce a dictionary, *Wörterbuch der bairischen Mundarten in Österreich [Dictionary of Bavarian Dialects in Austria]* (WBÖ); up to now five volumes (A–E, P and T) have been published. Today, about three-quarters of the collected paper slips are available in a digital format following several stages of digitisation. The available formats corresponding to the stages include scanned copies of the paper slips, a textual representation of the paper slips in TUSTEP, MySQL (Barabas et al., 2010) and TEI/XML (Schopper, 2015). This is an ongoing effort to make the data accessible and analyse them, including the use of semantic web technologies to make the data suitable for semantic publishing in the LOD platform.

3. Approach

There is an increasing focus on semantic publishing of traditional data using LOD platforms. To support this, different approaches are used to enrich and expose the data stored in legacy databases semantically. One such approach, direct conversion, converts structured databases (usually relational databases and XML files) directly to RDF triples (Berners-Lee, 1998). This approach mainly uses the schema of the legacy system to transform the data. The transformed data, usually in a triple format (subject, predicate, object), is published as a separate service to the legacy data or as a new layer on top of the legacy database. This approach allows a mass conversion of legacy data without the need for analysis beyond the available schema. However, one of the drawbacks of this approach is that it is restricted to the semantics available within the data and adds little semantics other than the one contained in the schema (Simpson & Brown, 2013). This approach is mainly applicable for general collections but requires a detailed analysis when the domain of interest becomes specialised.

The alternative to this approach focuses on the analysis of the domain of interest and generate/select one or more ontologies that describe the semantics of entities and their relationships. This approach is more rigorous in that experts define the semantics of each entity and its properties. Besides, it facilitates inclusion of the domain knowledge of the experts and opens up a way of accommodating entities that are relevant to the domain but not included in the dataset. The downside of this approach is that it requires a certain level of domain-expert involvement and may require more effort and expert agreement. However, this approach provides a robust semantics and significantly contributes to interoperability.

In our work, we merge the two approaches and use schema analysis to identify entities, attributes and their relationships and domain analysis to analyse and describe the domain and to understand the rationale of the data collection method.

3.1. Schema Analysis

The availability of the dataset in various formats motivates us to look into schema analysis. The questionnaires are available as analogue paper copies, flat text files, in TEI/XML format and a relational table format (dbo@ema). The schema analysis of the available datasets provides us with valuable information to build our semantic model. Research (Ferdinand, C. Zirpins, & D. Trastour, 2004; Deursen et al., 2008; Battle, 2006) has shown that schema analysis provides significant information. The quality of the resulting semantic data, however, depends on the completeness and expressiveness of the available schema and does not reflect the meanings of the entities. In many cases, even if the structural information is available, accurate interpretation of the meaning conveyed by a given schema and its mapping to a standard vocabulary is difficult to achieve. For example, a relational schema

which stores the year as "Year" requires accurate interpretation of whether the attribute "Year" refers to the year of publication of the questionnaire or the year it is distributed to data collectors or any other interpretation. Additionally, it requires an accurate description to resolve if "year" can be considered the same as "dcterms:date". Despite these drawbacks, schema analysis plays a significant role in identifying entities, attributes and their relationships.

```
*LT1* h-at~-in<<e [pl] *ANMO* Pl. ?
*LT2* h-at~-in [m,sg]
*BD/LT1* Abteilung für Heu
**********
*A* HK 157, d157^#3.1 = T1570317.sch^#3, korr. W.B.
*HL* (Ge—sott)tin:1
*QU* Matrei OTir. ob. Iselt., Aufn. Gabriel
*QDB* {1B.0f01} obIselt.:Iselt.:Iselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* ks-O((ut~-in
*BD/LT1* Abteilung für Gesott
**********
*A* HK 157, d157^#4.1 = T1570317.sch^#5, korr. W.B.
*HL* Stadel:1
*QU* Matrei OTir. ob.Iselt.
*QDB* {1B.0f01} obIselt.:Iselt.:Iselgeb.:OTir.:Tir.  Aufn. Gabriel *O* Matrei/O.
OTir.
===
*LT1* >st-..odl
**********
```

Figure 2. TUSTEP format

Schema description: through the life of the dataset, various software tools have been used to store and process the data. Currently, the software includes TUSTEP (Fig. 2), XML/TEI (Fig. 3) and MySQL (Fig. 4). Each of these tools keeps some schema of their own to describe the contents of the files. Having studied all these formats to understand the schema, we used the relational database schema as our

```
<entry xml:id="d157_qdb-d1e2" xml:lang="bar">
   <form type="hauptlemma">
      <orth>Tin</orth>
   </form>
   <gramGrp>
      <pos>Subst</pos>
   </gramGrp>
   <form type="lautung" n="1">
      <pron notation="tustep">|A t~-in</pron>
      <pron notation="ipa" resp="#JB" change="01">|A t~-in</pron>
      <gramGrp>
         <gram> [m,sg+U]</gram>
      </gramGrp>
   </form>
   <form type="lautung" n="2">
      <pron notation="tustep">t~-in;e</pron>
      <pron notation="ipa" resp="#JB" change="01">t~-in;e</pron>
   </form>
   <sense corresp="this:LT1">
      <def xml:lang="de">Abteilung für Heu</def>
   </sense>
   <cit type="kontext" n="1">
      <quote>in den t~-;in [m,sg4] hinein</quote>
      <quote resp="#JB" change="01">in den t~-;in hinein</quote>
   </cit>
```

Figure 3. XML/TEI format

main source containing 88 relational tables. In this paper, our focus is on the schema which is directly related to questionnaires (4), questions (2), authors (n=7) and answers (n=7).

From the schema analysis, entities such as questionnaire (Fragebogen), types of questionnaires, questions (Frage), answers and authors are identified. Attributes of these entities and their data types are also identified.

Field	Type	Null	Key	Default	Extra
id	int(11)	NO	PRI	NULL	auto_increment
nummer	varchar(28)	NO	UNI	NULL	
titel	varchar(512)	NO		NULL	
schlagwoerter	varchar(1024)	YES		NULL	
erscheinungsjahr	int(11)	YES		NULL	
person_id	int(11)	YES		NULL	
originaldaten	text	YES		NULL	
anmerkung	text	YES		NULL	
freigabe	tinyint(1)	NO		NULL	
checked	tinyint(1)	NO		NULL	
wordleiste	tinyint(1)	NO		NULL	
druck	tinyint(1)	NO		NULL	
online	tinyint(1)	NO		NULL	
publiziert	tinyint(1)	NO		NULL	
fragebogen_typ_id	int(11)	YES	MUL	NULL	

Figure 4. MySQL format

Each attribute of the entities is examined for the relevance of the conveyed information in addition to the availability of usable data. There are attributes that contain null values for all records and columns with redundant information. For example, the attribute "wordleiste" ("MS Word Bar") in Fig. 4 contains empty values across all the records in the table. Such attributes are identified and presented to the domain experts for further analysis. There are also attributes that contain null values for some of the records and are left as they are, as there are possibilities to populate them from other sources. Expert evaluation categorised these attributes as "relevant", "needs further investigation" and "not relevant". We included the first two categories but discarded the "not relevant" ones. Finally, the entities and attributes are used as an input for preparing the semantic model.

3.2. Domain Analysis

Domain analysis serves as another step for understanding the rationale of the data collection and the data collection process itself. It provides a solid foundation about why, how, when and by whom the data was collected, stored and processed. It further provides a solid base for understanding the core entities of the datasets, the relationship among the entities and across other entities of similar purpose. Our approach starts with the study of primary sources of information, investigating and examining original materials, interviewing users and maintainers of the dataset. It also includes secondary sources to complement and clarify the domain knowledge.

Following the approach used by Boyce & Pahl (2007), the domain analysis stage seeks information related to 1) Purpose - the rationale of the data collection, 2) Source - the data collection method used, 3) Domain - the nature of the collected data, and 4) Scope - what are the core entities of interest.

Purpose: The purpose of the data collection is to document the wealth of diversity of rural life and unite it under a Pan-European umbrella with a special focus on German language and diverse nationalities in the late Austro-Hungarian Monarchy (Gura, Piringer, & Wandl-Vogt,

forthcoming). The rationale of the data collection serves as a guidance for tuning our objectives and achieving the results. Thus, accordingly, our long-term interest is to capture the lexical data, represent it using standard vocabularies and interlink it with other collections.

Source: The primary data is collected using questionnaires with one or more questions. Questionnaires were distributed to the collectors, and the collectors filled the questionnaires by asking individuals and groups. In some cases, the collectors filled out the questionnaires themselves after observing teams of respondents. Then, collectors sent out the completed questionnaires to the centre where the data was further processed. The questions could be completed by one respondent or a group of respondents. In other cases, questions were filled by the data collectors themselves. Paper slips containing answers arrived at the centre even after several years and are stored in drawers alphabetically.

An interesting aspect of the domain analysis is the identification of the different question types which are not mentioned in any of the available schemas. A closer look at the questions resulted in the identification of patterns of questions used. The data collection is systematic in that it associates certain abbreviations to the questions that have asked similar types of questions. For example, phonological questions have abbreviations such as "*Aussprache, Ausspr.* or *Ltg.*" morphological questions have "*Komp.*" and synonym questions have "*Syn.* or *Synonym*" patterns. However, not all the questions have such abbreviations. The question types and their definitions are represented in detail in the next section. As the questions are linked to the answers, it is also possible to identify the different types of answers provided for a given question. The identification of question types by the domain experts will play a significant role for question-answering systems by exploiting these categories. However, modelling the answers is beyond the scope of this paper.

Domain: The primary data collected is lexical data in direct response to the questions of the questionnaire. It covers various aspects such as names, definitions, pronunciations, illustrations and other categories targeting a linguistic atlas and dictionary compilation (Arbeitsplan, 1912). However, there are other data generated during the process, including details of data collectors, the time and place of the data collection. Regarding the domain, the main interest is the linguistic data of historical and cultural importance.

Scope: From the above steps, we already identified the core entities contained in the datasets. These entities are defined and described by experts. The focus of this exercise is to use the questionnaires as the main entry point to semantically explore the data. Questionnaires contain individual questions of a particular topic which are linked to individual answers. However, in this paper, we will mainly focus on modelling questionnaires and the questions and explore obvious links to answers, authors, collectors and geographic locations. By doing so, we provide additional information which is relevant to answer research questions regarding gender-symmetry or spatiotemporal distributions. However, modelling the answers is complex and will not be discussed in detail in this paper. A pilot for modelling geographic locations is developed and treated separately (Scholz et al., 2016; Scholz, Hrastnig, & Wandl-Vogt, 2018).

4. Semantic Modelling

As a means of semantically enriching the datasets to publish it as a LOD, a semantic model was developed that incorporated the questionnaire model (Fig. 5) and the question model (Fig. 6) with a link to the associated entities. Both models are ontological models built using the Web Ontology Language (OWL2)[1] specification following ontological principles (Noy & Mcguinness, 2001; Edgar & Alexei, 2014). These models provide:

- A succinct definition of the entities and their relations,
- Interoperability with existing semantic resources to support LOD, and
- Extensibility to introduce new classes and relations.

There are many ontologies available to describe data of interest. These ontologies range from general purpose upper ontologies to lower, domain-specific ontologies to describe fine-grained knowledge for describing historical and cultural domains. After deciding the domain and the scope, the next step in the modelling stage is to consider reusing existing ontologies as this is preferable to developing an in-house ontology. However, for domain-specific description of datasets, it is difficult to find a suitable ontology and thus requires preparation either from scratch or extending existing ones.

We searched existing ontologies that can describe our domain of interest. The main repositories searched include LOV[2] ontology repository, Schema.org[3] and other specialised search tools such as Watson semantic web search engine.[4] We found terminologies related to questions, answers and questionnaires, but they do not fit our requirements, and such ontologies are not available yet. However, we will exploit some of the concepts defined in the Ontolex-Lemon model (McCrae et al., 2017) to describe the lexical data in the collection. We will further reuse vocabularies such as FOAF, SKOS and Dublin Core to describe authors, editors, collectors, places and publication. In addition to describing the entities, generic ontological constructs are used to create an interlinking with concepts from other repositories, and to compare our data with other similar data sets using meaningful interoperability.

A combination of top-down and bottom-up approaches as proposed by Uschold & Gruninger, (1996) is used to develop the model. The approach integrated domain analysis as a top-down approach and schema analysis as a bottom-up approach to build the ontology, in order to support our domain-specific requirements. We also used existing standardised vocabularies for entities that already have compatible representations. We developed our

[1] https://www.w3.org/TR/owl2-primer/
[2] http://lov.okfn.org
[3] http://schema.org
[4] http://watson.kmi.open.ac.uk/WatsonWUI/

ontology to represent both the structure and the meaning of the entities of interest.

4.1. Questionnaire Model

The questionnaire model is built based on the detailed analysis of the original and physically compiled book of sets of questionnaires and its electronic version (dbo@ema). Up to now, we have identified three questionnaire types. Each type has its characteristics and differs from the others in its purpose, the type of information it seeks and its format, including its physical appearance. Treating the different sets of questionnaires independently is crucial to preserve the historical importance and the structural and semantic relation each questionnaire set has with the collected data. The questionnaire types are discussed below:

1. Systematic: [*Systematischer Fragebogen*] is a questionnaire that is used to collect the original data. This type of questionnaire is used from the beginning of the data collection process.
2. Additional: [*Ergänzungsfragebogen*] is a questionnaire that is used as a supplementary questionnaire to the systematic questionnaire.
3. Dialectographic *[Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen] is a* questionnaire of the Munich and Vienna Dictionary Commissions.

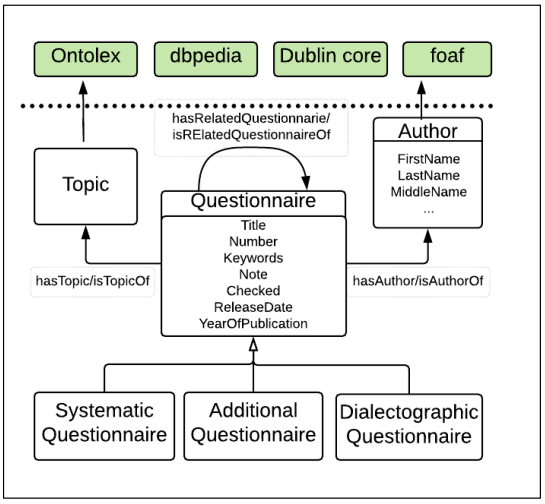

Figure 5. A Semantic model of questionnaire

A questionnaire may have one or more related questionnaires that deal with the same topic. We observed that questionnaires refer to other questionnaires. Such relationships are captured by an object property "hasRelatedQuesitonnaire" with an inverse property "isRelatedQuestionnaireOf". A questionnaire has at least one topic, and this relationship is captured by "hasTopic" with "isTopicOf" inverse object property. Furthermore, a questionnaire has at least one Author, and this relationship is captured by "hasAuthor" object property with "authorOf" inverse object property.

Topics (Questionnaire Topics). A topic is the main subject of the questionnaire or a given question. A questionnaire may focus on a general topic such as "Food" and a question may cover subtopics such as "Traditional Food". This information will be treated as a topic following a proper disambiguation technique and then relate to ontolex:lexicalConcept.

Author/collectors. Authors are defined in FOAF and Dublin Core. We will reuse the definition provided in FOAF Agent/Author classes.

4.2. Question Model

A question is a linguistic expression used to request information, or the request made using such an expression. The information requested is provided in the form of answer.[5] In this ontology, we categorise the questions mainly based on the content, the forms and the expected answers from the respondents. An analysis carried out by the experts, users and ontology engineers identified 12 different types of questions and added two more questions to accommodate future processing of additional questionnaire sets. It is important to note that these question types are not mutually exclusive to one another and there are instances of questions that belong to more than one type of questions, e.g. the question "*Kopf: Kopf/Haupt (in urspr. Bed.) in Vergl./Ra. (Kopf stehn, der Kopf mI¦chte einem zerspringen)*" is both semasiological and syntactic. The semantics of the question types are given below:

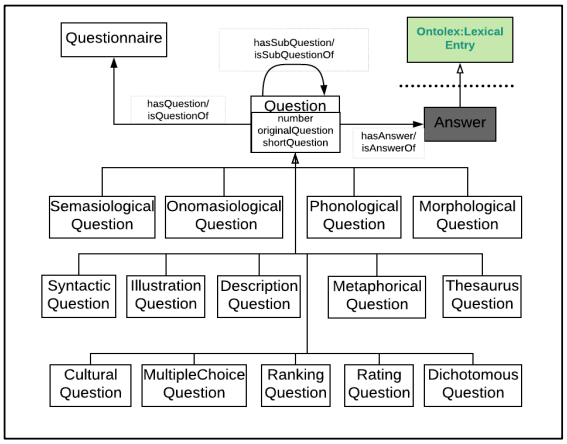

Figure 6. A Semantic model of question

1. *Onomasiological:* asks for the name of a given entity, e.g. "how do you call x?" where x represents an entity.
2. *Semasiological:* asks for the meaning of a given entity, e.g. "what does x mean?".
3. *Dichotomous:* asks for a selection of answers from a binary option. It includes yes/no or agree/disagree types of answers to stated questions.
4. *Description:* asks for a written representation of a given entity, e.g. "What would be the function of x?".

[5] https://en.wikipedia.org/wiki/Question

5. *Illustration:* asks for a pictorial or diagrammatic representation of a given entity, e.g. "What does x look like?".
6. *Morphological:* asks about the structure and the formation of words and parts of words. Based on the structure, morphological questions can take various forms.
7. *Phonological:* asks for the pronunciation, or phonetic representation of words.
8. *Syntactic:* asks for construction of phrases or sentences using a given word or a given idiom, e.g. "Provide a phrase/sentence for/using a word/idiom x".
9. *Metaphorical:* asks for some conveyed meanings given a word or an expression. Metaphorical questions are related to semasiological questions, but they ask for an additional interpretation of the expression beyond its obvious meaning.
10. *Thesaurus:* asks for a list of words or expressions that are used as synonyms (sometimes, antonyms) or contrasts of a given entity.
11. *Cultural:* asks for a belief of societies, procedures on how to make or prepare things and how to play games, contents of cultural songs, poems used for celebrations. Analysis of the existing questions shows that the cultural question type has its subtypes and has instances that significantly overlap with the other question types.
12. M*ultiple Choice:* asks for a selection of one item from a list of three or more potential answers.
13. *Rating:* asks the respondent to assign a rate (degree of excellence) to a given entity based on a predefined range
14. *Ranking Question:* asks the respondent to compare entities and rank them in a certain order.

It is commonly observed that a question may ask several other sub-questions, and this is captured by the "hasSubQuestion" object property. Thus, the object property "hasSubQuestion" relates one question with its subquestions. Each question is linked to its associated answer. A question may have several answers collected from different sources. This is captured by the "hasAnswer" object property with its inverse "isAnswerOf". Finally, a question is related to a questionnaire with the "isQuestionOf" object property where a single question is contained only in one questionnaire.

Answer: An answer is a written, spoken or illustrated response to a question. The different types of questions have answers either in a written, spoken or illustration format. In the case of questions that involve lexical data collection, the answer could be associated with some lexical category. For each types of questions, there are different types of answers including sentences, individual words, multiword expressions, affix, diagrams, etc. Modelling the answers is under investigation. However we will treat answers with single word, multiword expression or affixes as ontolex:lexicalEntries. For example, the answer to a thesaurus question is expected to be a word, or multiword expression in the OntoLex model.

Finally, an initial version of the ontology- Ontology for Lexical Data Collection and ANalysis (OLDCAN)[6] is developed following the approach discussed above. Since the project is at its development stage, a permanent URL has not been yet assigned to either the ontology or to the data. However, the ongoing results are available under a Creative Commons Licensing.[7]

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia: <http://dbpedia.org/ontology/> .
@prefix oldcan: <http://localhost/oldcan/OLDCAN#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

<#QuestionnaireTriplesMap>
rr:logicalTable [ rr:sqlQuery """

  SELECT Fragebogen_V1.*, (CASE QUESTIONNAIRE_TYP_ID
    WHEN '1' THEN 'SystematicQuestionnaire'
    WHEN '2' THEN 'AdditionalQuestionnaire'
    WHEN '3' THEN 'DialectographicQuestionnaire'
  END) QUESTIONNAIRETYPE FROM Fragebogen_V1
  """ ];
    rr:subjectMap [
        rr:template "http://localhost/dboe/Questionnaire/{ID}" ;
        rr:class oldcan:Questionnaire ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate rdf:type ;
        rr:objectMap [ rr:template "http://localhost/oldcan/OLDCA#
        {QUESTIONNAIRETYPE}";
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:title ;
        rr:objectMap [ rr:column "TITLE" ;rr:language "de";] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;] ;
    rr:predicateObjectMap [
        rr:predicate oldcan:publicationYear ;
        rr:objectMap [ rr:column "YEAR_OF_PUBLICATION" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    rr:predicateObjectMap [
        rr:predicate oldcan:note ;
        rr:objectMap [ rr:column "NOTE" ] ;
        rr:graph <http://localhost/dboe/Questionnaire_graph> ;];
    ];
    .
```

Figure 7. R2RML mapping excerpts

5. Semantic Up-lift

This stage focuses on the use of the semantic model and selected vocabularies to semantically enrich the data. It is used to annotate every data element with semantic information that states what it is, how it should be interpreted and how it is related to other elements within the datasets or across other datasets. There are various methods and tools used to transform relational databases to semantically compatible formats including direct mapping (Berners-Lee, 1998) and domain semantics-driven mapping (Michel, Montagnat, & Faron, 2013). We followed R2RML[8] to annotate our datasets due to its customisability for mapping relational databases into triples. Unlike direct mapping that depends on the database's structure, it is possible to use an ontology of the domain. Since R2RML is a vocabulary by itself, it stores

[6] http://exploreat.adaptcentre.ie/#Semantics
[7] https://creativecommons.org/licenses/by/3.0/at/deed.en
[8] https://www.w3.org/TR/r2rml/

the mappings from a relational database to RDF as RDF files and allows inclusion of provenance information. This facilitates knowledge discovery and reuse of mappings. However, it requires more effort compared to direct mapping. R2RML is used to map the relational data into a LOD. This phase includes the following steps:

1. Converting the major tables into classes,
2. Mapping object property relationships,
3. Mapping data property relationships,
4. Enriching the data with additional semantics.

To demonstrate the envisioned mapping, excerpts of the mapping file for both questionnaire and questions are generated. In the mapping (Fig. 8), each questionnaire is associated to oldcan:Questionnaire class using "a" ("rdf:type") property. The template defines the URL of the specific location of the questionnaire. The selected attributes are mapped to data properties, e.g. title is mapped to oldcan:title and the language of the title is included using a language tag "de".

The mapping of the questions is done similarly. Here the object property isQuestionOf is used to link the question with its questionnaire. In the ontology, the hasQuestion object property is defined as an inverse of isQuestionOf to achieve both brevity and searchability in the generated data. The different types of the questionnaires and the questions are captured. An excerpt of the resulting triple[9] is presented in Fig. 8.

```
<http://localhost/dboe/Questionnaire/1>
    a    <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
         <http://localhost/oldcan/OLDCAN#Questionnaire> ;
    <http://localhost/oldcan/OLDCAN#note>
         "resfb1" ;
    <http://localhost/oldcan/OLDCAN#publicationYear>
         "1920" ;
    <http://localhost/oldcan/OLDCAN#title>
         "Kopf (1)"@de .

<http://localhost/dboe/Questionnaire/2>
    a    <http://localhost/oldcan/OLDCA#SystematicQuestionnaire> ,
         <http://localhost/oldcan/OLDCAN#Questionnaire> ;
    <http://localhost/oldcan/OLDCAN#note>
         "bafb2" ;
    <http://localhost/oldcan/OLDCAN#publicationYear>
         "1920" ;
    <http://localhost/oldcan/OLDCAN#title>
         "Die Osterwoche (1)"@de .
.....
<http://localhost/dboe/Question/1-A11>
    a    <http://localhost/oldcan/OLDCAN#Question> ;
    <http://localhost/oldcan/OLDCAN#isQuestionOf>
         <http://localhost/dboe/Questionnaire/1> ;
    <http://localhost/oldcan/OLDCAN#number>
         "A11" ;
    <http://localhost/oldcan/OLDCAN#originalQuestion>
         "Kopf: breiter Kopf"@de ;
    <http://localhost/oldcan/OLDCAN#shortQuestion>
         "breiter Kopf"@de .

<http://localhost/dboe/Question/111-2>
    a    <http://localhost/oldcan/OLDCAN#Question> ;
    <http://localhost/oldcan/OLDCAN#isQuestionOf>
         <http://localhost/dboe/Questionnaire/111> ;
    <http://localhost/oldcan/OLDCAN#number>
         "2" ;
    <http://localhost/oldcan/OLDCAN#originalQuestion>
         "Altane im 1. Stock (Şller, Schrot, Laube, Br_ckel)"@de ;
    <http://localhost/oldcan/OLDCAN#shortQuestion>
         "Altane im 1.Stock (Şller, Schrot, Laube,*)"@de .
```

Figure 8. Questionnaire and question triples

6. Conclusion and Future Work

The effort to open up legacy databases to make them accessible, usable and researchable has increased with the development of LOD platforms. Such platforms facilitate publishing legacy data of a wide range of contents and formats. As the content becomes specialised, the need for finding and developing semantic models that describe the domain of interest become crucial. This paper has presented an approach which is currently used for building a semantic model for enriching and publishing traditional data of historical, cultural and lexical importance. It is argued that the use of such an approach for building semantic models to assist with semantic publishing of traditional data on the LOD platform is vital to the exploitation of data of historical importance. It further paves the way for researchers to understand and compare conceptualisation of entities at different times and their evolution through time. As the paper presents work in progress, our immediate focus is the enrichment of the semantic model by in-depth examination of the entities including answers to the questions to enable a strong semantic interlinking that will facilitate efficient question answering and comparison of the different types of questions. Furthermore, additional enrichment to interlink the data with other similar datasets and the visualisation of the dataset will be the next area to tackle.

7. Acknowledgements

This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! project and the ADAPT Centre for Digital Content Technology at Dublin City University which is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is cofunded under the European Regional Development Fund.

[9] http://exploreat.adaptcentre.ie/#APIs

Bibliographical References

Arbeitsplan (1912). *Arbeitsplan und Geschäftsordnung für das bayerisch-österreichische Wörterbuch. 16. Juli 1912.* Karton 1. Arbeitsplan-a-h Bayerisch-Österreichisches Wörterbuch. Archive of the Austrian Academy of Sciences. Wien.

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B. & Schwaiger, S. (2010). Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In H. Bergmann, M. M. Glauninger, E. Wandl-Vogt, & S. Winterstein (Eds.), *Fokus Dialekt. Analysieren – Dokumebattentieren – Kommunizieren. Festschrift für Ingeborg Geyer zum 60. Geburtstag.* (Germanistische Linguistik 199–201). Hildesheim, Zürich, New York: Olms, (pp. 47–64).

Battle, S. (2006). Gloze: XML to RDF and back again. *First Jena User Conference.* Bristol, UK.

Beretta, F., Ferhod, D., Gedzelman, S. & Vernus, P. (2014). The SyMoGIH project : publishing and sharing historical

data on the semantic web. *Digital Humanities 2014, July 2014*, Lausanne, Switzerland. (pp. 469–470).

Berners-Lee, T. (1998). *Relational Databases on the Semantic Web.* In *Design Issues for the World Wide Web.* Retrieved January 10, 2018, from https://www.w3.org/DesignIssues/RDB-RDF.html

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web Information Systems, 5*(3), 1–22.

Boyce, S. & Pahl, C. (2007). Developing Domain Ontologies for Course Content. *Educational Technology & Society, 10*, 275–288.

Chiarcos, C., Cimiano, P., Declerck, T. & McCrae, J. P. (2013). Linguistic Linked Open Data (LLOD) – Introduction and Overview. In C. Chiarcos, P. Cimiano, T. Declerck, & J. P. McCrae (Eds.), *2nd Workshop on Linked Data in Linguistics.* Pisa, (pp. i–xi).

Doerr, M. (2009). Ontologies for Cultural Heritage. In S. Staab & R. Studer (Eds.), *Handbook on Ontologies. International Handbooks on Information Systems.* Berlin, Heidelberg: Springer.

Ferdinand, M., Zirpins, C. & Trastour, D. (2004). Lifting XML Schema to OWL. In N. Koch, P. Fraternali, & M. Wirsing (Eds.), *Web Engineering. 4th International Conference, ICWE 2004. Munich, Germany, July 26-30, 2004. Proceedings.* (Lecture Notes in Computer Sciences 3140). Berlin, Heidelberg: Springer, (pp. 354–358).

Gura, C., Piringer, B. & Wandl-Vogt, E. (forthcoming). Nation Building durch Großlandschaftswörterbücher. Das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) als identitätsstiftender Faktor des österreichischen Bewusstseins.

Kansa, E. C., Kansa, S. W., Burton, M. M., & Stankowski, C. (2010). Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies, 6*(2), 301–326.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Leiden: Lexical Computing CZ, (pp. 587–597).

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. & Van Harmelen, F. (2015). Semantic Technologies for Historical Research: A Survey. *Semantic Web, 6*(6), 539–564.

Michel, F., Montagnat, J., & Faron Zucker, C. (2013). *A survey of RDB to RDF translation approaches and tools.* Retrieved January 16, 2018, from https://hal.archives-ouvertes.fr/hal-00903568v1

Noy, N. F., & McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology.* Technical, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Scholz, J., Hrastnig, E., & Wandl-Vogt, E. (2018). A Spatio-Temporal Linked Data Representation for Modeling Spatio-Temporal Dialect Data. In P. Fogliaroni, A. Ballatore & E. Clementini (Eds.), *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017).* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 275–282).

Scholz, J., Lampoltshammer, T. J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal Modeling of Linguistic Regions and Processes with Combined Indeterminate and Crisp Boundaries. In G. Gartner, M. Jobst, & H. Huang (Eds.), *Progress in Cartography.* (Lecture Notes in Geoinformation and Cartography) Cham: Springer, (pp. 133–151).

Schopper, D., Bowers, J. & Wandl-Vogt, E. (2015). dboe@TEI: remodelling a data-base of dialects into a rich LOD resource. Retrieved January 17, 2018 from *Text Encoding Initiative Conference and members' meeting 2015, October 28-31, Lyon, France. Papers.*

Serna Montoya, E., & Serna Arenas, A. (2014). Ontology for knowledge management in software maintenance. *International Journal of Information Management, 34*(5), 704–710.

Simpson, J. & Brown, S. (2013). From XML to RDF in the Orlando Project. In *Proceedings. International Conference on Culture and Computing. Culture and Computing 2013. 16-18 September 2013.* Kyoto: IEEE Xplore Digital Library, (pp. 194–195).

Uschold, M. & Gruninger, M. (1996). Ontologies: Principles, methods, and applications. *Knowledge Engineering Review, 11*(2), 93–155.

Van Deursen, D., Poppe, C., Martens, G., Mannens, E. & Van de Walle, R. (2008). XML to RDF Conversion: A Generic Approach. In P. Nesi, K. Ng & J. Delgado (Eds.), *Proceedings. Fourth International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution. Florence, Italy. 17 – 19 November 2008.* IEEE Xplore Digital Library, (pp. 138–144).

Wandl-Vogt, E. (2008). …wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen). In P. Ernst (Ed.), *Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. – 23. September 2006.* Wien: Praesens, (pp. 93–112).

Wandl-Vogt, E. (2012). *Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema).* Retrieved January 17, 2018 from https://wboe.oeaw.ac.at/projekt/beschreibung/

WBÖ (1970–2015). *Wörterbuch der bairischen Mundarten in Österreich. Bayerisches Wörterbuch: I. Österreich.* 5 vols. Ed. by Österreichische Akademie der Wissenschaften. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Language Resource References

[DBÖ] Österreichische Akademie der Wissenschaften. (1993–). Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria] (DBÖ). Wien. [Processing status: 2018.01.]

[dbo@ema] Wandl-Vogt, E. (2010) (Ed.). Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema). Wien. [Processing status: 2018.01.]

Annotating Sumerian: A LLOD-enhanced Workflow for Cuneiform Corpora

Christian Chiarcos*, Ilya Khait*, Émilie Pagé-Perron◇, Niko Schenk*, Jayanthλ, Lucas Reckling◇
*Goethe University Frankfurt, Germany, ◇University of Toronto, Canada,
λUniversity of California, Los Angeles
{chiarcos|khait|schenk}@informatik.uni-frankfurt.de,
{emilie.page.perron|lucas.reckling}@mail.utoronto.ca, jayanthj@ucla.edu

Abstract

Assyriology, the discipline that studies cuneiform sources and their context, has enormous potential for the application of computational linguistics theory and method on account of the significant quantity of transcribed texts that are available in digital form but that remain as yet largely unexploited. As part of the Machine Translation and Automated Analysis of Cuneiform Languages project (https://cdli-gh.github.io/mtaac/), we aim to bring together corpus data, lexical data, linguistic annotations and object metadata in order to contribute to resolving data processing and integration challenges in the field of Assyriology as a whole, as well as for related fields of research such as linguistics and history. Data sparsity presents a challenge to our goal of the automated transliteration of the administrative texts of the Ur III period. To mitigate this situation we have undertaken to annotate the whole corpus. To this end we have developed an annotation pipeline to facilitate the annotation of our gold corpus. This toolset can be re-employed to annotate any Sumerian text and will be integrated into the Cuneiform Digital Library Initiative (https://cdli.ucla.edu) infrastructure. To share these new data, we have also mapped our data to existing LOD and LLOD ontologies and vocabularies. This article provides details on the processing of Sumerian linguistic data using our pipeline, from raw transliterations to rich and structured data in the form of (L)LOD. We describe the morphological and syntactic annotation, with a particular focus on the publication of our datasets as LOD. This application of LLOD in Assyriology is unique and involves the concept of a LLOD edition of a linguistically annotated corpus of Sumerian, as well as linking with lexical resources, repositories of annotation terminology, and finally the museum collections in which the artifacts bearing these inscribed texts are kept.

Keywords: Linked Open Data, Sumerian, Linguistic Linked Open Data, linked dictionaries, syntactic parsing, annotation pipeline, CoNLL, RDF, pre-annotation

1. Introduction

1.1. Sumerian and Cuneiform Studies

The Sumerian language, an agglutinative isolate, is the earliest language recorded in writing. It was spoken in the third millennium BC in modern southern Iraq, and continued to be written until the late first millennium BC. This language was written with cuneiform, a logo-syllabic script with around one thousand signs in its inventory, formed by impressing a sharpened reed stylus into fresh clay. This script was employed in ancient Mesopotamia and surrounding regions to inscribe many different languages, notably the East Semitic Akkadian (Babylonian and Assyrian), the Indo-European Hittite, and others.

In order to make a text available for research, Assyriologists copy and transcribe it from the artifact bearing it. The results of this labor-intensive task are usually published on paper. A dozen projects which make various cuneiform corpora available online have emerged since the early 2000s, building on digital transcriptions created as early as the 1960s. Unfortunately, these initiatives rarely use shared conventions, and the toolset available to process these data is limited, thus vasts numbers of transliterated and digitized ancient cuneiform texts remain only superficially exploited.

1.2. Linked Open Data for Sumerian

Linked Open Data (LOD) defines principles and formalisms for the publication of data on the web, with the goal of facilitating its accessibility, transparency and reusability. The application of LOD formalisms to philological resources within the field of Assyriology promises two crucial advantages. First, we shall be able to establish interoperability and exchange between distributed resources that currently persist in isolated data silos – or that provide human-readable access only, with no machine-readable content. Among other benefits that LOD provides, one should also mention its federation, ecosystem, expressivity, semantics, and dynamicity potential (Chiarcos et al., 2013). Converting out data to an RDF representation is an essential step to open up the possibility of linking with other resources and integrating content from different portals. Further, using shared vocabularies allows us to publish structured descriptions of content elements in a transparent and well-defined fashion. Ontologies play a crucial role in this regard as these define shared data models and concepts.

1.3. The MTAAC Project

The "Machine Translation and Automated Analysis of Cuneiform Languages" (MTAAC) project[1] aims to develop state-of-the-art computational linguistics tools for cuneiform languages, using internationally recognized standards to share the resulting data with the widest possible audience. (Pagé-Perron et al., 2017) This is made possible through a collaboration between the Cuneiform Digital Library Initiative (CDLI)[2] and specialists in Assyriology, computer science and computational linguistics at the Goethe University Frankfurt, Germany, the University of California, Los Angeles (UCLA) and the University of Toronto, Canada.

The project entails the preparation of a methodology and an associated NLP pipeline for the Sumerian language. The

[1] https://cdli-gh.github.io/mtaac.
[2] https://cdli.ucla.edu.

pipeline processes, annotates and translates Sumerian texts, as well as extracts additional information from the corpus. In order to facilitate the study of the language and the historical, cultural, economic and political context of the texts, these data are to be made available both to designated audiences and machines.

In order to facilitate the reusability of these data, as well as to encourage reproducibility, we use linked data and open vocabularies, thereby contributing to interoperability with other resources[3]. Another aim in the application of LOD is to set new standards for digital cuneiform studies and to contribute to resolving data integration challenges both in Assyriology and related linguistic research. The (L)LOD edition for Sumerian and the linking of representative language resources uses lemon/ontolex for lexical data, the CIDOC/CRM for object metadata, lexvo for language identification, Pleiades for geographical information, and OLiA[4] for linguistic annotations. Brigning together corpus data, lexical data, linguistic annotations and object metadata breaks new ground in the field of Assyriology, and computational philology.

2. Corpus Data and Data Formats

2.1. Ur III Data in CDLI

One objective of our project is to complement the range of cuneiform corpora with morphologic, syntactic and semantic annotations for an extensive, but currently undertranslated genre, namely the administrative texts, especially for the Neo-Sumerian language of the Ur III period.

The Cuneiform Digital Library Initiative (CDLI) is a major Assyriological project which aims to provide information on all objects bearing cuneiform inscriptions kept in museums and collections around the world. The images, metadata, transliterations, transcriptions, translations and bibliography are made available online. At the moment the CDLI catalog contains entries for about 334,000 objects out of an estimated total of around 550,000.

The corpus we chose is a subset of these entries: 69,070 administrative and legal texts produced during the Ur III period (2100-2000 BC). These texts are available in transliteration but only 1,966 have parallel English translation. Textual data in the ATF format are presented as follows:[5]

```
&P142051 = WO 11, 21
#atf: lang sux
@tablet
@obverse
1. 2(gesz2) 2(u) 4(disz) udu bar-gal2
#tr.en: 144 sheep with fleece,
2. 4(disz) sila4 bar-gal2
#tr.en: 4 lambs with fleece,
3. 7(disz) udu bar-su-ga
#tr.en: 7 sheep without fleece,
4. 3(gesz2) 1(u) 2(disz) ud5 masz2 hi#-[a]
#tr.en: 192 mixed nanny and billy goats,
5. ki kas4-ta
#tr.en: from Kas
6. lu2-dsuen i3-dab5#
#tr.en: Lu-Suen took;
$ blank space
@reverse
$ blank space
1. mu us2-sa ki-maszki# ba-hul
#tr.en: year after: "Kimaš was destroyed".
```

These data are composed of lines of transliteration that start with a number; they also include structure tags, translation and comments which complement the content of each textual entry.

2.2. Other Sumerian Corpora

Previous research on Sumerian text has produced two corpora; of literary texts (ETCSL) (Black et al., 1998–2006) and royal inscriptions (ETCSRI, within ORACC)[6] respectively, but both corpora were limited to morphosyntactic annotation. To the best of our knowledge, this also corresponds to the state of the art in other branches of Assyriology, where representative morphosyntactic annotations (glosses) have been assembled, for example, within the ORACC[7] portal. Additionally, some other projects offer digital access to unannotated texts.[8]

2.3. Automated Annotation and Analysis

Experiments in automated syntactic annotation have been described by Jaworski (2008) and Smith (2010), but both focused on extracting automatically annotated fragments rather than on providing a coherently annotated corpus. The mORSuL ontology[9], developed to attach CIDOC-CRM to Ontomedia (Nurmikko, 2014; Nurmikko-Fuller, 2015),[10] has only reached the status of a case study. These experiments show the potential interest in Sumerian corpus data published in accordance with Semantic Web principles, but neither of these projects actually aims to provide Linked Data as an end product.

With respect to semantics, current research focuses on shallow techniques such as named entity recognition (SNER[11] on Sumerian), or entity linking and prosopography (Darmstadt on Hittite) – to the best of our knowledge, the annotation of cuneiform corpora with syntactic relations is limited to experiments[12], and semantic relations annotating has not

[3] E.g., Syriac http://syriaca.org, Hebrew http://tinyurl.com/guwe8kr, and Indo-European and Caucasian languages http://titus.fkidg1.uni-frankfurt.de/.

[4] http://www.acoli.informatik.uni-frankfurt.de/resources/olia/.

[5] Text published by Hruška (1980), CDLI entry prepared by Robert K. Englund. https://cdli.ucla.edu/P142051.

[6] http://oracc.museum.upenn.edu/etcsri/; as with all ORACC projects, ETCSRI uses a slightly different version of ATF as its core format.

[7] http://oracc.museum.upenn.edu.

[8] Apart from the CDLI, it is important to mention the Database of Neo-Sumerian Texts (BDTNS), a database of texts dating to the Ur III period http://bdts.filol.csic.es/.

[9] https://github.com/terhinurmikko/morsul.

[10] http://www.contextus.net/ontomedia.

[11] https://github.com/wwunlp.

[12] Karahashi and Tinney have previously worked on a rule based syntax annotator from which we expect to reuse some rules in the further development of our tool. https://github.com/oracc/oracc/tree/master/misc/ssa3. Unfortunately the documentation written by Karahashi is not available for con-

previously been attempted.

2.4. The CDLI-CoNLL Format

The CDLI-CoNLL format is an abridged version of the CoNLL-U format.[13] Because of the scarcity of specialists in the Sumerian language, our format was designed with ease and speed of annotation in mind.

The SEGM field contains information on the lemma, comprising a dictionary word and its sense, appended and in square brackets, e.g. udu[sheep] or dab[seize]. Affixes are standardized in conformity to a list of morphemes, following the ETCSRI project's morphological scheme. These morphemes are separated by a dash placed before the morpheme, except for the first element in the chain. When the analysis of the word demands a morpheme that is not explicit in the writing of the form, it is enclosed in square brackets.

The XPOSTAG field contains the part-of-speech tag associated with the morpheme present in the SEGM column. If the form represents a named entity, the named entity tag will take the place of the POS tag. The tags we employ are again those of the ETCSRI project. These tags are separated using a period placed before the morpheme, except in the case of the first element in the chain.

The information in these fields can easily be converted to the CoNLL-U format following rules and maps. Our converter uses maps to create the UPOSTAG from our domain-specific POS tags and for the conversion of morphemes to the verbose CoNLL-U FEATS column.

Figure 1 illustrates the CDLI-CoNLL format in comparison with the CoNLL-U format as far as morphology and morphosyntax are concerned:[14] the FORM column provides the transliteration of the original cuneiform signs, but without elements marking the state of the text on the medium (breaks, omissions, etc). The original SEGM column provides segmentation into morphological (rather than graphemic) segments. Because of the characteristics of the Sumerian noun, the LEMMA directly follows from this segmentation as its first substring. However, CoNLL-U does not allow us to preserve full SEGM information, so the LEMMA is used instead. The original XPOSTAG includes information about the part-of-speech and named entities categories (SN), as well as grammatical features. However, UD conventions allow us to preserve only parts of the morphological information in CoNLL-U: the last word of a Sumerian noun phrase aggregates all case morphology (its own as well as that of its – preceding – head), a phenomenon known as *Affixanhäufung*. In this case, the place name *Shuruppak* is a genitive attribute of an ergative argument. It is thus inflected for *both* genitive (*-ak*) and ergative (*-e*). In CoNLL-U, multiple case marking is not foreseen, so that here, a language-specific aggregate feature for multiple cases is introduced.[15] In addition, the SN tag marks Shuruppak as a site name, and we derive non-human animacy.

For the mapping between our morphological tags, the Universal dependencies tags and features (as well as Unimorph categories features), we adopt a Linked Open Data approach: we provide an ontological representation of the CDLI annotation scheme, and link its concepts via *skos:broader* (etc.) statements with the UD and Unimorph ontologies provided as part of the OLiA ontologies.[16] CDLI-CoNLL can also be converted to the Brat Standoff format through our pipeline described below, for further syntactic annotation, visualization, or using other tools geared to processing data in this format.

2.5. Linked Open Data Representations

Linked Open Data in Assyriology is limited at the moment to metadata on artifacts, which, however, seems practical when working on cuneiform corpora. The Modref project (Tchienehom, 2017)[17] is used in the classification of museum artifacts and employs CIDOC-CRM for that purpose. CDLI is among the three collections it connects[18]. Additionally, almost 22% of all CDLI artifacts are encompassed by the CIDOC-CRM-based SPARQL end point of the British Museum[19]. Linked Data technology allows us to query disparate artifacts across different collections using explicit links within such repositories. The SPARQL 1.1 federation allows us to query these metadata repositories and to link CDLI data with them.

Edition principles for philological corpora are only just emerging, with different alternative vocabularies (POWLA, NIF, TELIX) currently being discussed. In the MTAAC project, we generally base our proof-of-concept on the morphologically annotated ETCSRI corpus; the application of CoNLL-RDF serves as LOD representation within the CDLI.

2.6. CoNLL-RDF

CoNLL-RDF (Chiarcos and Fäth, 2017) is a rendering of RDF in CoNLL's tab-separated value format. It represents a convenient and human-readable data model that is close to conventional representations and can be serialized in RDF format. Crucially, it is comparably easy to read and parse as CoNLL: it provides the direct means to string-based manipulations that CoNLL is praised for, but in addition it allows

sultation. The only existing cuneiform corpus with (manual) annotation of syntax is the Annotated Corpus of Hittite Clauses, see (Molina, 2017);

[13] `http://universaldependencies.org/format.html`

[14] Both the CoNLL-U and CDLI-CoNLL formats have additional fields to handle relationships between words, such as syntax.

[15] This solution is problematic in that long chains of case markers can arise, and it is no longer possible to generalize over the resulting multitude of case features. Case combinatorics in the ETCSRI corpus yield 47 case chains resulting from only 15 case labels.

[16] `http://purl.org/olia`, for Unimorph, see `http://purl.org/olia/owl/experimental/unimorph/`..

[17] `http://triplestore.modyco.fr:8080/ModRef`.

[18] The other two are the ObjMythArcheo database `http://www.limc-france.fr` and `http://medaillesetantiques.bnf.fr`, a corpus of archaeological objects related to mythological iconography, and BiblioNum, a DL about France in the 20th century.

[19] `https://collection.britishmuseum.org/sparql`.

```
# ID    FORM                SEGM              XPOSTAG
o.0.4   szuruppak{ki}-ga-ke4 Szuruppag[1]-ak-e SN.GEN.ERG

# ID    FORM                LEMMA         UPOSTAG XPOSTAG FEATS
o.0.4   szuruppak{ki}-ga-ke4 Szuruppag[1]  PROPN   SN      Number=Sing|Case=Gen.Erg|Animacy=Nhum
```

Figure 1: CDLI-CoNLL annotation compared to CoNLL-U

us to seamlessly integrate LOD resources to process, manage, and manipulate CoNLL data with off-the-shelf technologies (Chiarcos and Schenk, 2018).

We argue for the use of CoNLL-RDF in our setting because of its suitability for LLOD integration. In fact, it is directly processable with Semantic Web technology insofar as it facilitates interoperability, interpretability, linkability, queryability, transformability, database support, and integration with web technologies. In the context of our corpus annotation workflow, CoNLL-RDF is used as an internal format for parsing and for the pre-annotation of syntax using SPARQL (Buil Aranda et al., 2013), cf. Section 3.3., but it can also serve as a future release format, cf. Mazziotta (2010) for Old French.

3. Annotation

3.1. Annotation Workflow

As explained in the corpus section (2.), the raw data entering the pipeline comprise unannotated textual data in the ATF format. Before conversion, this text is validated against structure rules and content. Structure is defined in the ATF format[20] specifications. Content is checked for word tokens and sign tokens against the existing data available at the CDLI.

When entering the pipeline, the text is first converted from ATF to the CDLI-CoNLL format. Like most members of the CoNLL format family, this is a TSV format with one word per line, newline-separated sentences. In comparison to, e.g., the widely used CoNLL-U format, it does come with project-specific columns. It is both more compact and more informative, but tailored to our specific use.

The CDLI-CoNLL file is then fed into morphological pre-annotation. A dictionary-based pre-annotation tool fills most of the morphological information for each form present in the text. The human editor goes over the result, filling the lines left incomplete, and verifying that the annotations are correct. Before storing the annotated CDLI-CoNLL text alongside the ATF text in the database, the content is again validated, both for content and conformity to the CDLI-CoNLL format. The resulting CDLI-CoNLL data are then stored in the database.

For syntactic parsing, the CDLI-CoNLL data are subject to the syntax pre-annotation tool described below, cf. Section 3.3.. The resulting data are serialized as CoNLL-U, but part of the conversion process is to replace CDLI-specific annotations with those conforming to the Universal Dependencies. For this purpose, we provide and consult an OWL representation of the CDLI annotation scheme and its linking with UD POS, feature and dependency labels. Using SPARQL update, these ontologies are loaded, their hierarchical structure traversed by property paths, and the corresponding tags replaced. We argue that the clear separation of (SPARQL) code and (OWL) data of different provenience (CDLI annotation model, UD annotation models, linking between both) facilitates the transparency, reproducibility and reversibility of our mapping in comparison to direct replacement rules.

Finally, the CoNLL-U data are converted to the Brat standoff format[21]; the human editor can thus verify and finalize the syntactic annotation of the text using the CDLI Brat server interface.

The completed Brat Standoff file is exported and converted back to CDLI-CoNLL. At this point, the novel annotations need to be merged with the original CDLI data. Although conflicts should not occur as long as the data was not *manually* manipulated, we need a robust merging routine in case such corrections have been applied. For this purpose, we employ CoNLL-Merge.[22] CoNLL-Merge performs a word-level diff on the FORM column. Beyond merely identifying mismatches, it also provides heuristic but robust merging strategies in case a mismatch occurred, e.g., if a word has been split, two words have been merged, or deletions or additions occurred.

Only the ATF and CDLI-CoNLL versions of the data are kept in the datastore as we can easily convert the CDLI-CoNLL format to CoNLL-U and CoNLL-RDF formats, according to need. While both will be important publication formats to facilitate usability and re-usability of our data, they will only be generated on demand. We are, however, exploring options to offer CoNLL-RDF as a dynamic view on the internal (relational) database via technologies such as R2RML (Das et al., 2012).

An illustration of the annotation workflow, including intermediate data formats, is shown in Fig. 2[23]

3.2. Morphological Pre-Annotation

As part of the pipeline, we have designed a morphological pre-annotation tool[24] to make the manual annotation process more efficient in respect to speed of annotation as well as consistency and actual morphological analysis correct-

[20] http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/index.html.

[21] http://brat.nlplab.org/standoff.html.

[22] https://github.com/acoli-repo/conll.

[23] Cuneiform text of the Ur III period from the settlement of Garshana, Mesopotamia (Owen, 2011, no. 851) and its transliteration as stored in the Cuneiform Digital Library Initiative (CDLI) database https://cdli.ucla.edu/P322539 (picture reproduced here with the kind permission of David I. Owen).

[24] The code for this tool and all the other tools we are designing for this pipeline are available in repositories kept under the CDLI organization page on Github https://github.com/cdli-gh.

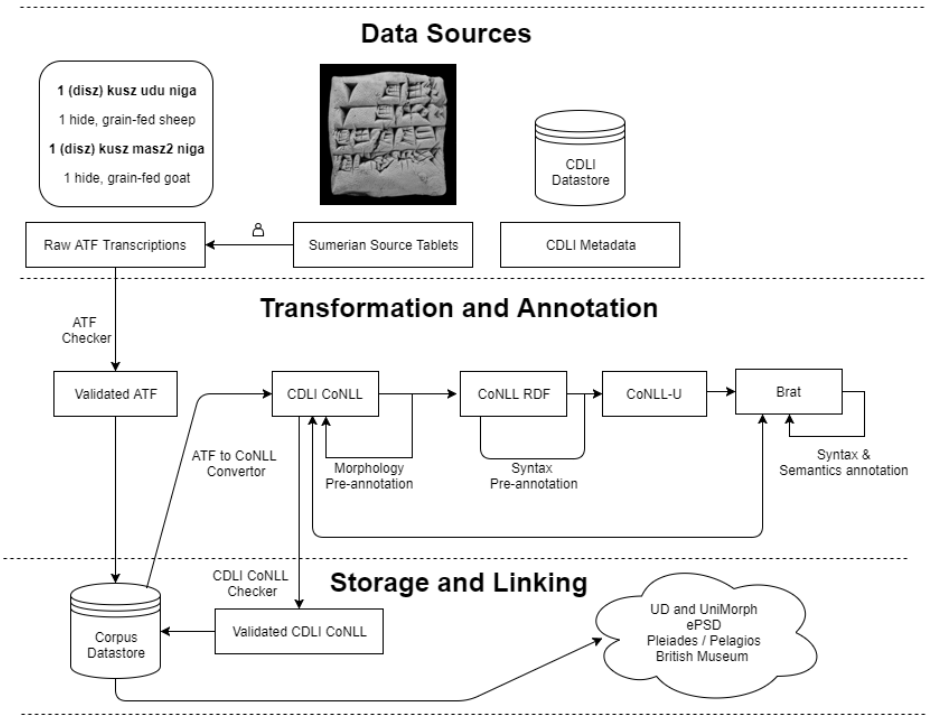

Figure 2: Corpus annotation pipeline: from ATF to RDF

ness. The tool is applied after the ATF text has been converted to the CDLI-CoNLL format. It uses the principle of a dictionary lookup to provide the most frequent annotation associated with the form it is annotating. For example, a text could contain the form ensi2 ("ruler"), without attached morphemes. All variant analyses of the form encountered to date, and their frequency, are stored in the dictionary.

```
"ensi2": [
  {
    "annotation": [
      "ensik[ruler]",
      "N"
    ],
    "count": 3
  },
  {
    "annotation": [
      "ensik[ruler][-ak][-ø]",
      "N.GEN.ABS"
    ],
    "count": 1
  },
  {
    "annotation": [
      "ensik[ruler][-ø]",
      "N.ABS"
    ],
    "count": 1
  },
  {
    "annotation": [
      "ensik[ruler][-ra]",
      "N.DAT-H"
    ],
    "count": 1
  }
],
```

When the pre-annotation tool encounters the form ensi2, it will add the first option of the example in the appropriate SEGM and XPOSTAG fields, based on frequency. The other choices are appended in subsequent columns so the human editor can easily copy and paste another option in the appropriate fields, if required. The additional columns will be destroyed while validating the contents. The pre-annotation tool can add new entries to the dictionary on demand, so it is best to perform this operation frequently to augment the accuracy of the tool.

The CDLI-CoNLL validator is integrated into the pre-annotation morphological tool. It performs checks on the syntax of the ID field, the existence of the lemmata in our dictionary, and the parallelism of the SEGM and XPOSTAG fields, based on a mapping of morphemes that can appear in the SEGM field and the morphological tags employed in the XPOSTAG field.

3.3. RDF-Based Pre-Annotation

CoNLL-RDF has been developed with the goal of flexible transformation of annotated corpora for the output of state-of-the-art NLP tools: every CoNLL sentence constitutes a graph, and parsing rules can be formulated as rewriting operations on this graph.

While this approach is qualitatively different from conventional parsing, we adopt the terminology of classical Shift-Reduce parsing (Nivre et al., 2007, 100-104). However, we model SHIFT and REDUCE as RDF properties that result from parsing operations rather than these parsing operations themselves. Along with that, parsing is no more sequential, and data structures such as QUEUE and STACK are no longer necessary; instead, both the 'queue' of tokens and the 'stack' of partial parses are marked by explicit SHIFT relations that represent their sequential order.

The method is initialized by adding a SHIFT relation for ev-

ery *nif:nextWord* property in the graph, i.e., the 'queue' of partial parses corresponds to the sequence of words. During parsing, language-specific rules are applied. Unlike classical Shift-reduce parsing, the words are not processed from left to right, but bottom-up. If an attachment rule applies for a word/partial parse X, it is removed from the 'queue' of words (which is no longer distinguished from the 'stack' of partial parses) by dropping its SHIFT relations. Instead, a REDUCE relation with its head is established, and the sequence of SHIFTS is restored by connecting the head of the partial parse with its SHIFT-precedent, or successor.
With any remaining SHIFT relations of the reduced elements being transferred to the (partial) parse, the sequence of SHIFTs takes over the functions of the traditional 'queue' and the traditional 'stack' at the same time, but elements are processed regardless of their sequential order; instead, the order of parsing rules plays a decisive role in the parsing process.
Parsing rules can be expressed as SPARQL Update statements, which are applied and iterated in a predefined order until there are no more transformations, i.e., because a single root for the sentence has been established. Finally, the SHIFT transitions are removed, whereas the REDUCE transitions are replaced by *conll:HEAD* properties.
Parsing, as defined here, is deterministic and greedy, and more or less context-insensitive. However, this is enough to provide a convenient means of implementing 'default' rules for syntactic attachment, which can be corrected afterwards during manual annotation.
In this sense, our basic rule-based parser provides a satisfactory syntactic *pre*-annotation with only 7 rules[25]:
1. Reduce adjective to preceding noun with adjectival modifier relation:
$\text{NOUN}_0\ \text{ADJ} \Rightarrow \text{NOUN} \xleftarrow{\text{amod}} \text{ADJ}$
E.g. nita $\xleftarrow{\text{amod}}$ kalag-ga "strong male".

2. Reduce noun in the genitive to preceding noun with appositional modifier relation:
$\text{NOUN}\ \text{NOUN}_{\text{GEN}} \Rightarrow \text{NOUN} \xleftarrow{\text{GEN}} \text{NOUN}$
E.g. lugal $\xleftarrow{\text{GEN}}$ $\text{urim}_5{}^{\text{ki}}$-ma "king of Ur".

3. Reduce noun with case marker to preceding noun with no case marker with appositional modifier relation:
$\text{NOUN}_0\ \text{NOUN}_{\text{CASE}} \Rightarrow \text{NOUN}_{\text{CASE}} \xleftarrow{\text{appos}} \text{NOUN}$
E.g. $^{\text{d}}\text{inana}_{\text{DAT}} \xleftarrow{\text{appos}}$ nin-a-ni "to Inanna, his lady".

4. Reduce noun to preceding noun with case relation:
$\text{NOUN}_0\ \text{NOUN}_{\text{CASE1+CASE2}} \Rightarrow \text{NOUN}_{\text{CASE1}} \xleftarrow{\text{CASE2}} \text{NOUN}$
This rule is applicable mostly for complex genitive chains.
E.g. $\text{lugal}_{\text{ERG}} \xleftarrow{\text{GEN}} \text{urim}_5{}^{\text{ki}}\text{-ma-ke}_4$ "king of Ur".

5. Reduce noun to preceding numeral with numeral modifier relation:
$\text{NUM}_0\ \text{NOUN}_{\text{(CASE)}} \Rightarrow \text{NUM}_{\text{(CASE)}} \xleftarrow{\text{nummod}} \text{NOUN}$
E.g. 3(u) $\xleftarrow{\text{nummod}}$ sila_3 "thirty sila (measuring unit)"

6. Reduce noun in case to following verb with absolutive relation:
$\text{NOUN}_{\text{ABS}}\ \text{VERB} \Rightarrow \text{NOUN} \xrightarrow{\text{ABS}} \text{VERB}$
E.g. numun-na-ni $\xrightarrow{\text{ABS}}$ he_2-eb-til-le-ne "may they end his lineage".

7. Reduce noun in case to following verb with case relation:
$\text{NOUN}_{\text{CASE}}\ \text{VERB} \Rightarrow \text{NOUN} \xrightarrow{\text{CASE}} \text{VERB}$

In part, these rules employ grammatical case features as dependency labels. After pre-annotation, however, these internal labels are to be mapped to CoNLL-U relations.
The graph-rewriting rules are implemented in SPARQL Update,[26] as illustrated by the example below, which matches the noun in the absolutive case to verb reduction rule (No. 6).

```
DELETE {
        ?last conll:SHIFT ?noun.
        ?noun conll:SHIFT ?verb.
} INSERT {
        ?noun conll:REDUCE ?verb; conll:EDGE ?case.
        ?last conll:SHIFT ?verb.
} WHERE {
        ?noun conll:POS ?pos FILTER(strends(str(?pos),'N')).
        ?noun conll:CASE ?case FILTER(?case="ABS")
        ?noun conll:SHIFT ?verb.
        ?verb conll:POS ?vPos FILTER(strstarts(str(?vPos),'V'))
        OPTIONAL {?last conll:SHIFT ?noun. }
}
```

Figure 3: SPARQL query for rule 6

An example of the output of the syntactic pre-annotation for a Sumerian royal inscription of Ur-Namma of Ur (approx. 2112-2095 B.C.)[27] is provided below.
We estimate that this method can be efficiently used for pre-annotation in order to enhance the syntactic annotation process; however, one cannot fully rely on its unsupervised result: mistakes and ambiguities are expected and these have to be resolved manually.

3.4. Manual Annotation

Manual annotation of the syntax is greatly simplified with the application of the pre-annotation tool. Using our Brat server[28], the human annotator must first verify that annotations generated by the pre-annotation tool are correct. When an annotation is faulty, the annotator removes the annotation and creates the appropriate one instead. Navigating the Brat interface is made easy as we modified the GUI to necessitate fewer clicks for each task. Finally, missing relationships must be added. The pre-annotation tool is

[25] Abbreviations follow Universal Dependencies; SHIFT and REDUCE relation are designated by whitespace (left) and arrow (right) respectively.

[26] The full code is available from https://github.com/cdli-gh/mtaac_work/tree/master/parse.

[27] See http://oracc.museum.upenn.edu/etcsri/Q000937.

[28] http://brat.nlplab.org/.

```
s1_1  . . / DAT-H---- ang          BASE an GW 1 ID 1 MORPH2 N1=NAME POS DN
s1_2  . . . \ appos-- lugal        BASE lugal GW king ID 2 MORPH2 N1=STEM POS N
s1_3  . . . . \ GEN-- dinjir-re-ne BASE dinjir GW deity ID 3 MORPH2
                                   N1=STEM.N4=PL.N5=GEN POS N
s1_4  . . . \ appos-- lugal-a-ni   BASE lugal GW king ID 4 MORPH2
                                   N1=STEM.N3=3-SG-H-POSS.N5=DAT-H POS N
s1_5  . . / ERG------ ur-{d}namma  BASE ur-{d}namma GW 1 ID 5 MORPH2 N1=NAME
                                   POS RN
s1_6  . . . \ appos-- lugal        BASE lugal GW king ID 6 MORPH2 N1=STEM POS N
s1_7  . . . . \ GEN-- urim5{ki}-ma-ke4 BASE urim5{ki} GW 1 ID 7 MORPH2
                                   N1=NAME.N5=GEN.N5=ERG POS SN
s1_8  . . / ABS------ kiri6        BASE kiri6 GW orchard ID 8 MORPH2 N1=STEM POS
                                   N
s1_9  . . . \ amod--- mah          BASE mah GW great ID 9 MORPH2
                                   NV2=mah.N5=ABS POS V/i
s1_10 . . \           mu-na-gub    BASE gub GW stand ID 10 MORPH2
                                   V4=VEN.V6=3-SG-H.V7=DAT.V11=3-SG-H-A.V12=gu
                                   b.V14=3-SG-P POS V/i

s1_11 . . / ABS------ barag        BASE barag GW dais ID 11 MORPH2
                                   N1=STEM.N5=ABS POS N
s1_12 . . / L2-NH---- ki           BASE ki GW place ID 12 MORPH2 N1=STEM POS N
s1_13 . . . \ amod--- sikil-la     BASE sikil GW pure ID 13 MORPH2
                                   NV2=STEM.N5=L2-NH POS V/i
s1_14 . . \           mu-na-du3    BASE du3 GW build ID 14 MORPH2
                                   V4=VEN.V6=3-SG-H.V7=DAT.V11=3-SG-H-A.V12=ST
                                   EM.V14=3-SG-P POS V/t
```

Figure 4: Syntactic pre-annotation of Ur-Namma 5

improved from the feedback of the human annotators along the way. Generally, annotations will be correct as they are created using the rules described in 3.3.; more complex cases are not covered by the rules, so they are to be created by the annotator. Figure 5 shows a screenshot of three examples of relationships between words. Clicking on one term and then another one opens up a panel for choosing the nature of the relationship and creates it on confirmation; selecting a word or a relationship and pressing DEL removes the annotation.

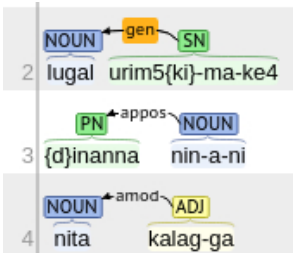

Figure 5: Brat annotation example

3.5. Linking Resources

As described above, morphosyntactic and syntactic annotations of the CDLI corpus have been linked with models of UD parts-of-speech, features and dependency labels, and this information is actively used during syntactic pre-annotation. To facilitate interpretability of our data, it can also be provided as part of the RDF edition of the annotated CDLI corpus.

In addition, morphological features have also been defined in language-independent terms, by linking the existing CDLI/ETCSRI morphological annotation scheme[29] with an ontological model of the UniMorph specifications available as part of the OLiA ontologies[30]. This effectively positions Sumerian among the language corpora that are linked by their linguistic annotations, and employing this schema will also facilitate translation since Unimorph is able to define morphological features in language-independent terms (Sylak-Glassman, 2016, 3). The only digital resource for Sumerian vocabulary is the ePSD[31]. We have prepared an index of deep links, modeled as a lemon dictionary[32]. Until the proper integration of Linked Data into the anticipated upcoming ePSD2 edition, this acts as a placeholder. Despite being a preliminary resource, at best, this index, comprising lemon-compliant lexical entries, forms and senses, already serves to illustrate linking with lexical resources. Additional local lexical resources will abe provided as we prepare the Ur III research corpus.

The CDLI catalog provides metadata on objects bearing cuneiform inscriptions which, especially when integrated into the text analysis method, can prove to be useful for the discovery and study of the artifact. It is stored in a MySQL database and is exported daily in CSV format. We convert the data to RDF with the csv2rdf tool[33] supplemented with embedded custom turtle templates, and link to external metadata repositories: the Modref project and the British Museum.

4. Discussion and Outlook

4.1. Limits of Morphological Pre-Annotation

The first limitation of morphological pre-annotation concerns word identification. Since a word can have different meanings, identifying the right one requires an awareness of the context. The same problem occurs when dealing with forms where case markers were not inscribed; they must be inferred based on the analysis of the whole sentence, or in the case of the Ur III administrative texts, the order of words, since it is often stereotyped. To counteract those limitations, the human annotator analyses the text and corrects and refines the generated annotations.

Because of the sheer quantity of the texts to annotate, semi-automated annotation using the morphological pre-annotation tool coupled with the input of an annotator to prepare all texts is not feasible. As discussed elsewhere, we are developing a machine-learning pipeline for the automated annotation and translation of texts, based on the translation and annotations prepared to form the required gold corpus, using the method described in section 3.1.

4.2. Limits of Syntactic Pre-Annotation

The implementation of syntactic pre-annotation is not a fully-featured parser, but a simple deterministic and greedy algorithm to assist manual annotation. This process, based on 'default rule', allows us to automatically pre-annotate *most* of the material and then correct it, rather than manually annotate everything from scratch.

[29] `http://oracc.museum.upenn.edu/etcsri/parsing/index.html`.

[30] `purl.org/olia/owl/experimental/unimorph/`, also cf. `http://unimorph.org/`.

[31] `http://psd.museum.upenn.edu/`.

[32] Our index is hosted on the Oracc server, home of the ePSD: `http://oracc.museum.upenn.edu/ttl/epsd1.ttl`.

[33] `http://clarkparsia.github.io/csv2rdf/`.

Some examples in which the syntactic pre-annotation analysis will likely be incorrect, are presented below (from Jagersma (2010)):

1. Nominal clause. Clauses that does not contain an independent verbal form might not be parsed correctly in some cases, e.g.:

   urdu$_2$ lu$_2$-še lugal-zu-u$_3$
   urdu2.d lu$_2$ =še =Ø lugal =zu =Ø
   slave man=that=ABS master=your=ABS
   'Slave! Is that man your master?'
   (Jagersma, 2010, 716, no. 7)

2. Word order. Sumerian normally has a SOV word order, with the verb at the final position. However, exceptional right-dislocated clauses are known, e.g.:

   i$_3$-ĝu$_{10}$ i$_3$-gu$_7$-e d nisaba-ke$_4$
   ì =ĝu =Ø 'i -gu$_7$-e nisaba.k=e
   fat=my=ABS VP-eat -3SG.A:IPFV Nisaba =ERG
   'She will eat my cream, Nisaba.'
   (Jagersma, 2010, 300, no. 27)

   Clause boundaries will not be correctly recognized in such cases.

3. Enclitic copula. The Sumerian copula *me* can be both independent and enclitic. In the latter case the analysis of the token in context of other words is ambiguous, as it contains both nominal and verbal annotation, e.g.:

   še dub-sar-ne-kam
   še dub.sar=ene=ak =Ø ='am
   barley scribe =PL =GEN=ABS=be:3N.S
   'This is barley of the scribes.'

   nagar-me-eš$_2$
   nagar =Ø =me-eš
   carpenter=ABS=be -3PL.S
   'They are carpenters.'
   (Jagersma, 2010, 681-2, nos. 24 and 27)

4. Enclitic possessive pronouns and dimensional prefixes. To facilitate subsequent dependency parsing, enclitic possessives are analyzed in terms of their *morphosyntactic* characteristics, not on grounds of their *semantics*: In their function, enclitic possessives are referential and this could be explicitly expressed with explicit links between possessor and possessum within UD using the language-specific but popular `nmod:poss` relation. However, such links cannot be easily integrated into UD-compliant syntactic annotation as it may easily lead to non-projective trees (i.e., crossing edges):

   sipa-de$_3$-ne / gu$_2$-ne-ne-a / e-ne-ĝar
   sipa.d =enē=r(a) gu$_2$ =anēnē='a 'i -nnē -n -ĝar -Ø
   shepherd=PL =DAT neck=their =LOC VP-3PL.OO-3SG.A-place-3N.S/DO
   'He placed this (as a burden) on the shepherds, on their necks.'
   (Jagersma, 2010, 686, no. 21a) In this example, the locative argument syntactically depends on the verb; at the same time, the enclitic possessive (glossed as 'their') refers to the preceding argument. Therefore, these semantic relations are to be captured in a subsequent processing step akin to anaphor resolution in other languages.

This incomplete list gives examples of cases where the analysis by the pre-annotation tool would be incorrect at this time in the development of the tool. But the bulk of these grammatical elements occur very rarely in Ur III administrative texts and royal inscriptions. Still, the pre-annotation algorithm will be extended with more elaborate rules in the future to improve its performance and to incorporate more complex features and constructions since we aim to make this tool useful to annotate all genres of the Sumerian language.

4.3. Conclusions

The workflow that brings ATF raw textual data to publication as Linked Open Data, and the pipeline for text annotation–in particular the annotation of morphology and syntax–described in this paper, draws a roadmap for further development in the processing and analysis of ancient cuneiform languages. Improving and automating the annotation process for Sumerian sources is foundational for future work on cuneiform corpora, while the generation of annotations using a semi-automated annotation process for Sumerian syntax is generally unprecedented and innovative. We find the implementation of new standards for Assyriology as a digital discipline hardly meaningful without compatibility with existing LLOD standards on the one hand, and their adaptation to the particular languages and the material under scrutiny on the other, hence the choice of the CoNLL formats, RDF, UD, and the CIDOC-CRM. Building the machine translation pipeline for Sumerian, the ultimate goal of the MTAAC project, is greatly dependent on this work.
These altogether are crucial steps towards LLOD editions of Sumerian and other cuneiform languages. We hope that our work will help to provide Assyriogists and researchers from other fields with new open access annotated textual datasets, and reusable infrastructure that can significantly contribute to the study of ancient languages and cultures.

Acknowledgements

The Machine Translation and Automated Analysis of Cuneiform Languages project is generously funded by the German Research Foundation, the Canadian Social Sciences and Humanities Research Council, and the American National Endowment for the Humanities through the T-AP Digging into Data Challenge.[34]
Our appreciation goes to Heather D. Baker and Robert K. Englund for their insights and suggestions.

[34] `https://diggingintodata.org/`.

5. Bibliographical References

Black, J. A., Cunningham, G., Ebeling, G., Flückiger-Hawker, J., Robson, E., Taylor, J., and Zólyomi, G. (1998–2006). The Electronic Text Corpus of Sumerian Literature. http://etcsl.orinst.ox.ac.uk.

Buil Aranda, C., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., Harris, S., Hawke, S., Herman, I., Humfrey, N., Michaelis, N., Ogbuji, C., Perry, M., Passant, A., Polleres, A., Prud'hommeaux, E., Seaborne, A., and Williams, G. (2013). SPARQL 1.1 overview. https://www.w3.org/TR/sparql11-overview.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge*, pages 74–88. Springer.

Chiarcos, C. and Schenk, N. (2018). The ACoLi CoNLL Libraries: Beyond tab-separated values. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards Open Data for linguistics: Linguistic Linked Data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

Das, S., Sundara, S., and Cyganiak, R. (2012). R2RML: RDB to RDF mapping language. Technical report.

Hruška, B. (1980). Drei neusumerische Texte aus Drehem. *Die Welt des Orients*, 11:27.

Jagersma, A. H. (2010). *A descriptive grammar of Sumerian*. Ph.D. thesis, Faculty of the Humanities, Leiden University.

Jaworski, W. (2008). *Ontology-based knowledge discovery from documents in natural language*. Ph.D. thesis, Warszawa: Uniwersytet Warszawski.

Mazziotta, N. (2010). Building the Syntactic Reference corpus of Medieval French using NotaBene RDF annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 142–146. Association for Computational Linguistics.

Molina, M. (2017). Syntactic annotation for a Hittite corpus: problems and principles.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):9 5–135.

Nurmikko-Fuller, T. (2015). *Telling ancient tales to modern machines: ontological representation of Sumerian literary narratives*. Ph.D. thesis, University of Southampton.

Nurmikko, T. (2014). Assessing the suitability of existing OWL ontologies for the representation of narrative structures in Sumerian literature. *ISAW Papers*, 7(1):1–9.

Owen, D. I. (2011). *Garshana studies*. CDL Press.

Pagé-Perron, É., Sukhareva, M., Khait, I., and Chiarcos, C. (2017). Machine Translation and Automated Analysis of the Sumerian Language.

Smith, E. (2010). *Query-based annotation and the Sumerian verbal prefixes*. Ph.D. thesis, University of Toronto.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (Unimorph schema). Technical report, Technical report, Department of Computer Science, Johns Hopkins University.

Tchienehom, P. L. (2017). ModRef project: from creation to exploitation of CIDOC-CRM triplestores. In *The Fifth International Conference on Building and Exploring Web Based Environments (WEB 2017)* , Barcelona, Spain, May.

Electronic Dictionaries – from File System to *lemon* Based Lexical Database

Ranka Stanković, Cvetana Krstev, Biljana Lazić, Mihailo Škorić
{Faculty of Mining and Geology, Faculty of Philology } University of Belgrade
{Djušina 7, Studentski trg 3} Belgrade, Serbia
{ranka.stankovic, biljana.lazic, mihailo.skoric}@rgf.bg.ac.rs, cvetana@matf.bg.ac.rs

Abstract

In this paper we present our approach for lexical data migration from textual e-dictionaries to a lexical database. After years of development, Serbian Morphological Dictionaries (SMD), developed as a system of textual files, have become a large and complex lexical resource. As a consequence, LeXimir, the application that has been used for SMD development and management, was no more suitable. We thus started developing an on-line application for dictionary development and management, based on a central lexical data repository (lexical database). In this paper we present the model for the SMD lexical database developed following the *lemon* model, and the thesaurus of data categories, to be used for enabling links to other (lexical) data. The new database offers various possibilities for improvement of SMD, e.g. control of data consistency and introduction of explicit relations between lexical entries. Besides the procedure used for mapping the existing data model to the new one, we present sets of rules developed to establish relations between lexical entries. We also present some additional improvements – automatic generation of dictionary candidates, with their lexical and derivation variants. This automatic procedure enabled migration of all 26 simple word and 15 multi-word unit Serbian dictionary files with more than 150,000 lexical entries.

Keywords: lexical database, lemon, electronic dictionaries, lexical model, lexical relations

1. Introduction

An application dubbed WS4LR (Krstev et al., 2006), subsequently upgraded and renamed LeXimir (Stanković, Ranka and Krstev, Cvetana, 2016), was designed and implemented for the purpose of further development and management of morphological electronic dictionaries of Serbian (SMD), presented in more details in Section 3.. However, with the growing number of dictionary developers, and given the variety of dictionaries and information stored in them (proper names, domain-specific terms, etc.), the need arose for a more robust application. The main shortcoming of LeXimir, being a desktop application, was that dictionary updates by one user could not be synchronized with other users in real time. Thus, we decided to develop a web application for dictionary management, and enhance the development environment from singe-user to multi-user. In addition to that, LeXimir did not offer support for complex constraints that the development of large dictionaries with rich information needs. The format used in LeXimir did not support the establishment of relations between lexical entries, nor cross-linking with other lexical models, such as Serbian WordNet, another important lexical resource for Serbian (Koeva et al., 2008). This was the main motivation for transforming SMD dictionaries from the existing file system to a *lemon* based lexical database. The model for this lexical database was developed in compliance with the state of the art standards for lexical resources. In this paper we describe how the lexical database was designed following the *lemon* model. We also present how dictionaries were automatically improved and enriched by introducing new lexical entries and/or lexical relations, and by checking the existing ones.

An NLP lexicon has little in common with human-oriented e-dictionary. Data structures in these two types of e-dictionaries are quite different. However, it proved to be very useful to use NLP applications and components in human-oriented e-dictionaries. There are also some NLP-lexicons that can be used by humans. One of such examples is WordNet. A growing number of e-dictionaries pinpointed the need for data standardization, interchange and reusability. In addition to that, the development of the Semantic Web emphasized the importance of enriching ontologies with lexical information. These developments motivated the NLP community to join efforts in standardization. The resulted are widely-used guidelines and standards for dictionary description and lexical databases such as TEI (Tutin and Véronis, 1998), LexInfo (Cimiano et al., 2011), LMF (Francopoulo, 2013), *lemon* (McCrae et al., 2011) etc. The *lemon* model was implemented in several well-known and widely used resources (BabelNet, DBpedia, WordNet), proving that it can be useful in bringing complementary lexical resources together within a single framework.

2. Related work

In order to develop a concrete and general model of dictionaries, it is essential to distinguish between the formal model itself and the encoding or database schema that may ultimately instantiate it (Ide et al., 2000). Having in mind interoperability and standardization issues, three options for the lexical model were considered. The first one were TEI (Text Encoding Initiative) Guidelines for dictionary description. TEI is a widely accepted standard for text encoding that proposes solutions for many text types, one of them being dictionaries. However, it seems that TEI is more often used for traditional human-oriented digitized dictionaries (Khemakhem et al., 2017, Bański et al., 2017).

The second option considered was the LMF (Lexical Markup Framework) model, as it pays special attention to language resources interoperability and re-usability. It provides description of lexical objects, including morphological, syntactic and semantic aspects (McCrae et al., 2012).

This model offers special solutions for the description of lexical information that is used in NLP. Many papers present examples of converting different lexical resources, such as monolingual (Attia et al., 2010) and bilingual (Maks et al., 2008) lexicons, to LMF based multifunctional and reusable electronic lexical databases. LeXimir provided for export of e-dictionaries to XML files compliant to LMF model, but further exploitation of these files was not implemented, neither for lexical database development nor for further processing (Stanković et al., 2013).

Finally we considered the *lemon* model (Lexicon Model for Ontologies), which was derived from LMF, and has been designed for ontology lexicons on the Semantic Web. It is aimed at enriching the conceptualization represented by a given ontology by means of a lexico-terminological layer (McCrae et al., 2012). In order to enable sharing on the semantic web, and for interface with tools *lemon* is based on RDF. Its semantic modeling is more lightweight than that of LMF. One of the advantages is that grammatical annotations are obtained by the use of separate linguistic description ontologies (ISOcat (Kemps-Snijders et al., 2008), GOLD (Farrar and Langendoen, 2003), LexInfo (McCrae et al., 2011)).

The *lemon* approach has been successfully used for comprehensive NLP resources (Bosque-Gil et al., 2016, Villegas and Bel, 2015). The *lemon* model was also implemented in well-known resources such as BabelNet and DBpedia. A paper dealing with WordNet conversion to lemon model (McCrae et al., 2011) demonstrated that *lemon* is an interchange format that can be used to bring complementary lexical resources together under a single framework. The main advantage of the *lemon* model for the research outlined in this paper was its support for linking with other (lexical) data and the possibility to access data by using the standardized SPARQL query language. The model presented is based on the *lemon* model, but some modifications and extensions were necessary to enable full migration of complex grammatical structures and numerous inflected forms for Serbian. MULTEX-East lexicons (Krstev et al., 2004) represent another important NLP lexical resource for Serbian, besides Serbian WordNet. However, both of them are not comparable with SMD either in size or in content, which is why SMD was chosen as the first lexicon for Serbian to be converted into a lexical database.

3. Morphological electronic dictionaries

Morphological electronic dictionaries of Serbian for NLP are being developed for many years now (Vitas et al., 1993) (Krstev, Cvetana and Vitas, Duško, 2015). They cover general lexica, proper names (persons and toponyms), general knowledge (famous or fictitious persons, places and organizations), and domain terminology. For practical reasons they are kept in a number of files, according to different criteria.

These dictionaries are in the so-called DELA format: in the dictionary of lemmas each lemma is described in full detail, so that the dictionary of forms containing all necessary grammatical information can be generated from it, and subsequently used in various NLP tasks (Courtois and Silberztein, 1990). A dictionary of lemmas can contain simple-word lemmas (DELAS) or multi-word lemmas (DELACF), producing, respectively, a dictionary of simple-word forms (DELAF) or multi-word forms (DELACF).[1] Traditionally, dictionaries of lemmas are prepared and maintained as one or more textual files, while dictionaries of forms are generated automatically, also as textual files. The structure of a simple word lemma is:

```
lemma,POS#fst[+Marker]*
```

Mandatory parts of this structure are a lemma, its POS, and identification of a finite-state transducer that will produce all lemma's inflected forms with associated grammatical information (e.d. case, number, gender, etc.). Markers are not mandatory, but they are nevertheless assigned to the majority of lemmas. Formally, they can be of two types:

- *switches*: if a marker of this type is present, then it indicates that a lemma has a certain feature, but if it does not exist, that indicates the absence this feature for the lemma. For instance, the marker `+Hum` indicates that a lemma represents a human being (e.g. `profesorka,N661+Hum` – 'woman professor', as opposed to `krava,N601` – 'cow');
- *attribute/value pairs*: an attribute indicates the type of the feature, while a value makes it more specific. For instance, in the marker `+DOM=Math` the attribute `+DOM` indicates that a lemma is related to a certain domain, whereas the value specifies this domain to be mathematics (e.g. `diedar,N3+DOM=Math`). Values assigned to a certain attribute can belong to a closed set (e.g. `+CC2=RS` is a two character country code marker assigned, for instance, to geopolitical names), or to an open set (e.g. `+Val=Vaughn` is assigned to a surname *Von*, Serbian transcription of the English surname *Vaughn*).

Semantically, markers can be of various types:

- *semantic/ontology* – these markers denote lemmas as belonging to a certain ontological class, e.g. `+Hum` (humans), `+Body` (body parts), etc.;
- *syntactic* – these markers provide some syntactic information about a lemma, e.g. a marker `+Ref` assigned to a verb indicates it is a reflexive verb;
- *pronunciation* – these markers are assigned to lemmas specific to a certain pronunciation, e.g. `+Ek` for Ekavian, `+Ijk` for Ijekavian pronunciation;
- *derivation* – these markers are assigned to lemmas derived from other lemmas, e.g. the marker `+GM` assigned to *profesorka* 'woman professor' denotes that it is derived from *profesor* 'professor' by gender motion; [2]

[1] Serbian e-dictionaries, SMD, have reached a considerable size: they comprise more than 150,000 simple-word lemmas, generating more than 5 million forms and 18,000 multi-word lemmas.

[2] In the lexical database described in this paper these markers are converted from switch to attribute/value markers, e.g. `+DER=GM`.

- *variation* – these markers indicate that a lemma has a variant, and how this variation is produced. In Serbian, many words have lexical variants that do not bear any specific meaning – they may be preferred in certain regions or in certain period of time (Klajn, 2005, Stanojčić and Popović, 2008). For instance, *afirmisati* and *afirmirati* 'to establish' are two such variants, to which markers `+VAR=SatiRati` and `+VAR=RatiSati` are assigned, respectively;

- *domain* – these markers indicate the domain of use of lemmas to which they are assigned;

- *information* – these markers provide some additional information about a lemma, e.g. the lemma *deci*, shortened for 'deciliter', has a marker `+SI=dl` assigned to it, indicating that its abbreviation in the International System of Units is *dl*.

Relations can exist between certain markers. For example, the hyperonymy/hyponymy relation exists between semantic markers: river (`+River`), which is a hydronym (`+Hyd`), which is a geographic concept (`+Top`), and thus all three are assigned to the lemma *Dunav* 'Danube'. Some lemmas are related by some sort of "inverse" relation, which indicates that if one lemma has a certain feature, then at least one other lemma exists with an "inverse" feature. These relations are sometimes explicitly encoded by appropriate markers (e.g. variation and pronunciation markers presented before), while in most cases they are implicit. For instance, lemmas for *profesorka* and *profesorica*, both meaning 'woman profesor' are derived from *profesor*, and they both have a marker `+GM`, while lemma for *profesor* does not have a marker indicating that forms derived from it by gender motion exist.
All the entries in a DELAF dictionary of forms are in the following format:

```
form,lemma[:categories]*
```

where `form` is a simple word form of a `lemma`, represented by its DELAS entry form, and `:categories` are the possible grammatical categories of the word form, each category represented by a single character code (Krstev and Vitas, 2007).
LeXimir, a tool for development and maintenance of e-dictionaries enabled development of Serbian morphological dictionaries in the past decade. However, with the enhancement of dictionaries and enrichment of their content some serious drawbacks of this tool became evident. Besides being a desktop application, discouraging cooperative work, it also does not have appropriate support for the treatment of duplicates (e.g. should *atlas* be one lemma or two lemmas that have same inflectional behavior, one denoting a book with maps and having markers `+Conc+Text`, the other denoting a type of a fabric and having markers `+Conc+Mat`). The consistency check is missing as well (e.g. can a marker `+Hum` be assigned to a lemma whose grammatical category `q` indicates it is inanimate, like *lonac* 'pot'?), as well as a check establishing the correctness of "inverse" relations (e.g. does a variant lemma *duhan* indicated by the marker `+VAR=VH` assigned to a lemma *duvan* 'tobacco' exist?). Finally, the lack of all these features was an impediment to production of special purpose dictionaries: for instance, for purely morphological dictionaries, *atlas* should be one lemma, while for dictionaries aiming at semantic processing, two lemmas are necessary.

4. The Model and Implementation of the Lexical Database

The main goal of the research presented in this paper was to produce a central lexical repository that will enable multiuser distributed management of lexical data, overcoming the main problem of the existing solution – local, single-user editing of dictionaries in textual form. The new lexical database should also enable of its content export in various formats. The Unitex[3] format for DELA dictionaries (dictionaries of lemmas and dictionaries of inflected forms presented in the previous section), supported by LeXimir, will be only one of the formats supported by the lexical database. The database will also provide for automatic production of dictionary editions for different profiles of users: full dictionaries, public-domain oriented, filtered by different criteria (e.g. pronunciation: Ekavian and Ijekavian), etc.
In the new lexical database model for Serbian Morphological Dictionaries, based on the *lemon* model, main classes for lexical entries, morphological, syntactic and semantic features are controlled by the internal thesaurus of data categories, outlined in (Krstev et al., 2010). During the whole period of the development of Serbian morphological dictionaries, the corresponding metadata were documented by a simple textual file. This file was the base for the creation of a dictionary of markers, that is, data categories and their values (Figure 1). Transition to the database that supports the control of field domains revealed inconsistencies among markers: same markers used for different purposes, different markers used for the same purpose, missing markers, markers associated to wrong categories, etc. Presently, there are 23 semantic markers in the database (e.g. `+Hum` for human beings), 17 syntactic (e.g. `+Ref` for reflexive verbs), 24 grammatical (`V` for verbs), with a total of 836 different values. There are also special domain markers (at present 104), which relate the lexical entry (and a particular sense) to its domain of use. For instance, the lexical entry *jezik* 'language, tongue' has three different senses (presently recorded in SMD), and their textual representation in DELA format is:

```
jezik,N9+DOM=Ling//communication media
jezik,N9+Conc+Body+DOM=Anatomy//body part
jezik,N9+Conc+Food+Prod+DOM=Culinary//food
```

Each of these entries is connected to a different domain (linguistics, anatomy, and culinary, respectively).
Since the use of *lemon* is complemented with LexInfo, as an ontology of types, values and properties to be used with the *lemon* model (partially derived from ISOcat), one of the goals was to map categories used in existing SMD to LexInfo, as a catalog of data categories (e.g., to denote gender, number, part of speech, etc.).

[3]Unitex is a lexically-based corpus processing suite that offers strong support for finite-state processing using morphological dictionaries –`http://unitexgramlab.org/`

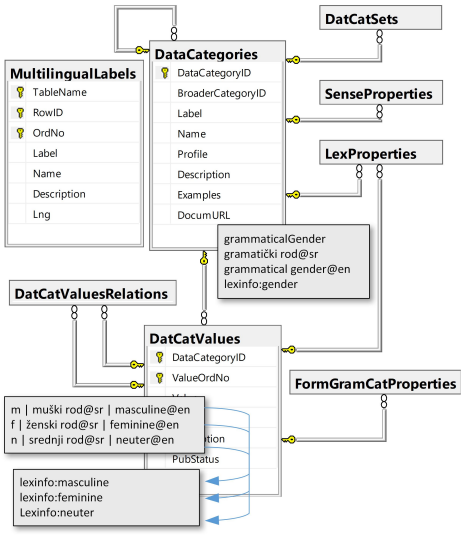

Figure 1: Data categories (markers) dictionary.

The main class of the core of the lexicon model is the class `LexicalEntry`, representing a unit of analysis of the lexicon, which encompasses a set of inflected forms that are grammatically related, and a set of base meanings that are associated with all of these forms (Figure 2). A lexical entry is a (single) word, multi-word expression, acronym or affix with a single part-of-speech, a morphological pattern, or a set of senses.

The `LexicalRelation` class relates lexical variants (for instance, *istorija* and *historija* 'history'), full forms and their abbreviation (e.g. *kilogram* and *kg*), derivationally related lexical entries (e.g. *istorija* and *istorijski* 'relating to the study of history'), and different pronunciations (Ekavian *dete* and Ijekavian *dijete* 'child'). `LexicalSense` is used to represent a particular sense of a lexical entry (e.g. for instance, three senses of *jezik*), and link a lexical entry with an ontology by connecting the set of markers denoting one sense with individual markers in the `SenseProperties` table.[4]

`SenseRelation` provides for connecting various senses with others that are narrower, broader, synonymous and so on, while `SenseRef` and `SenseExample` contain information about provenance and usage.

For languages with rich morphology, such as Serbian, the maintenance of dictionaries of inflected word forms is very important. For instance, inflected forms of *jezik* are: *jezik*, *jezika*, *jeziku*, *jeziče*, *jezikom*, *jezici*, *jezike*, *jezicima*. In the model presented, the table `Forms` is used to store all forms that are inflected from a lemma, together with sets of grammatical categories assigned. Since one lexical form can represent one or more grammatical realization of a lexical entry, it is described with one or more sets of grammatical categories stored in `FormGramCats`. For instance, the form *jezikom* has one set of grammatical categories assigned to it `:ms6q` (the instrumental case, singular), while two sets of grammatical codes are assigned to *jezika*: `:ms2q` and `:mp2q` (the genitive case, singular and plural). In addition, sets of grammatical categories are represented as individual categories in the table `FormGramCatProperties`, as presented in the left side of Figure 2.

The class `Forms` is used in the *lemon* model to indicate a non-semantic relationship between two lexical entries, for instance, cases when a term is derived from another term: "lexical" and "lexicalize". In the model presented, the class `LexicalEntry` is used for canonical forms of different variants, and the class `LexicalRelation` for relations between variants.

Dictionary production in different formats is also envisaged. For instance, compiled dictionaries to be used by Unitex, or textual inflected files to be further utilized by users. RDF serialization (e.g. Turtle, RDF/XML) is under development, and Linguistic Linked Open Data (LLOD) publishing will also be supported, while the same lexical database will be used for query expansion Web APIs used for information retrieval and indexing support. The master lexical data repository is stored in a relational database management system, but the use of triple-stores, e.g. graph databases Neo4J and DBGraph, is being investigated. The use of triple-stores will be read-only in this phase of development, and they will be used for querying and linking to external resources, while CRUD (create, read, update, delete) operations will remain in the relational system, given the required stability and the implementation experience so far.

5. Migration of Dictionary Data

The procedure for transferring data from existing dictionaries into the lemon-based model is integrated in the existing tool for dictionary management LeXimir, in order to support parallel development for a certain period of time, and to enable smooth transition of development environment. The database contains all currently used markers, but these markers have not a "flat" structure anymore, but rather a hierarchical structure that can serve as a controller for domains of some fields in a database.

As previously mentioned, DELAS dictionaries are distributed in more than 40 files for practical reasons, and information about the file a lemma comes from is stored in the `Lexicon` table for development purposes. Lemma entries from a DELAS dictionary are generally mapped to entries in `LexicalEntry` and `LexicalSense` (Figure 2), where a lemma, its POS, the inflectional class (governing production of all inflected forms) are stored in the `LexicalEntry` table, while associated markers – syntactic, semantic, domain and other – can be separated if needed. Identical lexical entries from DELAS sharing the same inflectional class are merged into one `LexicalEntry`, while their semantic markers indicating different senses are separated into more entries. For instance, in the new database `jezik,N9` is an entry in the table `LexicalEntry`, while associated markers that differentiate senses are recorded in the `LexicalSense` table. Entries that are part of a MWU, which is entered in the same tables `LexicalSense` and `LexicalEntry`, are

[4]The terms *class* and *table* are used respectively to indicate a model class and a physical table in a database.

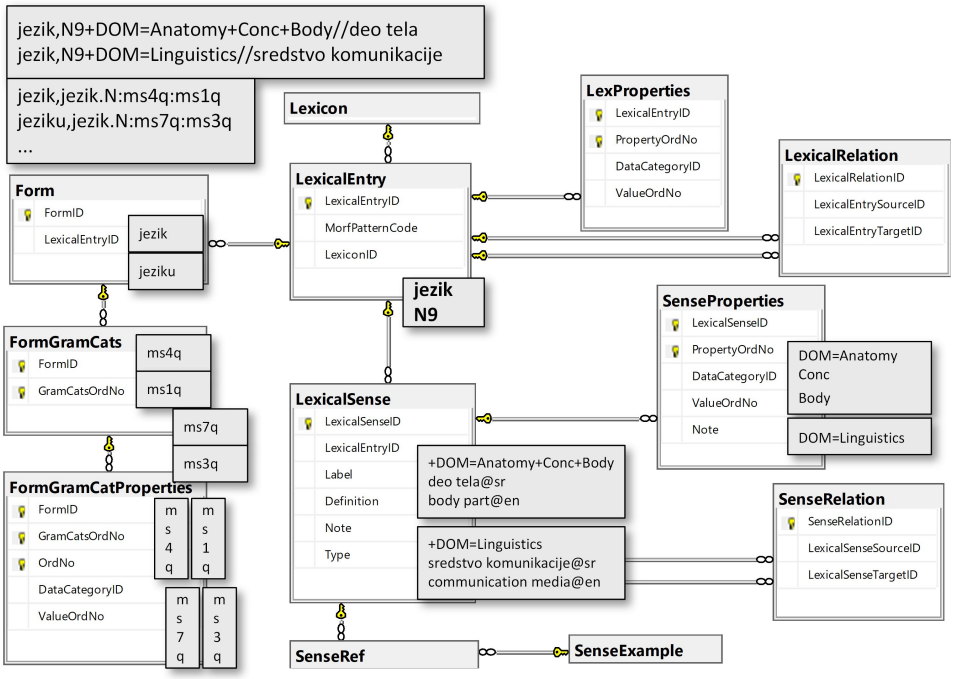

Figure 2: Lexical database core model.

related with the corresponding MWU. Examples of such entries are (simplified):

```
maternji jezik +DOM=Ling      'mother tongue'
jezik za zube  +DOM=Anatomy
     'tongue behind teeth (keep mouth shut)'
telećі jezik   +DOM=Culinary 'veal tongue'
```

The same example in the *lemon* form is:

```
lex_jezik a ontolex:LexicalEntry;
 lexinfo:partOfSpeech lexinfo:Noun;
 jezik ontolex:morphologicalPattern :N9;
 form_jezik ontolex:writtenRep "jezik"@sr;
 ontolex:canonicalForm :form_jezik;
 ontolex:sense :jezik_sense1;
 ontolex:sense :jezik_sense2;
 ontolex:sense :jezik_sense3.

:jezik_sense1 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Linguistics>;
 ontolex:reference
 <http://dbpedia..../Category:Language>.
:jezik_sense2 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Cooking>;
 ontolex:reference
  <http://dbpedia.../Tongue_(foodstuff)>.
:jezik_sense3 a ontolex:LexicalSense;
 dct:subject
  <http://dbpedia.org/page/Anatomy>;
 ontolex:reference
  <http://dbpedia.org/page/Tongue>.
```

Sense linking to WordNet synsets is planned, but is not yet implemented.

6. Dictionary improvement based on lexical variations and derivations

6.1. Corrections and Additions

The newly implemented lexical database (presented in Section 4.) introduced new possibilities for the improvement of valuable existing resources. Besides relatively trivial task of finding and correcting all incorrect markers (mostly typos), duplicate markers (denoting same concepts), it enabled the conversion of all markers that indicate links between lemmas (see Section 3.) into true relations between lexical entries. For instance, dictionary entry for *kućica* 'small house' had a marker for the diminutive `+Dem` assigned to it, but no indication of its basic form; at the same time, for the dictionary entry *kuća* 'house' it was not possible to determine whether it had a diminutive and if so, what it was.

Two approaches were used to establish relations between lexical entries. The first approach was used for explicit inverse relations, mostly for lexical variants or two different pronunciations, Ekavian and Ijekavian. In this approach one or more target lemmas are constructed based on the type of the relation, using some simple string matching and replacement, and the newly constructed lemmas had to (a) exist in dictionaries; and (b) have an inverse marker.

For instance, verbs *afirmisati* and *afirmirati* are two variants (the first one being preferred today in Serbian) of the same verb 'to establish'. Similarly, *hleb* and *leb* are two variants of the same noun 'bread' (the second one being

non-literary). The Ijekavian lemma for the Ekavian lemma *devojka* 'girl' is *djevojka*. These lemma pairs were recorded in DELAS entries of e-dictionaries, in the following manner:

```
afirmirati,V1+Imperf+Perf+Tr+Iref+Ref
           +VAR=RatiSati
afirmisati,V21+Imperf+Perf+Tr+Iref+Ref
           +VAR=SatiRati
hleb,N81+VAR=H0+Ek+Conc+Course+Food
     +DOM=Culinary
leb,N81+VAR=0H+Ek+Conc+Food+Prod
     +DOM=Culinary
devojka,N617+Hum+Ek
djevojka,N617+Hum+Ijk
```

The marker `+VAR=RatiSati` indicates that the suffix *-rati* can be substituted in the lemma *afirmirati* by the suffix *-sati* to produce the lemma *afirmisati*, which is recorded in e-dictionaries and has an inverse marker `+VAR=SatiRati` assigned to it. The marker `+VAR=H0` indicates that an *h* can be deleted in the lemma *hleb* to produce the lemma *leb*, which has an inverse marker `+VAR=0H` assigned to it in e-dictionaries. The marker `+Ijk` indicates that the reflection of an Old Slavic *yat* can be substituted in the Ijekavian lemma *djevojka* by *e* to produce the Ekavian lemma *devojka*, which has an inverse marker `+Ek` in e-dictionaries. It should be noted that for each lemma pair all other markers assigned to them are identical. Also, it is sometimes irrelevant which is the initial lemma used for producing the other lemma by substitution/deletion (the first example), while in some other cases one of the lemmas in a pair is a better initial choice (in examples above, lemma containing *h* for markers `+VAR=H0`/`+VAR=0H`, and Ijekavian lemma for markers `+Ek`/`+Ijk`). In the first two cases a *variation relation* is established between the pair of lexical entries, while in the third case it is a *pronunciation relation*. Namely, entries for which a lexical variant exists have a special marker in the form `+type=value` where, in this case, `+VAR` indicates a variation marker and `value` indicates a type of variation, which also gives a hint how one variant can be derived from another. As a rule, these relations should be inverse, as is the case with examples given above. With dictionaries maintained as textual files, one could rely only on a developer to enforce this rule.

The second approach is used for implicit inverse relations: a lemma that has a derivation marker is used to generate the source lemma, origin of the derivation, which is then sought in dictionaries. The generation is sometimes quite simple, as is the case with verbal nouns (gerunds) that are derived from most of imperfective verbs (marker `+Imperf` in DELAS dictionaries), and marked with `+VN`. The simple rule here is, to remove verbal noun suffix *-nje* and add an infinitive suffix *-ti*. Also, adjectives (past participles) are derived from most transitive verbs (marker `+Tr` in DELAS dictionaries), and marked with `+PP`. This procedure would establish a *derivation relation* between two above-mentioned verb variants, as well as respective verbal nouns and adjectives, starting from the following four e-dictionary entries:

```
afirmiranje,N300+VN+VAR=RatiSati
afirmiran,A6+PP+VAR=SatiRati
afirmisanje,N300+VN+VAR=SatiRati
afirmisan,A6+VAR=SatiRati
```

However, these verbal nouns and adjectives also come in variation pairs, so a *variation relation* is established between them also by using a procedure similar to the one described above.

6.2. Procedures for establishment of relations

We have developed a set of Unitex graphs, SQL procedures and C# tools to automate the task of explicit linking of existing entries. Even though our main goal was to connect existing entries, these automation tools introduced new possibilities for further expansion and annotation of dictionaries, including detection of missing markers and production of new entries. Here we will present, in more detail, two approaches that have been applied to actually connect lexical entries; first, the approach applied to produce and connect derived entries, and then the approach to connect lexical variants.

Establishing derivation relation is, in general, far from simple. So far, we have dealt with possessive adjectives derived from surnames, leaving other cases – diminutives, augmentatives, relational adjectives, gender motion, and so on – for future work. E-dictionaries contain a large number of surnames, both typical Serbian surnames (close to 18,000) and surnames of foreign origin transcribed according to Serbian orthography (close to 7,500). Possessive adjectives are often derived from surnames, e.g. *Lazić* ← *Lazićev* and *Ešton* ← *Eštonov* (Serbian transcription for 'Ashton'), as well as, in some cases, feminine nouns, *Lazić* ← *Lazićka* and *Ešton* ← *Eštonka* (women with surnames *Lazić* and *Ešton*, respectively). However, only a small number of these related lemmas were actually recorded in e-dictionaries (850 possessive adjectives and 25 feminine surnames). To systematically produce these derived lemmas we developed finite-state transducers (16 different FSTs), similar to those used for inflection, to derive possessive adjectives and feminine counterparts from all surnames, if they exist. One such FST is presented in Figure 3 and it derives a possessive adjective *Černijev* and a feminine surname *Černijka* from a surname *Černi* 'Czerny' (and 332 more surnames, mostly those ending with *i*). Derivation markers `+Pos` and `+GM` are added to the produced lemmas together with codes of inflectional transducers that should be applied to them (`A1` and `N661` in this case).

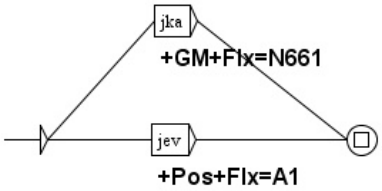

Figure 3: A FST for the derivation of a possessive adjective and a feminine counterpart from a surname belonging to a class of surnames.

As a result, the dictionary entry for the surname *Černi* produced two more derived and connected lemmas – the value

of the marker +BASE= explicitly establishes a *derivation relation* between the original and derived lemmas.

```
Černi,N1064+NProp+Hum+Last+SR
Černijev,A1+Pos+NProp+Hum+Last+SR
         +BASE=Černi_N1064
Černijka,N661+GM+NProp+Hum+Last+SR
         +BASE=Černi_N1064
```

The comparison of generated lemmas with those already in the dictionaries proved the correctness of this approach. In this way we generated more than 24,000 possessive adjectives and nearly 20,000 feminine counterparts of surnames that are all connected to basic lemmas.

For treatment of lexical variants, as well as simple derivation processes, as described in Subsection 6.1., we applied a different approach. In SMD, 5,592 lexical entries are annotated with one of 86 VAR markers.

This procedure, based on the set of rules, will be illustrated with two rule sets for variations: suffix_variations (124 rules) and affix_variations (44 rules) of a single lexical entry, but a similar approach is used for other types of variations, as well as for some simple derivation relations and pronunciation relations. Each rule is represented with the following set of attributes:

1. *RelationName* is a unique rule name and its identification, built upon a unique combination of other attribute values (e.g. *VAR=IratiOvati_V_V*);

2. *RelationType* is a type a variation: suffix_variations, affix_variations, etc.;

3. *SuffixFrom / SuffixTo* indicates suffixes (suffix_variations) or substrings (affix_variations) that a source/target lexical entry must contain;

4. *MarkerFrom / MarkerTo* is a required dictionary marker that a source/target lexical entry must have;

5. *Group* relates rules that are used in pairs.

The group attribute is used to relate a rule with its pair that generates a lexical entry in the opposite direction, e.g. from *oksidirati* → *oksidovati* can be generated, and conversely, from *oksidirati* → *oksidovati* (both meaning the same – *to oxidate*). In this way rule groups were introduced containing rules that come in pairs.

An example of a rule from this rule set is *VAR=ArisatiIrati_V_V*, which is applied to a verb that ends in *-arisati* and contains the marker *VAR=ArisatiIrati* (e.g. *komentarisati* 'to comment'). The rule can be used to generate its variation with suffix *-irati* and an inverse marker – *VAR=IratiArisati* (e.g. *komentirati*). This rule is in a group with five other rules: *VAR=ArisatiIrati_N_N* that generates a noun variation, *VAR=ArisatiIrati_A_A* that generates an adjective variation and three others with inverse markers *VAR=IratiArisati_V_V, VAR=IratiArisati_N_N,* and *VAR=IratiArisati_A_A*. A POS is an important part of these rules since it dictates the *SuffixTo* and *SuffixFrom* values which differ from rule to rule. For example, *-arisati* and *-irati* are related verb suffixes, *-arisanje* and *-iranje* are corresponding noun suffixes (*komentarisanje* vs. *komentiranje* 'commenting') and *-arisan* and *-iran* (*komentarisan* vs. *komentiran* 'commented') are adjective suffixes.

The second rule set (affix_variations) locates candidates that have a certain substring (one or more letters, but also an empty string indicating that a substring may be omitted) anywhere in the lexical entry, and an appropriate marker. For these rules a POS is irrelevant, but must be the same in both the origin and the target lexical entry. There are 22 two-rule inverse groups, which gives a total of 44 rules in this rule set. One example is the rule *VAR=0H* that describes lexical entries in which the letter *h* is missing and can be inserted to obtain a variant, for example *ladan* vs. *hladan* 'cold'. The corresponding inverse rule from the same group is *VAR=H0* indicating that a letter *h* may be omitted. The rule *VAR=CS* operates in a similar way, but in this case the operation is not omission/insertion but substitution – the letter *s* may be replaced with the letter *c*, thus generating, for example, *sufinanciranje* from *sufinansiranje* 'co-financing'.

These rules are not too successful for finding candidates for dictionary expansion because a large number of possible candidates may be generated due to unspecified position of the substring on which the rule operates. For example, *sufinansiranje* with *AffixFrom* being letter *s* and *AffixTo* being letter *c* can result in any of the following: *sufinanciranje, cufinansiranje* and *cufinanciranje*, with only the first one, in this case, being correct.

Developed rules were used to solve three subtasks:

1. Finding lexical entries that are missing in the dictionary (provided that their existence is indicated by markers of existing entries);

2. Finding lexical entries that exist in the dictionary, but lack the expected lexical marker (which is indicated by a marker assigned to a related existing entry);

3. Finding two lexical entries that exist in the dictionary and are expectedly marked (indicating a relation between them).

For the first option, a generated target entry becomes a candidate for a new lexical entry in the dictionary; for the second, a candidate for a marker annotation of an entry is generated; while for the third, a relation is established between two related lexical entries. This procedures also found a few errors in already assigned +VAR markers.

6.3. Statistics and Evaluation Results

The first subtask returned a total of 103 new candidates for dictionary entries through the suffix_variations rule set, of which 50 were accepted and 53 rejected. This may not seem as a very good result, but analysis revealed that the majority of the rejected candidates were actually marked with an incorrect +VAR marker, e.g. IratiOvati/OvatiIrati instead of a CiratiKovati/KovatiCirati. After these markers were corrected, 50 new candidate entries were accepted and only 3 were rejected. For the set of affix_variations rules, 119 candidates were returned for the first subtask, only 38 (32%) of them suitable. Most of affixes are very short (one

letter) and it is not easy to detect which letter should be affected by a rule if several of them occur in a single entry. Most of the rejected candidates were found due to unspecified number of replacements and their position (in cases when there is more than one replacement in the marked lexical entry).

The second subtask found only 35 lexical entries with missing markers. Since in each case both related entries existed in the dictionaries, and one is a possible variation of the other, there is just a small margin for errors. It was confirmed that all but one of the candidates were correct, and that this one occurred because one lexical entry variant was a homograph of another entry.

The third, and most important subtask, established relations between lexical entries using the produced rule sets to find properly marked pairs of entries (both having `+VAR` markers and a POS needed to activate a specific rule that generates their pair). A total of 5,129 symmetric relations was established, 4,411 through the suffix_variations rule set, and 718 through the affix_variations rule set. Frequencies of the most common variations used to connect entries are presented for suffix variations in Figure 4 and for affix variations in Figure 5.

Similar procedures are produced to connect some derivationally related entries (e.g. verbs and verbal nouns and adjectives) and to produce explicit inverse relations from originally implicit ones (in DELAS format).

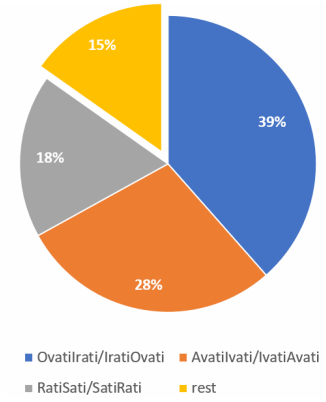

Figure 4: The frequency of established connections by relations from the suffix variations rule set.

7. Conclusion

In this paper we presented a new database model, developed upon the *lemon* model, as well as its application for migrating electronic morphological dictionaries from a single-user file system to a multi-user environment based on a relational database management system. The new lexical data model implemented as a lexical database has various advantages over the previously used file-based system. The introduced logical constraints will prevent omission of markers and enable their controlled use in the future. This will facilitate the enrichment of existing lexical entries with new markers and lexical relations, as we plan to establish as

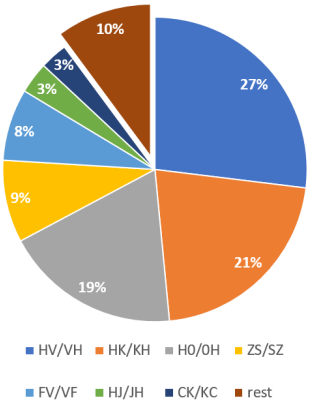

Figure 5: The frequency of established connections by relations from the affix variations rule set.

many explicit relations between lexical entries as possible, on the basis of information already given in SMD.

We adapted the *lemon* model in order to transfer all information stored in existing electronic dictionaries. Therefore, in our model the class `Form` is used for inflected forms instead of variant forms, which is important for Serbian as a highly inflective language. Also, we adapted the *lemon* model to store all existing markers as a thesaurus of data categories and their values, which enabled linking them to LexInfo and other ontologies, like SUMO. Mapping of grammatical categories as well as their values from existing dictionaries to LexInfo, using the catalog of grammatical categories that is complemented with the *lemon* model, is almost complete: for instance, grammaticalGender → lexinfo#gender, while m → lexinfo#masculine, f → lexinfo#feminine, n → lexinfo#neuter. However, for some categories the appropriate mapping still needs to be defined. The mapping of semantic markers to SUMO has also started, for instance `+DOM=Bot` → FloweringPlant and `+DOM=Culinary` → Cooking, but an exact match is not always possible. Future activities also include the use of linked data principles to enable open publishing and linking of language resources on the Web, integrating them with Linguistic Linked Open Data. After that novel application for dictionary management are planned, which will enable not only dictionary development and maintenance, but also their export to different dictionary schemata and formats, to support various NLP application needs.

The first part of the evaluation of the presented model was successfully completed, since all existing data were stored in the new database. Cross-linking was initiated, and some data-inconsistencies were detected and resolved. However, the final evaluation report will be given once the application is fully developed and database exploitation stars. Given that language resources for more than 22 languages, currently distributed with Unitex/GramLab, were developed in the same DELA format and that the presented migration approach is language independent, it is safe to say that it will prove useful for other languages as well.

8. Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178003.

9. Bibliographical References

Attia, M., Tounsi, L., and Van Genabith, J. (2010). Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic. Technical report, The NCLT Seminar Series, DCU, Dublin, Ireland.

Bański, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19—21 September 2017*, pages 485 – 494.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In Ilan Kernerman, et al., editors, *Proc. of GLOBALEX'16 workshop at LREC'16, Portoroz, Slovenia*, pages 65–72. European Language Resources Association, May.

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.

Courtois, B. and Silberztein, M. (1990). *Dictionnaires électroniques du français*, volume 87 of *Langue français*. Larousse, Paris.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international*, 7(3):97–100.

Francopoulo, G. (2013). *LMF Lexical Markup Framework*. John Wiley & Sons.

Ide, N., Kilgarriff, A., and Romary, L. (2000). ITRI-00-30 A Formal Model of Dictionary Structure and Content. In *Proceedings of EURALEX 2000*, pages 113–126. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2000.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Wright, S. E. (2008). ISOcat: Corralling Data Categories in the Wild. In *LREC*.

Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In Iztok Kosem, et al., editors, *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19 – 21 September 2017*, pages 598–613, Leiden, Netherlands, September.

Klajn, I. (2005). *Gramatika srpskog jezika*. Zavod za udžbenike.

Koeva, S., Krstev, C., and Vitas, D. (2008). Morpho-semantic relations in wordnet–a case study for two slavic languages. In *Proceedings of Global WordNet Conference 2008*, pages 239–253. University of Szeged, Department of Informatics.

Krstev, C. and Vitas, D. (2007). Extending the Serbian E-dictionary by using lexical transducers. In *Formaliser les langues avec l'ordinateur : De INTEX à Nooj*, pages 147–168.

Krstev, C., Vitas, D., and Erjavec, T. (2004). MULTEXT-East resources for Serbian. In *Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004*. Erjavec, Tomaž and Zganec Gros, Jerneja.

Krstev, C., Stanković, R., Vitas, D., and Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 1692–1697.

Krstev, C., Stanković, R., and Vitas, D. (2010). A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Maks, I., Tiberius, C., and van Veenendaal, R. (2008). Standardising Bilingual Lexical Resources According to the Lexicon Markup Framework. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, pages 1723–1727.

McCrae, J., Spohr, D., and Cimiano, P., (2011). *Linking Lexical Resources and Ontologies on the Semantic Web with Lemon*, pages 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg.

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.

Stanković, R., Obradović, I., and Utvić, M. (2013). Developing termbases for expert terminology under the TBX standard. In *35th Anniversary of Computational Linguistics in Serbia*, pages 12–26. University of Belgrade, Faculty of Mathematics.

Stanojčić, Ž. and Popović, L. (2008). *Gramatika srpskog jezika*. Zavod za udžbenike.

Tutin, A. and Véronis, J. (1998). Electronic dictionary encoding: Customizing the TEI guidelines. In *Proc. Euralex*.

Villegas, M. and Bel, N. (2015). PAROLE/SIMPLE 'lemon' ontology and lexicons. *Semantic Web*, 6:363–369.

Vitas, D., Pavlović-Lažetić, G., and Krstev, C. (1993). Electronic dictionary and text processing in Serbo-Croatian. *Sprache–Kommunikation–Informatik*, 1:225.

10. Language Resource References

Krstev, Cvetana and Vitas, Duško. (2015). *Serbian Morphological Dictionary - SMD*. University of Belgrade, HLT Group and Jerteh, Lexical resource, 2.0.

Stanković, Ranka and Krstev, Cvetana. (2016). *LeXimir*. University of Belgrade, HLT Group, Software Toolkit, 2.0.

Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web

Fahad Khan
CNR-Istituto di Linguistica Computazionale "A. Zampolli"
Pisa, Italy
fahad.khan@ilc.cnr.it

Abstract

In this article we take a detailed look at a number of issues relating to the publication of etymological data as linked data. We then put forward our proposal for an RDF-based model for representing etymologies that, as we will show, helps to answer at least some of the problems and requirements outlined in the initial part of the paper. We also take a more general look at the representation of diachronic lexical data as linked data.

Keywords: Etymology, Linked data, Ontolex-lemon

1. Introduction

Linked data with its core emphasis on linking together different and sometimes heterogeneous resources seems to be perfectly suited to the representation of etymological data since such data relies on the bringing together of evidence from diverse sources. In the case of etymology these can be primary sources that attest to the appearance, in a text, of a given word or phrase under a specific form or with a particular meaning, or they can be secondary sources that refer to salient hypotheses made by scholars in the past. In this article we take a detailed look at a number of issues relating to the publication of etymological data as linked data. We then put forward our proposal for an RDF-based model built on top of ontolex-lemon for representing etymologies that, as we will show, helps to answer at least some of the problems and requirements outlined in the first part of the paper. In addition we will take a more general look at the representation of diachronic lexical data as linked data.

In the next section, Section 2. we give an overview of some of the main challenges to modelling etymology in linked data. Then in Section 3. we make a first proposal of a model for a model for etymology. Next, in Section 4. we discuss the addition of temporal information to lexcal linked datasets.

2. The Challenges of Modelling Etymology in Linked Data

The word *etymology* has at least two different senses. In the first of these it is a sub-discipline of historical linguistics that concerns itself with the development of individual words (and other lexical entries) over time and attempts to trace their origins as far back as the evidential record will support – and sometimes even beyond. In addition *etymology* can also refer to a single such history of a word (or other lexical item). Etymologies in this latter sense can be found in many dictionaries and lexicons although typically in a condensed or abbreviated form. Note that we will be using both senses of the word in what is to follow, although we will focus predominantly on the latter.

Three important points which, as we argue below, have a significant impact on the modeling of etymologies in RDF, can be seen to immediately follow from the preceding definitions. The first point is that etymologies are essentially diachronic and call for the explicit representation of the unfolding of historical processes. In particular we often need to model the fact that a word w had the sense s during period the t, i.e., that a given property (having the sense s) holds for a certain period of time – something which is notoriously difficult to do, in a human-intuitive way, with a formalism like RDF that is limited to unary and binary predicates. There are several design patterns that can be used to overcome this expressive difficulty, none of which however turn out to be wholly satisfying on all or most accounts. Etymologies can potentially represent more than one kind of change as occurring at (around) the same time, so that as well as showing how a word's meaning alters over a given period, we might also want to depict the kinds of sound changes which it undergoes along with any shifts in written form and grammatical properties that might have occurred. Furthermore, the temporal information given in etymological sources is frequently underspecified (and of course it cannot be otherwise when it comes to reconstructed roots/words) and in many cases we lack a precise year or even century – or it is the case that whatever dates we do have are qualified with the modifier "circa". As these issues are very typical of etymological data, both in general purpose dictionaries and in specialist etymological works, we will need to take them into consideration when designing our model.

This leads us onto our next point, which is that etymologies have a marked tendency towards the speculative and in many cases there is no settled consensus as to a word's origins or the different twists and turns that it might have undergone during its historical development. In fact it's not unusual to find more than one etymology in a lexical entry and for etymologies to differ substantially for the same word across according to different sources. This

is due to the dearth of evidence relating to the earlier stages of modern day languages or to extinct languages and the frequent use of reconstructions in building up etymologies. It is therefore important to have a means of explicitly representing different hypotheses concerning a word's origin and development, as well as an accurate means of citing and, in general, describing the secondary literature. We will discuss this briefly in what follows. For reasons of space, the more general issue of how to represent attestations and citations in RDF versions of lexical/lexicographic resources, will not be covered here, although we do plan to discuss this in forthcoming work.

Another consideration to be borne in mind in the present regard is that, as was mentioned earlier, etymologies encompass different levels of linguistic description, typically the phonological or the semantic levels, and can concern more than one level at the same time. It is therefore an important precondition for an RDF based model for etymologies that there already exist a framework of different modules for representing these levels of linguistic description. In theory linked Data offers us much of the expressivity that we need to represent information at each descriptive level (at least in the case of a large number of etymological examples) but we currently lack specific, specialised, vocabularies; this is especially the case when it comes to representing different kinds of semantic shift.

We intend for our model to be used both in the creation of new lexical resources, or at least in cases where a significant amount of source material has yet to be integrated into a meaningful resource-wide organisational structure, as well as for retrodigitised lexicons and in consequence our model needs to be as fairly flexible. However as the conversion of retrodigitised print dictionaries into RDF is likely to be one of the most popular use cases for such a model[1] we have tried, as far as possible, to take the most common conventions of print etymological resources into consideration when designing our model.

2.1. Two Example Etymologies for the Word *girl*

Before we go on to describe our proposed model and in order to make our discussion a little more concrete than it has been up to this point we will take a look at the etymology of the word *girl* from two different sources. The first etymology is taken from Walter Skeat's influential etymological dictionary of English originally published in 1886,(Skeat, 1910):

> GIRL, a female child, young woman. (E.)
>
> ME. *gerle*, *girle*, *gyrle*, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) *girl* is a young woman; but in C.T. 666 (A 664), the pl. *girles* means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means 'young women;' in P. Plowman, B. i.33, it means 'boys;' cf. B. x. 175. Answering to an AS. form **gyr-el-*, Teut. **gur-wil-*, a dimin. form from Teut. base **gur-*. Cf. NFries. *gör*, a girl; Pomeran. *goer*, a child; O. Low G. *gör*, a child; see Bremen Wörtebuch, ii. 528. Cf. Swiss *gurre*, *gurrli*, a depreciatory term for a girl; Sanders, G. Dict. i. 609, 641; also Norw. *gorre*, a small child (Aasen); Swed. dial. *gårrä*, *guerre* (the same). Root uncertain. Der. *girl-ish*, *girl-ish-ly*, *girl-ish-ness*, *girl-hood*.

The second etymology is taken from Eric Partridge's single volume 'Origins: A Short Etymological Dictionary of Modern English' (Partridge, 1966)

> girl
> , whence **girlish**, derives from ME *girle*, varr *gerle*, *gurle*: o.o.o.: perh of C origin: cf Ga and Ir *caile*, EIr *cale*, a girl; with Anglo-Ir *girleen* (dim *-een*), a (young) girl, cf Ga-Ir *cailin* (dim -in), a girl. But far more prob, *girl* is of Gmc origin: Whitehall postulates the OE etymon **gyrela* or **gyrele* and adduces Southern E dial *girls*, primrose blossoms, and *grlopp*, a lout, and tentatively LG *goere*, a young p/erson (either sex). Ult, perh, related to L *puer*, *puella*, with basic idea '(young) growing thing'.

The first entry presents the word *girl* as having undergone a semantic change of narrowing from its original meaning of 'young man or woman' (as attested by a passage in the Canterbury Tales) to its modern meaning of 'young female person'. Skeat offers a number of possible cognates to *girl*, that is words that are probably derived from the same root as *girl*, in other Germanic languages, adding citations to the literature in support. He considers the origin of the word *girl* to be uncertain however, too uncertain, at least, to suggest any plausible hypotheses. Partridge on the other hand – and in spite of the fact that he labels the word as 'o.o.o.' (of obscure origin) – gives three different hypotheses as to the word's origin, citing the literature in support of a postulation from a reconstructed old English etymon.

3. A First Proposal for a Linked Data-Based Model for Etymology

Having prepared the ground in the preceding sections with a discussion of relevant topics, it is finally time to present our proposal for a model, an extension of ontolex-lemon, to represent etymological data in RDF. We do this in Section 3.2.; in the next subsection, Section 3.1., however, we will describe other relevant and/or related work in the area of language resources and technologies.

[1] Indeed this seems to be a very timely moment for the definition of such a model given the growing interest in converting lexicographic resources into formats such as TEI and RDF. C.f. the current European project ELEXIS. The fact that lexicography stands at the crossroads of several different humanistic disciplines – in particular historical linguistics, lexicography and philology – makes it an interesting and salient case study from the point of view of the ongoing development of the digital humanities (as well of course as raising a variety of non-trivial challenges from a computational point of view).

3.1. Related Work

Previous work on defining a framework for representing etymological data in digital lexical resources includes Salmon-Alt's proposal for an LMF based etymology model, (Salmon-Alt, 2006), as well as Bowers and Romary's work on the deep encoding of etymological information in TEI (Bowers and Romary, 2016). We have been influenced by both of these works in the development of our own model, though we will not detail the differences and similarities between their models and ours here.

With respect to modeling etymologies in RDF, previous work includes (De Melo, 2014) and (Moran and Bruemmer, 2013). In (Chiarcos et al., 2016) Chiarcos et al. defined a minimal extension of the lemon model with two properties for encoding and navigating etymological data: these were the symmetric and transitive `cognate` and the transitive `derivedFrom`. The adoption of such a minimal vocabulary for etymological data is likely to be sufficient for a good number of use-cases. Other cases, such as e.g., in the modeling of entries from more scholarly dictionaries, will necessitate a fuller representation of the evolution of a word, taking into consider its various linguistic properties at different points in time as well as the different hypotheses relating to each of them. Our intention in this article is to propose such a model, one that allows for the kind of so called 'deep' etymological modeling as described in bowers2016deep.

3.2. The Core Entities of our Model

Note that as we alluded to above, our proposed model is an extension or module of ontolex-lemon[2], the latest version of the popular lemon model(McCrae et al., 2017).

To begin with we will fix on the most important kinds of entity that we should, ideally, be able to refer to and to describe, i.e., to define predicates over, when modeling etymologies and that we will therefore want to make into classes[3].

The utility of being able to refer to etymologies themselves – their component parts, their provenances, and perhaps even their likelihoods as possible hypotheses – should be clear from the preceding discussion. It will therefore come as no surprise that we have made `Etymology` a class in and of itself[4]. Indeed this seems like an even more obvious move when you consider the frequency with which it is possible to find two or more different etymologies for the same entry in the same dictionary (c.f. the first etymology of *girl* presented above) or to have provenance information associated with individual etymologies (c.f. the citation of secondary literature in the second etymology of *girl*). We have decided not to limit members of the class `Etymology` to being associated with lexical entries only (for instance a sense or a morphological variant can each have their own separate etymologies).

The second main class which we propose is `Etymon`; the name is taken from the term in linguistics referring to words or morphemes from which other words or morphemes derive. The existence of the class `Etymon` enables us to make a distinction between the 'official' lexical entries in a lexicon and other lexemes whose main or only role is to describe etymological information relating to an entry (of course there are also 'official' lexical entries which also play the roles of etymons to other entries but these are still regarded as first-class entries). Why is this a useful distinction to make? Well, most comprehensive monolingual general purpose dictionaries for a language like English will contain etymological information – but we don't necessarily want, in the case of English, thousands of French and Latin words to appear in a list of all the instances of `LexicalEntry` in the resource – or at least not without being able to filter them out and distinguish them in some way. On the other hand it isn't enough to distinguish members of the class `Etymon` from lexical entries by the bare fact of their having been assigned a different language from the language(s) of the lexicon, since this wouldn't allow us to differentiate between cognates and etymons. In fact we also define the class `Cognate` in order to distinguish lexemes that play the role of cognates in an entry[5]. We have chosen to make both `Etymon` and `Cognate` subclasses of `LexicalEntry`.

Returning to the etymologies themselves: how do we relate together an instance of `Etymology` with the `LexicalEntry` whose history it describes and the instances of `Etymon` (or other elements) which it relates together? One option is to represent an `Etymology` as an ordered sequence of elements using one of the data structures provided by RDF, containers or collections or lists. But this might be too restrictive for our purposes since we may want to elaborate on the relationships between the different elements which have been ordered together in the etymology. In order to illustrate this point further we shall take as an example the English word *friar*, with an etymology adapted from Philip Durkin's Oxford Guide to Etymology (Durkin, 2009).

Although the word *friar* ultimately derives from *frāter*, the Latin word for 'brother', it first entered into English from Old French, from the polysemic word *frere* which means both 'brother' (as in the Latin) as well as 'member of a religious fraternity'. This latter sense was borrowed into Middle English as *frere* (with the same pronunciation as in the French) where it meant both 'member of a religious fraternity' as well as the more specialized meaning of 'member

[2] `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[3] With the obvious proviso that this can be done in different ways, and that the proposal we make is only one of several options in accord with the core necessities of describing etymologies.

[4] We are considering making `Etymology` a subclass of the class `Hypothesis` from the Linked Science Vocabulary (`http://linkedscience.org/lsc/ns/`) – but are still undecided on this point.

[5] We will not discuss the class `Cognate` further here, but will develop it in forthcoming work.

of a mendicant order' (but not 'brother' as in sibling), before finally coming, in modern English, to take on the latter sense. We can identify a number of different relationships between the various etymons identified above: Old French ***inherits*** the word *frāter* which, after having undergone a sequence of sound changes in the meantime, becomes *frere*, then Middle English ***borrows*** the word *frere* into its vocabulary, indeed borrows only a single sense of the word, eventually this changes its meaning through a process of ***specialisation***. The following shorthand description for the whole process which uses the '<' symbol [6] is again taken from (Durkin, 2009):

> Latin *frāter* brother<Old French *frere* brother, also member of a religious order of 'brothers'<Middle English *frere*, *friar*<modern English *friar*

By explicitly representing, indeed, reifying the shifts between instances of `Etymon` (and between an `Etymon` and a `LexicalEntry`) we can include important information on the type of etymological process that leads from one element to the other. For this purpose we have defined the `EtyLink` class that represents an etymological relationship between two elements. The `EtyLink` class can be seen as equivalent to the etymological symbol <. An instance of `Etymology`, then, consists of a series of such instances of `EtyLink`. This leads to a model (part of) which is represented in Figure 1.

Note the presence of the object property

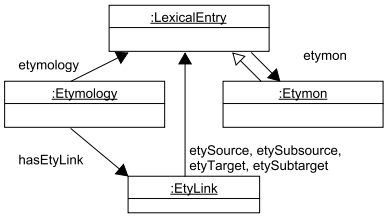

Figure 1: The relationship between some of the classes in our proposed model.

`etySource` which relates an `EtyLink` to a `LexicalEntry`/`Etymon` as its source, similarly with `etyTarget`. The two properties `etySubsource` and `etySubtarget` are designed to further specify the source and targets of an etymological relation between two entities. This is useful in case we want to elaborate on the sense or form which a lexical entry derives from. Using this model we can represent the Durkin *friar* example as in Figure 2.

Here we've given the first `Etymon` in the series the special status of root as the earliest `Etymon` to which we can trace

[6] Note that '<' is overloaded because it stands both for the development of one word from another or of a word borrowing from another language

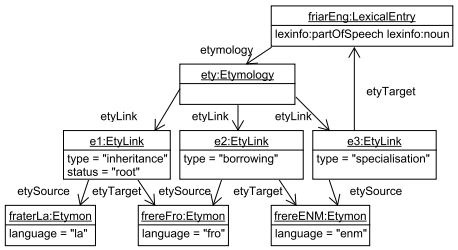

Figure 2: Modelling the *friar* example.

the word back to. If we wish to further specify the fact that the word *frere* in Early Modern English derives from the sense of the word in Old French in which it meant 'member of a religious order of brothers' then we can proceed as in Figure 3.

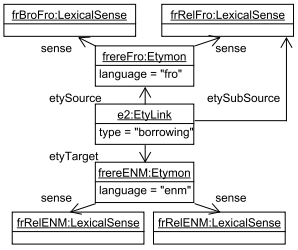

Figure 3: Modelling the *friar* example.

We can also specify, in a similar way, that the word *frere* in Old French derives from the accusative singular form of the Latin *frāter*, namely, *fratrem*.

As we stated above etymologies often include reconstructed words/root forms of words in reconstructed languages (as well as in historical languages for which there is a lack of relevant attestations) such as proto-Indo-European and proto-Germanic for which we have no surviving written attestations. Often etymologists will assign a meaning to these reconstructions on the basis of the evidence of words in other, attested languages; these reconstructed meanings however should be distinguished from other lexical entries for which there actually exists direct evidence. As Watkins (quoted by Durkin(Durkin, 2009)) points out 'reconstructed words are often assigned hazy, vague or unspecific meanings...The apparent haziness in meaning of a given Indo-European root often simply reflects the fact that with the passage of several thousand years the different words derived from this root in divergent languages have undergone semantic changes that are no longer recoverable in detail.'(Watkins, 2000).

In such cases the use of the ontolex-lemon `LexicalSense` class (and the `sense` relationship) would usually be inappropriate – on the other hand though we **would** like to be able to include semantic information

associated with the root or reconstructed word in question. Therefore, and given that this issue is an especially pertinent one in the encoding of etymologies, we have defined a new class in our model, `LexicalDomain`, in order to provide a weaker notion of meaning than that of `LexicalSense`, although as with the latter class `LexicalDomain` is intended to link a `LexicalEntry` with an ontology concept. So for instance the reconstructed root **ker-tā-* which has been assigned the meaning 'fire' and which is hypothesised to be the root of the English word *hearth* can be modelled as in Figure 3.2.. Note that the object relations `lexicalDomain` and `domainField` play a role that corresponds to `sense` and `reference`, respectively, in ontolex-lemon.

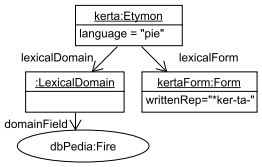

Figure 4: An example using the `LexicalDomain` class.

4. Adding Temporal Information to lexical data in RDF

Up until now we have avoided the issue of how to include temporal information in RDF etymologies and for good reason too. That is, as we mentioned above, it is not immediately obvious what the best way of doing this in RDF actually is. However in this section we will discuss one particular strategy for doing this.

In previous work(Khan et al., 2014), (E Díaz-Vera, 2014) we have opted for a perdurantist/four-dimensionalist(4D) approach when modeling sense shift[7], along with other diachronic lexical information, in RDF, and this is also what we propose in the present work. What, then, does this approach entail in the current case? Simply put, the idea is to treat elements such as senses, forms, and even whole lexical entries as having an inherent temporal extension. And so by making temporal extent a property of these elements we do not need to reify the original relation in order to introduce a temporal parameter[8].

We can think of it as follows. In ontolex-lemon the relation `sense` holds between a lexical entry l and each one of its lexical senses s. Now if it were an ideal world we could simply add a temporal parameter, specifying the interval, t, in which the `sense` relation holds, i.e., $\texttt{sense}(l, s, t)$. Obviously we can't do this in RDF. On the other hand, however, since a sense is already a reification of the meaning relation between a lexical entry and a reference (representing the extension of the entry) we *can* 'attach' this temporal information to s itself, that is rather than adding an extra parameter to the `sense` relation, without wreaking too much conceptual havoc as a result[9]. To reiterate then, we can represent a lexical sense as a entity with an extension in time that can be associated with a lexical entry and that describes one of its meanings as if it were a process in time, i.e., $\texttt{sense}(l, s)$ and $\texttt{hasTime}(s, t)$. It may be useful to distinguish senses that have temporal extent from 'normal' senses by referring to them as *p-Senses* (the 'p' here stands for *perdurant*) and creating a new subclass of `LexicalSense` called `LexicalpSense`. We can do something similar with the class `Form` and the object property `lexicalForm` in ontolex-lemon. Indeed one can go further and define an `Etymon` as a perdurant. This would give us much more expressivity in representing etymologies.

One of the advantages of explicitly representing temporal information in RDF is that it becomes much easier to query for such data. By making use of OWL axioms we can also reason over such data. The fact that the temporal information in etymological datasets is often vague and underspecified need not necessarily prove to be an insurmountable barrier to the use of OWL-based reasoning over such data. As we demonstrated in (Khan et al., 2016) it is fairly straightforward to reason with and query over such data by using Allen relations to describe the relationships between temporal intervals and by using other Semantic Web standards such as e.g., the Semantic Web Rule Language and the Semantic Query-Enhanced Web Rule Language.

5. Conclusion

Our intention in this article has been to present a first proposal on how to model etymologies as well as diachronic lexical data more generally in RDF through an extension of the RDF-native lexical model ontolex-lemon. Some of our proposals will, no doubt, be controversial but we hope the present work will serve to stimulate discussion on this issue and thereby help to contribute towards the definition of a standard (or recommendation) for modeling such data, one that will gain some measure of acceptance within the various communities that find themselves working with etymological data as part of their research.

6. Bibliographical References

Bowers, J. and Romary, L. (2016). Deep encoding of etymological information in tei. *arXiv preprint arXiv:1611.10122*.

Chiarcos, C., Abromeit, F., Fäth, C., and Ionov, M. (2016). Etymology meets linked data. a case study in turkic. In *Digital Humanities 2016. Krakow*.

[7] An good introduction to the 4D perspective can be found in (Welty and Fikes, 2006). We favour the slightly altered formulation given in (Krieger, 2014).

[8] C.f. `https://www.w3.org/TR/swbp-n-aryRelations/`

[9] In contrast a statement like 'Rome is the capital of Italy' is a little bit more difficult to model, and in the 4D view we would have to define a relation 'isCapitalOf' between a time slice of the referents of 'Rome' and 'Italy', i.e., we have to create two new entities that separately encode this temporal aspect. Other RDF temporal representation strategies have their own specific drawbacks.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. Citeseer.

Durkin, P. (2009). *The Oxford guide to etymology*. Oxford University Press.

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).

Khan, A. F., Bellandi, A., and Monachini, M. (2016). Tools and instruments for building and querying diachronic computational lexica. *LT4DH 2016*, page 164.

Krieger, H.-U. (2014). A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in rdf and owl. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 1.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. pages 587–597, September.

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.

Partridge, E. (1966). *Origins : a short etymological dictionary of modern English / by Eric Partridge*. Routledge and Kegan Paul London, 4th ed. (with numerous revisions and some substantial additions). edition.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *BULAG*, 31:1–12.

Skeat, W. W. (1910). *An etymological dictionary of the English language / Rev. Walter W. Skeat*. Oxford University Press London, 4th ed. revised, enlarged and reset. edition.

Watkins, C. (2000). *The American Heritage Dictionary of Indo-European Roots*. Houghton Mifflin Harcourt, second edition edition, September.

Welty, C. and Fikes, R. (2006). A reusable ontology for fluents in owl. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, pages 226–236, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Towards Temporal Reasoning in Portuguese

Livy Real, Alexandre Rademaker, Fabricio Chalub, Valeria de Paiva
USP, IBM Research, Nuance Communications
livyreal@gmail.com, alexrad@br.ibm.com, fchalub@br.ibm.com, valeria.depaiva@nuance.com

Abstract

This paper describes our ongoing work to create an open temporally annotated corpus in Portuguese and how this task helped to improve and evaluate linked open lexical resources, namely OpenWordNet-PT and TempoWordNet. We use the Linguateca's Bosque corpus, one of the most used open Portuguese corpora, and the system HeidelTime, the open tool that represents the current state of the art for time tagging, to build Bosque-T0. We compare the output of this work to what is present in the linked resources cited and discuss strengths and weaknesses of combining these knowledge bases.

Keywords: Portuguese, temporal expressions, HeidelTime, WordNet, corpus

1. Introduction

Although time and temporal reasoning pose many problems in language and logic (Steedman, 2005), much improvement has been achieved on temporal information retrieval (T-IR) in the last decade. At least since the first TempEval, 2007, there is an explicit effort and advance towards temporal tagging. Systems are performing close to the inter-annotator reliability for entity recognition (UzZaman et al., 2014), different domains are being explored (Bethard et al., 2015) and more complex tasks are addressed, such as temporal relation typing (Derczynski, 2016). While much progress can be found for English processing, the situation for languages other than English is not so optimistic. Recently, HeidelTime (Strötgen and Gertz, 2015) was made available for 13 languages, including Portuguese, with an automatically built expansion that promises to deal with more than 200 languages.

Here we will concentrate on verifying how much of the traditional wisdom in dealing with time in multilingual projects can be re-purposed, wholesale, for dealing with time in Portuguese. We focus on the HeidelTime system and linguistic linked open resources, namely, OpenWordNet-PT (de Paiva et al., 2012) and TempoWordNet (Dias et al., 2014), linked through the OpenMultilinguaWordNet project (Bond and Foster, 2013). Using non-language-specific tools for bootstrapping the creation of preliminary systems and linguistic resources to less resourced languages is useful in many ways. It creates baselines to compare further work to and it serves to start investigating applications that depend on the kind of data desired. Our applications depend on temporal data, so a preliminary investigation of tools and data for dealing with it is a requirement for our project.

We start by investigating what is the state-of-the-art for recognizing time expressions in Portuguese and progress to verify how good our lexical resources are for this first level of investigation. We aim at a fully fledged description of a temporal logic system, similar to the one in (Crouch and de Paiva, 2014), but we need to make sure that the basics are in place for Portuguese.

Steedman (Steedman, 2005) and Crouch (Crouch, 1998) start by discussing what a very naive approach to modelling temporal effects in natural language could be, simply using logical operators for the past and future. In the simplest possible case this would give us a modal logic with two tense operators, P (for past) and F (for future), applying to propositions ϕ that are evaluated in a model M.

When using logic to represent the meanings of natural language sentences, it is assumed that the temporal index of evaluation for the whole proposition is set to be the time s at which the utterance is made — the speech time. Thus, for example:

1. "John was in London" is true uttered at s iff $[P(in(john, london))]$ holds for $\{M, s\}$,
2. "John is in London" is true uttered at s iff $[in(john, london)]$ holds for $\{M, s\}$,
3. "John will be in London" is true uttered at s iff $[F(in(john, london))]$ holds for $\{M, s\}$.

where M is a model. The past tense formula evaluated at the speech time s shifts the temporal index to an earlier time — call this the event time — and evaluates the embedded (present tense) proposition relative to the event time. The absence of any operator, as in the present tense formula, means that the speech and event times are identical.

Although there are a number of shortcomings to this particular approach as a linguistic representation, we still want to have for Portuguese the ability to discuss these paradigmatic simple examples of sentences, in the most direct form possible. Thus the direct translations of the sentences above

1. "João esteve em Londres" is true uttered at s iff $[P(in(joao, londres))]$ holds,
2. "João está em Londres" is true uttered at s iff $[in(joao, londres)]$ holds,
3. "João vai estar em Londres" is true uttered at s iff $[F(in(joao, londres))]$ holds.

need to define a completely trivial temporal system in Portuguese, the same way that they do in English. While it seems clear that the *tense systems* are very different in English and Portuguese and that hence temporal markings might need to be modified and adapted, we are surveying the commonalities between the problems and solutions. We aim, just like (Costa and Branco, 2012a), to import open good tools we may find to help with the task at hand.

Here we describe our first steps towards a temporal reasoning in Portuguese, that surely are needed for temporal IR in Portuguese. We verify how well HeidelTime works for Portuguese and how much of temporal information is present in OpenWordNet-PT (OWN-PT), an open wordnet that we are working on since 2012. For this task, we pay special attention to open linked resources (LLOD) (Chiarcos et al., 2012). OWN-PT is linked to OpenMultilinguaWordNet (OMW), which links several WordNet projects, including TempoWordNet (TempoWN). We expected that the temporal information present in TempoWN would be valuable to us to improve OWN-PT and to help make sure that the basics are in place to allow temporal extraction and, thereafter, reasoning in Portuguese. The contributions of this preliminary investigation are: 1) Bosque-T0, a Portuguese corpus tagged by HeidelTime and a manual assessment of the data produced; 2) the improvement of OpenWordNet-PT's synsets related to temporal information; 3) an assessment of the quality found in TempoWordNet and of the usefulness of using this linked knowledge for Portuguese processing.

1.1. Related Work

Different approaches to T-IR arose in the last years. Many of them are libraries or specific modules of Natural Language Processing pipelines that normalize those expressions. Not so many, but still a notable number of lexical resources have also been paying attention to this issue. Here we briefly outline some libraries and resources available for Portuguese processing. As usual, much work has been done for English, and we can also find several recent works using a multilingual strategy. However, very few works are specifically concerned with Portuguese processing and most of those are not open source.

While TempEval shared tasks have produced much good work for English temporal evaluation, the only work we know that discusses time recognition for Portuguese is the HAREM evaluation. HAREM (Mota and Santos, 2008) is a series of shared tasks organized by Linguateca[1] for Named Entity Recognition, its last edition was held in 2008. It pays special attention to time expressions and uses a specific tagset that was built considering the state of the art of Portuguese processing at that time aiming to be useful to the Lusophone NLP community. Thus, the exactly tagset used in HAREM is not shared with any community, which makes difficult the task of comparing HAREM results with any other tools or data, as discussed in (Real and Rademaker, 2015).

Other work on Portuguese time expressions includes the LX-TimeAnalyzer (Costa and Branco, 2012b), the STRING system (Mamede et al., 2012) and specifically their temporal analyzer (Hagège et al., 2010). Mostly this work is based on proprietary systems and hence re-using it is difficult. The LX-TimeAnalyzer, for example, is made available for the community in a browsable version,[2] but its code is not open.

Turning to open tools, there is the work on Freeling (Padró and Stanilovsky, 2012) and on the HeidelTime (Strötgen and Gertz, 2015) frameworks. Freeling offers a date recognition module and two modules for Named Entities recognition, but we have not seen data about their accuracy. Since HeidelTime offers dates normalization, but also offers other kinds of temporal expressions recognition and uses the same annotation as the TempEval evaluations, we opt to start our investigation with HeidelTime.

2. Resources

Although many systems for T-IR do not rely on re-using information present in lexical resources, we believe, as do (Costa and Branco, 2012b), that combining the knowledge of wordnets with the knowledge of temporal oriented systems can improve the quality and coverage of both kinds of systems. This needs to be a two-way road: one can improve the coverage of the lexical resource considering the output of the temporal system and conversely one can improve the temporal tags, if we have more lexical knowledge. For instance, one needs to recognize adverbial expressions — such as *yesterday, today, tomorrow*, respectively *ontem, hoje, amanhã* — and these temporal expressions are not always recognized as such. More difficult is to correctly detect highly ambiguous words, as and , similarly ambiguous in Portuguese, whether they are used in temporal contexts or not. For this kind of sub-problem, lexical resources can be very helpful for T-IR. We discuss below the two resources we use in this work, as well the Bosque corpus and the HeidelTime system.

2.1. OpenWordNet-PT

OWN-PT[3] is an open access WN for Portuguese, originally developed as a syntactic projection of Universal WordNet of (De Melo, 2009). OWN-PT is linked to OpenMultilingual Wordnet(OMW)[4](Bond and Foster, 2013). Due to the construction of the Portuguese wordnet, all the original English synsets are already present in OWN-PT, but not all of them have Portuguese words. Many have not a single word form in Portuguese, or they miss translated glosses and examples. We are engaged in completing the translation of the empty OWN-PT synsets, but since this consists of a long term work, we focus on subsets of synsets related to specific tasks. Considering the synsets related to time expressions seems an interesting and productive idea, which is also related to our work on Portuguese processing of historical data (Paiva et al., 2014).

Princeton WordNet (PWN) classifies as temporal nouns 1028 synsets. Of these, more than 200 synsets have no Portuguese translations at the moment[5].

2.2. TempoWordNet

TempoWN[6] (Dias et al., 2014) is a free lexical knowledge base for temporal analysis where each synset of PWN is assigned to an intrinsic temporal value. TempoWN is

[1] https://www.linguateca.pt/HAREM/

[2] http://nlxserv.di.fc.ul.pt/lxtimeanalyzer.

[3] http://wnpt.brlcloud.com/wn/

[4] http://compling.hss.ntu.edu.sg/omw/

[5] March, 2018.

[6] https://tempowordnet.greyc.fr/

also linked to OMW, so the use of its base for improving OWN-PT is easily achieved. Each synset of TempoWN is semi-automatically time-tagged with four labels: atemporal, past, present and future. Temporal classifiers were learned from a set of time-sensitive synsets (manually curated) and then applied to the whole resource to give rise to TempoWN. So, each synset is augmented with its calculated temporal value. Perhaps the main difference between TempoWN and other resources and tools for temporal expressions recognition is the fact that TempoWN always tags a synset with a temporal value, even if most of the synsets have the 'atemporal' time value assigned.
Using PWN domain classification for nouns, we know which of the 82,115 noun synsets are related to time, the `noun.time` ones. Adjectives, verbs and adverbs can be related to temporal features, but this classification does not exist in PWN itself. Thus the use for us of TempoWN and its link to OMW would be to check how many temporal adjectives, adverbs and verbs should be in OWN-PT. We aim to detect, amongst the many adjectives, verbs and adverbs that exist in English and that are empty in Portuguese the ones that are temporally cogent.

2.3. HeidelTime

HeidelTime[7](Strötgen et al., 2013) is a multilingual, cross-domain temporal tagger that extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard. This standard uses the markup language TimeML (Pustejovsky et al., 2003). HeidelTime uses different normalization strategies depending on the domain of the documents that are to be processed, be them news, narratives (e.g., Wikipedia articles), colloquial (e.g., SMS, tweets), or scientific (e.g., biomedical studies). The tool is a rule-based system and its source code and the resources (patterns, normalization information, and rules) are strictly separated. Since 13 languages are supported with manually developed resources and Portuguese is one of these, we have decided to investigate it for our work.

2.4. The Bosque corpus

The Bosque corpus is a subset of 'Floresta Virgem', a collection of treebanks distributed by Linguateca[8]. According to the creators in their website, Bosque is "fully revised and corrected in the scope of the project, with a current size of 162,484 lexical units". The Bosque corpus has 9,368 sentences, corresponding to 1,962 different extracts from mostly newspaper text. But many of these 9,368 sentences are no grammatical sentences. Since the corpus was extracted from newswire, there are many headlines that are simply noun phrases like *PT no governo* (The Workers Party (PT) in Power). There are also dialogues, recognizable through the use of the names of the interlocutors, and answers to questions, which tend not to be full sentences. Still, Bosque is probably the most used corpus in the Lusophone community, it has both Brazilian and European Portuguese variants and has been annotated using several different linguistic theories. Most recently it has been converted to Universal Dependencies version 2.0 (Rademaker et al., 2017).

3. Bosque-T0

We call the temporally annotated version of Bosque of Bosque-T0[9]. The main purpose of Bosque-T0 is to be used as a baseline for future work. We ran the stand alone version of HeidelTime in our corpus, creating a temporally annotated corpus in Portuguese. This is similar to the work on TimeBank-PT (Costa and Branco, 2012c), but uses an open source temporal tagging system that is officially the state-of-the-art and that is available to all. TimeBank-PT is 'the result of translating the English corpus used in the first TempEval challenge to the Portuguese language'. While TimeBank-PT is TimeML annotated, it is a translation of an English corpus, not originally Portuguese texts. By contrast, the HAREM data collection is 'truly' Portuguese, but it does not use TimeML guidelines. Therefore, as far as we know, our work is the first open corpus that uses the TIMEX3 tagset, from the TimeML temporal markup language, in an original Portuguese corpus.
Out of the 1962 extracts, HeidelTime says 741 have no time annotations at all. Many of the sentences on these extracts do have temporal expressions, but these were not found by the tool. For instance, in the extract[10]

> Em relação ao mesmo mês do ano passado, quando os negócios atingiram 139,8 toneladas de ouro, a redução é de 61,37%. A média diária naquele mês foi de 6,6 toneladas, segundo dados da Bolsa de Mercadorias e Futuros.

no timex was found.
Considering that HeidelTime is rule-based, we expected that it would be able to detect all expressions composed by digits or expressions that tend to be always related to time, as the name of the months. But this does not always happen. For example, in the following examples, no timex was found either.

> A cotação para maio ficou em 20.000 pontos[11]

> *Empresa funciona das 9h às 19h, diariamente*.[12]

In total HeidelTime identified 2464 tags, 644 unique ones, of different types. Most of the ones identified were dates. Almost 300 timex occurrences were the word *ontem* (yesterday). Several temporal expressions were correctly marked, from full dates such as *dia 23 de maio de 1972* (day 23 of May of 1972) to some complex phrases such as *há cerca de 20 anos* (around 20 years ago).

[7] `https://github.com/HeidelTime/heideltime`.

[8] `http://www.linguateca.pt/floresta/info_floresta_English.html`

[9] Available at `https://github.com/own-pt/portuguese-time`.

[10] In comparison to the same month last year, when business achieved 139,8 tons of gold, the reduction was of 61,37%. The daily average in that month was 6,6 ton, according to data from the Bolsa de Mercadorias e Futuros.

[11] The price for May stood at 20,000 points.

[12] Company operates from 9am to 7pm, daily.

Nevertheless amonsgt the expressions found, we also find (interesting) mistakes. In[13]

> Manifestações espontâneas em protesto contra o facto de Daniel Cohn-Bendit, líder do Maio de 68, ter sido proibido de residir em França.

Maio de 68, a relevant French movement, which is also present in Wikipedia-PT[14], was tagged as DATE.
To see the kinds of issues that are problematic with the tagging, we choose some random 20 extracts from the Bosque-T0 to verify HeidelTime choices. Many temporal expressions are missed or half-marked. For example, in the sentence[15]

> A mudança do local de jogo que deve acontecer também na partida contra o Corinthians, no <TIMEX3>próximo</TIMEX3> dia 17 foi determinada pela CBF, que não viu garantias de segurança no estádio santista.

the term *próximo* (next) is correctly tagged, but the actual "day 17" *dia 17* is not.
Simply looking at the expressions produced by HeidelTime, we can see that a traditional way of referring to the past in Portuguese is missing altogether from the terms produced. For example the sentence[16]

> Monique, 37, disse que descobriu a marquinha, que não é pedra no rim quando se separou do marido, em junho passado.

should have "junho passado" (last June) marked. Not a single "passado" (last, just passed) appears in our HeidelTime terms.
It is also clear that more subtle ways of referring to time are much harder to tag. For example in the sentence [17]

> Eles se dizem oposição, mas ainda não informaram o que vão combater.

the word *ainda* (yet) can be a temporal marker, indicating that a event has not happened so far. While a full date, such as *dia 23 de maio de 1972* is easy to recognize and tag, a partial date, such as the year *1995* in the sentence [18]

> A seca que atingiu as áreas produtoras de grãos não deve causar grandes estragos na safra <TIMEX3>1994</TIMEX3>/95.

does not get recognized as a date.
Several of the holidays that we have been trying to complete in OWN-PT are not marked by HeidelTies as temporal events, yet. For example the sentence[19]

> Pizzaria oferece cardápio especial para Páscoa.

needed to mark 'Páscoa' (Easter) as a temporal noun, as it's marked in English. We recognize that what the HeidelTime developers call "temponyms" (Kuzey et al., 2016) are not fully developed, yet for other languages. They only exist for English, hence given the sentence[20]

> Muito mais do que nos tempos da ditadura, a solidez do PT está, agora, ameaçada.

we would not expect the expression *tempos da ditadura* (dictatorship times) to be marked. However at least the word *tempos* (times) we thought would be recognized as a temporal marker and tagged.
We are now in the process of checking the markings we have and verifying their accuracy. We plan to 'triangulate' information provided by OWN-PT for the sentences, with the HeidelTime tags in the near future.

4. Linked Open Data for Temporal IR

In this section we discuss how to improve the annotated corpus making use of the linked resources we have at hand, as well as, how OWN-PT can benefit from this work. Since TempoWN scores all PWN synsets with a temporal value, for this preliminary work, we considered only the synsets whose probability of being `PAST` or `FUTURE` according to TempoWordNet is above 90 percent. This represents already more than 3K synsets. Since TempoWN is not manually curated, as PWN and OWN-PT are, we started to manually check the quality of these probability assignments and unfortunately we found many labels that we do not agree with and that do not seem very useful for the present task. For example, the synset that has the higher probability, 0.998, of being `PAST` is `00012689-a: ideal | constituting or existing only in the form of an idea or mental image or conception`. While one can try to force the interpretation that this abstract image needs to be formed in the past to exist, there is nothing that really connects it to the usual notion of `PAST`. Another example of inconsistent score is the pair *00130151-a — retrograde — of amnesia; affecting time immediately preceding trauma* and *00130281-a — anterograde — of amnesia; affecting time immediately following trauma*, with probability of being `PAST` of 0.996 and 0.994 respectively. It does not matter if one fixes the `PRESENT` as the moment of speech or the moment the related trauma happened, one cannot have *retrograde* and *anterograde* tagged with the same label.
At first glance, TempoWN has a large coverage that seems to be useful for temporal tagging, but its information is too noisy to be useful. Checking simply the

[13] Spontaneous demonstrations protesting against the fact that Daniel Cohn-Bendit, leader of May 1968, was banned from residing in France.

[14] `https://pt.wikipedia.org/wiki/Maio_de_1968`

[15] The change of place for the match, which should happen also in the match against the Corinthians on the next 17th, was determined by the CBF, which did not see guarantees of security measures in the Santos stadium.

[16] Monique, 37, said that she discovered the little mark, not a kidney stone, when she got divorced from her husband last June.

[17] They say they're the opposition, but have not informed us, yet, what they will oppose.

[18] The drought that hit the grain growing areas should not cause a big disaster in the harvest year 1994/95.

[19] Pizzaria offers special menu for Easter

[20] More than in the times of the dictatorship, the existence of the PT is now threatened.

most frequent timex expressions in Bosque-T0 in TempoWN and OWN-PT, we could complete some missing synsets in Portuguese, but we should not use the extra time score offered by TempoWN. While the synset for *ontem*(yesterday) has more than 0.99 probability of being `PAST` and *agora*(now), is also scored as 0.99+ `PRESENT` some other probability assignments seem dubious; The synset for *hoje* `00207366-r | today | on this day as distinct from yesterday or tomorrow`, appears in TempoWN with 0.99+ probability of being `FUTURE` and *próximo* `00054212-r | next | at the time or occasion immediately following;`, has 0.99+ probability of being `PAST`.

We reap the benefits of linked linguistic open data through the connection established between TempoWN, OMW and OWN-PT. But it is harder to decide if the TempoWN information is useful for the task at hand or not. The markings of adjectives and adverbs should be useful for reasoning with texts in Portuguese, if the probability assignments are reasonable. Many of them seem good, but how to improve TempoWN scores is future work.

Many of the timex expressions found in Bosque-T0 were missing in OWN-PT at the beginning of this work, for instance the synset `00065748-r | last | most recently`. While in English, this is clearly an adverb, in Portuguese, we need an adverbial phrase to convey the same kind of meaning *por último* ("by last").

For this preliminary work more than 300 temporal synsets were completed in OWN-PT. Many language or culture specific ones are still missing. Some of these empty Portuguese synsets are typical holidays in the United States, such as the synset *15189982-n* for *Father's Day*. There is a holiday called Father's Day (*Dia dos Pais*) in Portuguese. But it happens at different times in Brazil (August) and Portugal (March), while it happens in June in the US and England. Thus in PWN this synset holds a relation with *June*, which only makes sense for the English wordnet. This hints at the issues of the intersection of multilingual and multicultural aspects of lexical and world knowledge Looking at these translations also helps to notice smaller differences between the languages. A typical and principled difference between the wordnets is that we do not use a prefix like "mid" in the synset *15211711-n* for *mid-May*; we say instead *meados de maio*, which although can be seen as a multi-word expression, is compositional in Portuguese and therefore it may not necessarily be included in a Portuguese lexical base if multilingual alignment was not a previous goal.

5. Conclusions

We presented our ongoing work towards temporal reasoning in Portuguese. Since not much is available for Portuguese natural language processing, we started providing an open corpus temporally tagged by the HeidelTime tool, which we call Bosque-T0. In the process of analyzing the annotations of Bosque-T0 we improved the coverage of OpenWord-Net temporal synsets and discussed how its link to a temporal wordnet, TempoWordNet, could be useful for this task.

Due to the different building processes of OpenWordNet-PT and TempoWordNet, the quality of those resources are radically different. While OpenWordNet-PT has less, but reliable information, TempoWordNet offers temporal scoring for every synset of Princeton WordNet, but most of the scores are controversial. We briefly discussed the issues found in Bosque-T0, which show that much work still needs to be done to address temporal IR in Portuguese — at least as far as using open-source tools and resources is considered. We aim to use Bosque-T0 as a baseline for this future work.

For future work we would like to improve the Portuguese HeidelTime system, using the insights gained from analyzing the issues found in Bosque-T0. We also want to manually annotate a small part of the Bosque corpus with TIME3 tagsets to make it available as a small golden corpus. Checking how well HeidelTime deals with TimeBank-PT and the HAREM corpora are also possible next steps. Finally maybe one should try a deep analysis of the proposed adaptation of the TimeML guidelines to Portuguese, as proposed by (Hagège et al., 2010).

As said before, we are interested about temporal reasoning, not only in Temporal Information Retrieval. As a long term goal, we aim to merge temporal information with other linguistic levels. We plan to do so using Bosque-UD, the human revised version of Bosque corpus annotated with Universal Dependencies.

Although this experiment has shown that using TempoWordNet does not improve our processing, we still believe in the benefit of Linguist Linked Open Data. For example, using the information of DBPedia Português [21], we could solve the discussed issue of extracting culturally specific holidays.

6. Bibliographical References

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., and Verhagen, M. (2015). Semeval-2015 task 6: Clinical tempeval. *SemEval 2015*.

Bond, F. and Foster, R. (2013). Linking and extending an Open Multilingual Wordnet. In *ACL, 2013*.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). Linking linguistic resources: Examples from the open linguistics working group. *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer:p. 201–216.

Costa, F. and Branco, A. (2012a). Extracting temporal information from Portuguese texts. In *PROPOR 2012*, volume 7243 of *Lecture Notes in Artificial Intelligence*, pages 99–105, Berlin, Germany. Springer.

Costa, F. and Branco, A. (2012b). LX-TimeAnalyzer: A temporal information processing system for Portuguese. Technical Report DI-FCUL-TR-2012-01.

Costa, F. and Branco, A. (2012c). TimeBankPT: A TimeML annotated corpus of Portuguese. In *LREC'12*, pages 3727–3734, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Crouch, D. and de Paiva, V. (2014). If, not when. *Elec-*

[21] `http://pt.dbpedia.org`

tronic Notes in Theoretical Computer Science, 300:3 – 20. IMLA 2013, {UNILOG} 2013.

Crouch, D. (1998). Temporality in natural language. In *ESSLLI class notes*, Saarbruecken, Germany.

De Melo, G.and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *18th ACM conference on Information and knowledge management*. ACM.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In *COLING 2012*.

Derczynski, L. (2016). *Automatically Ordering Events and Times in Text*. Studies in Computational Intelligence. Springer International Publishing.

Dias, G., Hasanuzzaman, M., Ferrari, S., and Mathet, Y. (2014). Tempowordnet for sentence time tagging. In *23rd International Conference on World Wide Web*, pages 833–838. ACM.

Hagège, C., Baptista, J., and Mamede, N. J. (2010). Caracterização e processamento de expressões temporais em português. *Linguamática*, 2:63–76.

Kuzey, E., Strötgen, J., Setty, V., and Weikum, G. (2016). Temponym Tagging: Temporal Scopes for Textual Phrases. In *TempWeb '16*, pages 841–842. ACM.

Mamede, N., Baptista, J., Diniz, C., and Cabarrão, V. (2012). String: An hybrid statistical and rule-based natural language processing chain for portuguese.

Cristina Mota et al., editors. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *LREC 2012*, Istanbul, Turkey, May. ELRA.

Paiva, V. D., Oliveira, D., Higuchi, S., Rademaker, A., and Melo, G. D. (2014). Exploratory information extraction from a historical dictionary. In *IEEE 10th e-Scienc)*, volume 2, pages 11–18. IEEE, oct.

Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003). TimeML: robust specification of event and temporal expressions in text. In *(IWCS-5*.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal dependencies for portuguese. In *Depling 2017*, pages 197–206.

Real, L. and Rademaker, A. (2015). Harem and klue: how to compare two tagsets for named entities annotation. In *NEWS 2015)*, Beijing, China, July.

Steedman, M. (2005). The productions of time: Temporality and causality in linguistic semantics, (available from his webpage).

Strötgen, J. and Gertz, M. (2015). A baseline temporal tagger for all languages. In *EMNLP 2015*, September.

Strötgen, J., Zell, J., and Gertz, M. (2013). Heideltime: Tuning english and developing spanish resources for tempeval-3. In *SemEval 2013*, pages 15–19, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., and Pustejovsky, J. (2014). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv:1206.5333v2*.

Towards LLOD-based Language Contact Studies: A Case Study in Interoperability

Christian Chiarcos, Kathrin Donandt, Hasmik Sargsian, Jesse Wichers Schreur, Maxim Ionov
Goethe University Frankfurt, Frankfurt am Main, Germany
{chiarcos|donandt|ionov}@cs.uni-frankfurt.de,
{sargsyan|wichersSchreur}@em.uni-frankfurt.de

Abstract

We describe a methodological and technical framework for conducting qualitative and quantitative studies of linguistic research questions over diverse and heterogeneous data sources such as corpora and elicitations.
We demonstrate how LLOD formalisms can be employed to develop extraction pipelines for features and linguistic examples from corpora and collections of interlinear glossed text, and furthermore, how SPARQL UPDATE can be employed
(1) to normalize diverse data against a reference data model (here, POWLA),
(2) to harmonize annotation vocabularies by reference to terminology repositories (here, OLiA),
(3) to extract examples from these normalized data structures regardless of their origin, and
(4) to implement this extraction routine in a tool-independent manner for different languages with different annotation schemes.
We demonstrate our approach for language contact studies for genetically unrelated, but neighboring languages from the Caucasus area, Eastern Armenian and Georgian.

Keywords: Linguistic Linked Open Data, language contact, Georgian, Armenian, syntax, corpus interoperability

1. Motivation

We describe a methodological and technical framework for qualitative and quantitative investigations of linguistic research questions which heavily depend on data such as corpora, elicitations, etc. It can be used for all research areas, but is primarily suitable for typological, historical and comparative studies. We demonstrate our approach using a specific research question in language contact studies as a case study.

For such research, there are usually several data sources, e.g. a dictionary, a number of elicitations, or even a corpus. All of these may be in different formats without an interface to query over them simultaneously. Furthermore, these linguistic resources may not even share a tagset, and may have different annotations for the same grammatical categories. We show that by applying (Linguistic) Linked Open Data (LLOD) principles, we are able to unify different types of resources, and query these heterogeneous sources as a single united resource.

Linguistic Linked Open Data (LLOD)[1] describes the application of Linked Open Data principles and methodologies for modeling, sharing and linking language resources in various text- and knowledge-processing disciplines. These disciplines range from artificial intelligence and computational linguistics via lexicography and the localization industry to linguistics and philology. For these areas, a number of benefits of LLOD and the underlying RDF technology over traditional representation formalisms have been identified (Chiarcos et al., 2013). Most notable for the work described here, this includes *representation* (linked graphs can represent any kind of linguistic annotation), *interoperability* (RDF graphs can easily be integrated), *ecosystem* (broad support by off-the-shelf database technology), and explicit *semantics* (links to commonly used vocabularies provide community-approved meanings for concepts and data structures).

LOD interoperability and the ability to use its shared vocabularies provides the possibility to integrate and enrich different and heterogeneous language resources. In our project, we focus on applying this methodology to studies in various areas of linguistics: Armenian and Kartvelian studies, language contact studies, syntax, and typology.

In this paper, we show the application of this approach on the study of similar syntactic constructions in Standard Eastern Armenian and Modern Georgian using heterogeneous resources. In order to use those resources we convert them to a unified representation. Using RDF conversion and further SPARQL UPDATE queries, we create a pipeline that dynamically annotates a data stream (with a help of `CoNLLStreamExtractor`, a part of the CoNLL-RDF library (Chiarcos and Fäth, 2017)[2]). The enriched annotation can then be used to conduct the research at hand.

The remainder of the paper is structured as follows: Section 2 introduces the linguistic problem under consideration, Section 3 presents the corpus data and explains its conversion to a unified format which is a necessary preparation step for the experiment described in Section 4. Section 5 reflects on the results of the experiment and the insights gained, and discusses its relevance for approaching the linguistic problem at hand.

2. Linguistic Background

2.1. Introduction

Georgian and Armenian are genealogically unrelated languages that have been spoken in neighboring areas for centuries. Hence, they are expected to share a number of features on different levels of linguistic analysis, among which

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

[2] https://github.com/acoli-repo/conll-rdf

syntax. One of the common syntactic-pragmatic features of Georgian and Armenian is pre-verbal focus (Comrie (1984, pp.1-2); Harris (1981, pp.14-18)). With pre-verbal, we mean the position directly before the finite verb, which can be either a main verb or an auxiliary. See Section 2.2 for a short discussion of focus.

As a case study, we look into common analytic predicative constructions in these languages, namely those that consist of an auxiliary verb and a main verb. More specifically, we consider the position of the auxiliary with respect to the main verb. This will serve as a basis for a further research on the effects of word order on the focus of the clause. If the results are similar in both languages, this would be a possible testament to syntactic convergence in the history of these neighboring languages.

We restrict our preliminary research on word order samples to the *to-be*-auxiliary and a modal auxiliary in Armenian, and three modal auxiliaries in Georgian.

2.2. Terminology

There is hardly a completely unambiguous and cross-linguistically valid definition for the term 'auxiliary (verb)' (Ramat, 1987, pp. 3-19). In the present paper, however, we use the term in its broader sense of a finite verb (with full or defective inflection), which is used in combination with the lexical verb and expresses features such as person, number, and TAM[3].

Focus is the grammatical category that determines which part of the sentence provides new or contrastive information (see further Zuo and Zuo (2001)). In many languages, e.g. in Armenian, instead of (or in addition to) stress, word order can be used to express focus[4], see the example below:

a. *Kat´oɫikos-ə* *ut-um* *ēr*
Catholicos-DEF eat-IPFV AUX.PST.3SG

'Catholicos was eating.' (And not doing something else)[5]

b. *Kat´oɫikos-n* *ēr* *ut-um*
Catholicos-DEF AUX.PST.3SG EAT-IPFV

'Catholicos was eating.' (It was Catholicos, who was eating.)

2.3. Georgian and Armenian

Eastern Armenian forms some of its tenses by combining certain non-finite forms of the verb with the unstressed *to-be*-auxiliary, which originates from the copula and is inflected for person/number and tense (present/past) (cf. e.g. Comrie (1984); Tamrazian (1991); Kahnemuyipour and Megerdoomian (2017)). While the context-independent citation form of this predicative construction is V AUX, the auxiliary can attach enclitically to any constituent before the main verb in a given context to mark the syntactic focus of the clause. However, it cannot attach to full words following the verb (this was verified by the results of the corpus search, see Section 7.1.).

In Modern Georgian, just as in English, the notions of possibility, necessity and desire are expressed by auxiliary verbs: *unda*[6] 'must', *minda* 'I want', *mč'irdeba* 'I need', *šemiʒlia* 'I can'. Georgian natural sentential word order fluctuates between SOV[7] and SVO (Vogt, 1974)[8] with a preference for OV in shorter sentences (Apronidze, 1986, p. 26). In languages with dominant SOV order, one would expect the auxiliary to follow the main verb (Greenberg, 1963, universal 16). However, a cursory corpus-based investigation (looking at the verbs 'must', 'to want' and 'to be able to' in the GNC[9]) shows that appr. 80% of clauses with an auxiliary show the order AUX V, which corresponds to the citation form of Armenian modal verbs (e.g. *piti gnam* must go.1SG 'I must go').

Thus, the prevalent order is V AUX (where AUX is a form of 'to be') in Armenian and AUX V (where AUX is a modal verb) in Georgian. A further investigation will consider conditions under which word order deviates from these prevalent patterns and the frequency of certain order types. One such condition could be focus, since the element directly before the AUX is expected to have syntactic focus. Furthermore, the influence of different types of focus (besides syntactic focus) could be examined. If both Armenian and Georgian show similar strategies regarding the expression of focus with use of the placement of the auxiliary, syntactic convergence due to language contact could be considered.

In the scope of the present paper, we only conduct a preliminary experiment in order to check the operability of the pipeline.

3. Language Resources

3.1. Eastern Armenian National Corpus

With its 110 million tokens, the Eastern Armenian National Corpus (EANC)[10], contains written texts in different genres (fiction, news, scientific texts, and other non-fiction), transcripts of oral communication, and logs of electronic communication. Nearly all genres are represented as fully as possible (except for the electronic communication and online news). All the texts are morphologically parsed without disambiguation. A tagset used for the corpus was developed specifically for the EANC project.

From a technical perspective, texts are represented in a CoNLL-like format (TSV[11]). The main difference from the traditional CoNLL is the presence of alternative parses: since there is no disambiguation in the EANC corpus, an

[3] Tense, aspect, mood.

[4] Here, we refer only to syntactic focus; Comrie (1984, pp.3-4) distinguishes this from pragmatic and intonational focus.

[5] Vrt'anes P'ap'azyan, Stories. EANC

[6] Although discussion may arise as to whether this word is truly verbal (since it is not inflected), it does fulfill the same function as the other modal verbs.

[7] S(ubject), O(bject), V(erb).

[8] The same uncertainty as to basic SOV-SVO order applies to Armenian, cf. Comrie (1984, p. 4).

[9] Georgian National Corpus, see Section 3.3.

[10] http://eanc.net/EANC/search

[11] Tab-separated values

annotation of each word is repeated for every possible morphological parse. To represent this in CoNLL, the authors output every possible parse as a separate word (on a new line) but with the same word ID. This non-standard format required updating the CoNLL-RDF conversion (see section 4.1.) to correctly handle this design decision.

3.2. Interlinear Glossed Georgian Text in FLEx

Fieldwork Language Explorer (FLEx)[12] is a tool designed for field linguists to create interlinear glossed text and lexicons, and also features some (limited) corpus query functionalities. The user can completely customize its part of speech tagsets, and the glosses of grammatical morphemes can be viewed as further annotation tags. The output is an XML file with the extension .flextext, which contains one annotated text. A collection of short stories by Erlom Akhvlediani (1986) called "Vano & Niko" have been glossed and exported accordingly. This sample consists of approximately 900 sentences and reflects the modern standard Georgian literary language.

3.3. Georgian National Corpus

The Georgian National Corpus (GNC)[13] is developed by researchers at the universities of Frankfurt, Bergen, and Tbilisi, and contains over 227 million tokens. The corpus, which is still under development, contains subcorpora of Old, Middle and Modern Georgian, plus two subcorpora of transcribed recordings of spoken language. Corpora of Megrelian and Svan texts are under construction as well. A large Georgian reference corpus (GRC) is included that contains less thoroughly processed texts from various fictional and non-fictional domains. The Georgian texts (within GNC and GRC) are fully morphologically annotated (lemma forms and morphosyntactic features), and all texts in the GNC subcorpora have comprehensive metadata.

4. Conversion to RDF

In a first step, we convert the source formats to an isomorphic rendering in RDF, which then represents the basis for further normalization.

4.1. CoNLL ⇒ RDF

To facilitate the processing of TSV formats such as the EANC format, the CoNLL format family, or popular infrastructures such as the corpus workbench, the **CoNLL-RDF** package (Chiarcos and Fäth, 2017)[14] uses RDF technology. In this way, it enables the advanced manipulation of annotated corpora (graph rewriting) with SPARQL UPDATE, their quantitative evaluation with SPARQL SELECT, off-the-shelf database support with RDF Triple/Quad Stores, sentence-level stream processing and access with a W3C standardized query language (SPARQL). CoNLL-RDF provides an isomorphic, but shallow reconstruction of CoNLL data structures in RDF:

- Every row — which in standard CoNLL corresponds to a word — is mapped to a *nif:Word* (using the NIF vocabulary, Hellmann et al. (2013)). As mentioned above, the EANC corpus is not disambiguated and therefore, there can appear several lines for one and the same word in the TSV files, each line containing the word with a different possible parse. This problem was solved by joining the different annotations into a triple group containing the same subject (the URI of the word) and predicate (the annotation type), while having several objects — one for each annotation possibility (e.g. `:s1_1 conll:GRAM "cvb conneg", "sbjv pres sg 3", "imp sg 2".`).
- Consecutive words are connected by *nif:nextWord*.
- Rows which are not separated by an empty line are represented as a *nif:Sentence*.
- Consecutive sentences are connected by *nif:nextSentence*.
- The actual annotations in the original CoNLL files are stored in columns. Every column with a user-provided label, say, *WORD, POS*, etc., is rendered as a property in the conll namespace (*conll:WORD, conll:POS*, etc.).

The EANC corpus files and the GNC data are converted to CoNLL-RDF, because the GNC — in addition to its native XML format — is also available in CoNLL-U. An example of the resulting RDF data displayed in the Turtle syntax is given in Fig. 1.

4.2. FLEx ⇒ RDF

For the RDF rendering of the FLEx data, we use the FLEx LLODifier tool,[15] which converts to the so-called FLEx-RDF format. The LLodifier is a collection of tools for converting language resources into an RDF representation (Chiarcos et al., 2017). In comparison to CoNLL, the FLEx data model is complex, as it allows annotations on three levels of granularity: *flex:phrase, flex:word*, and *flex:morph*. These are furthermore organized hierarchically (a *flex:phrase flex:has_word* some *flex:word*, a *flex:word flex:has_morph* some *flex:morph*) as well as sequentially (*flex:next_phrase*, *flex:next_word*, *flex:next_morph*).

5. Harmonization

These different, source-specific RDF renderings of our respective data are now transformed into uniform representations by anchoring them in more general LLOD vocabularies and terminology bases.

To represent linguistic data structures in general, we use POWLA (Chiarcos, 2012), an OWL2/DL reconstruction of the Linguistic Annotation Framework (LAF).[16] From the LAF, POWLA inherits the claim to represent *any* linguistic data structures applicable to textual data.

[12] https://software.sil.org/fieldworks/

[13] http://gnc.gov.ge/gnc/page?page-id=gnc-main-page

[14] implemented in Java and available under Apache 2.0 license, https://github.com/acoli-repo/conll-rdf

[15] https://github.com/acoli-repo/LLODifier/tree/master/flex

[16] https://www.iso.org/standard/37326.html

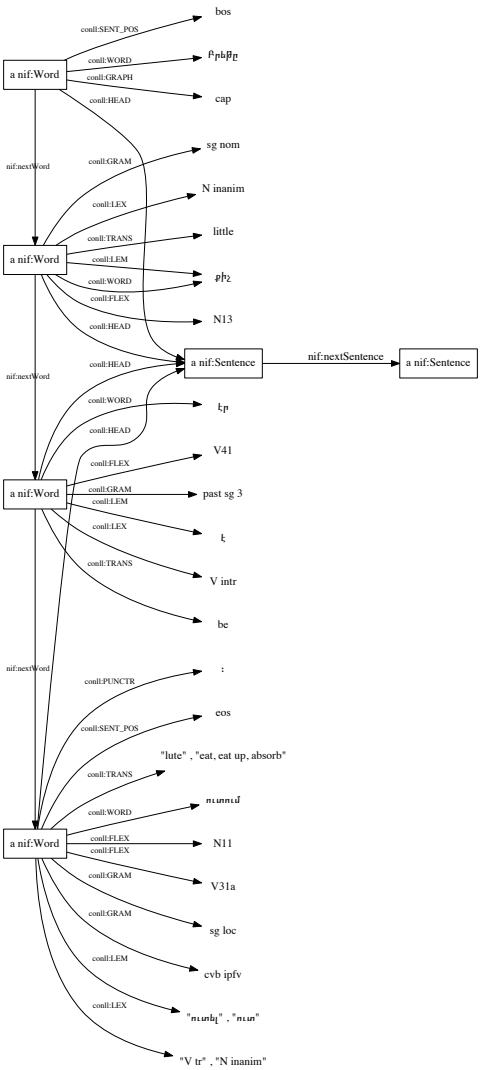

Figure 1: Example of the EANC data converted to CoNLL-RDF

To represent linguistic annotations while guaranteeing interoperability, we apply the Ontologies of Linguistic Annotation (OLiA)[17] which allow us to derive a structured, ontology-based representation from plain tags as used during the annotation.

5.1. POWLA and the LAF

It is generally accepted that any kind of linguistic annotation can be represented by means of directed (acyclic) graphs (Bird and Liberman, 2001; Ide and Suderman, 2007): Aside from the primary data (text), linguistic annotations consist of three principal components, i.e., segments (spans of text, e.g., a phrase), relations between segments

[17] http://purl.org/olia/

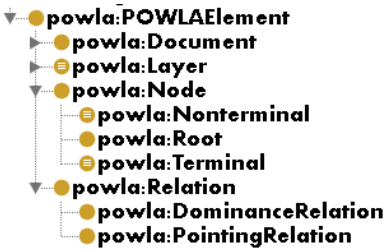

Figure 2: The POWLA data model

(e.g., dominance relation between two phrases) and annotations that describe different types of segments or relations. In graph-theoretical terms, segments can be formalized as nodes, relations as directed edges and annotations as labels attached to nodes and/or edges. These structures can then be connected to the primary data by means of pointers. A number of generic formats have been proposed on the basis of such a mapping from annotations to graphs, most importantly the Linguistic Annotation Framework (LAF) developed by ISO TC37/SC4. Such formats are traditionally serialized as standoff XML, e.g., in the GrAF format, but as these are poorly supported by off-the-shelf technology and highly domain-specific, serializations of this data model in RDF have been developed. Here, we focus on POWLA (Chiarcos, 2012), an OWL/DL serialization of the data model of the PAULA XML format (Dipper, 2005; Chiarcos et al., 2008; Chiarcos et al., 2011), a generic interchange format that originates from early drafts of the Linguistic Annotation Framework, and which is closely related to the later ISO TC37/SC4 format GrAF. PAULA was designed to support the lossless representation of arbitrary kinds of text-oriented linguistic annotation, and in particular the merging of annotations produced by different tools (e.g., multiple independent syntax annotations (Chiarcos, 2010) , or syntax, coreference and discourse structure annotation at the same time, (Chiarcos et al., 2011)). With POWLA, these annotations can also be represented by means of Semantic Web standards.

The POWLA data model, as illustrated here (Fig. 2), is relatively minimalistic. Aside from corpus structure (*powla:Document, powla:Layer*), annotations are grounded in *powla:Nodes* which can be linked by *powla:Relations* (hierarchical dominance relations, or non-hierarchical pointing relations with explicit *hasTarget/hasSource* properties). Hierarchical relations are accompanied by a *powla:hasChild* (resp. *powla:hasParent*) property between the parent and child node, which can also be used without *powla:Relation* for an unlabeled hierarchical relation.

For our use case, POWLA allows us to generalize over both data models (CoNLL-RDF and FLEx-RDF): The mapping of the format-specific *nif/flex* categories into POWLA categories is listed in Tab. 1.

This generalization is done by a SPARQL UPDATE script which loads an ontology providing the rdfs:subClassOf, rdfs:subPropertyOf statements for the FLEx (resp. CoNLL (Fig. 3)) categories. Using this ontology, the update replaces the original CoNLL (FLEx) data structures with

EANC, GNC (via CoNLL-RDF)	**Georgian IGT** (FLEx, via FLEx-RDF)	**POWLA**
nif:Word, nif:Sentence	flex:word, flex:phrase, flex:morph	powla:Node
nif:nextWord, nif:nextSentence	flex:next_word, flex:next_phrase, flex:next_morph	powla:next
conll:HEAD (to nif:Sentence)	flex:has_word, flex:has_phrase, flex:has_morph	powla:hasChild / hasParent

Table 1: Harmonization of corpus formats via POWLA

POWLA data structures (Fig. 4). The actual annotations of these data structures are, however, left in their original namespace, as they are extensible in the original formats/tools. Fig. 5 illustrates an extract of the data resulting from running this SPARQL UPDATE script.

```
...
    <owl:ObjectProperty rdf:about="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#next
        <rdfs:subPropertyOf rdf:resource="http://purl.org/powla/powla.owl#nextNode"/>
    </owl:ObjectProperty>
...
    <owl:Class rdf:about="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Sentence">
        <rdfs:subClassOf rdf:resource="http://purl.org/powla/powla.owl#Root"/>
    </owl:Class>
```

Figure 3: Extract of the ontology conllpowla.owl

5.2. Mapping to OLiA

After unifying the data formats by converting to an RDF format and mapping to the POWLA data structure, the values of the annotation must also be harmonized. Therefore, we needed to define mapping rules for *conll:GRAM, conll:LEX, flex:gls,* etc. to a unified annotation model. This is done by employing OLiA (REF), the Ontologies of Linguistic Annotation. It provides:

1. a modular architecture of ontologies for annotation models for different languages,
2. the OLiA reference model and
3. linking models.

The linking models connect the annotation models (1.) to the OLiA reference model with *rdfs:subClassOf* (etc.)

```
# assume that the graph <http://purl.org/powla/> contains a mapping to POWLA
# (cf. conllrdf.owl )
INSERT {
        ?a ?powlaProp ?b
} WHERE {
        ?a ?prop ?b.
        GRAPH <http://purl.org/powla/> {
                ?prop rdfs:subPropertyOf ?powlaProp.
                FILTER(contains(str(?powlaProp), "http://purl.org/powla"))
        }
};

INSERT {
        ?a a ?powlaClass.
} WHERE {
        ?a a ?Class.
        GRAPH <http://purl.org/powla/> {
                ?Class rdfs:subClassOf ?powlaClass.
                FILTER(contains(str(?powlaClass), "http://purl.org/powla"))
        }
};
```

Figure 4: Extract of the SPARQL UPDATE to complement CoNLL-RDF data structures with POWLA data structures

statements. OLiA already provides several annotation models (e.g. for the Universal Dependencies (UD)), but for Georgian FLEx, GNC, and the Armenian EANC data, we had to develop novel annotation models[18].

Since an annotation tag in all the given corpora consists of several features (e.g. "V intr"), we used the *hasTagContaining* property of the OLiA System Ontology[19] to attribute the features to its Named Individual in our annotation models (e.g. `eanc:intr system:hasTagContaining intr^^xsd:string .`). This property, however, is unsuitable for features, whose strings partially coincide with others (e.g. tr for transitive and intr intransitive). To solve this ambiguity, the hasTagMatching property with a regular expression was used instead (e.g. `eanc:tr system:hasTagMatching ^(.*)*tr( .*)*$^^xsd:string.`).

Figure 6 illustrates how the OLiA mapping for a specific tag (in this example marking a cardinal numeral) in the EANC corpus functions by linking the EANC annotation model class (EANC CardinalNumber) to its super class in OLiA (OLiA CardinalNumber).

The implementation of the OLiA mapping is done by a SPARQL UPDATE, similarly to the POWLA mapping. The update inserts unified annotations according to the corresponding annotation model. For the EANC annotation model, the query is shown in Figure 7.

The features used in the GNC (303 in total) have a shallow hierarchy. They are divided into two categories, i.e. Part of Speech, and Grammatical Features, and have been mapped to OLiA as such. Similarly, the tags used in FLEx are divided into PoS (annotated in FLEx as Word Category) and other grammatical features (annotated in FLEx as glosses). Because of the large number of superfluous features, only basic PoS features and their OLiA mapping have been used for the experiment, i.e. Verb, Noun, Modal.

The linking of our annotation models to the OLiA reference model faced certain challenges. On the one hand, the linking requires to find the OLiA category which best generalizes over a language-specific category, and an agreement between specialists of the language needed to be found. On the other hand, the OLiA coverage is by nature incomplete, and when linking a new language which contains concepts not yet covered in OLiA, its extension becomes necessary. This was the case for the Converb, appearing both in the EANC and GNC annotation model. Finally, a class in the annotation model is not always linkable to just one class in OLiA. It can be linked to multiple OLiA classes at once, or there can be several alternative OLiA classes to which one might want to link (e.g. the EANC class Determination/Possession is either a subclass of the OLiA DefiniteArticle or of the OLiA PossessiveDeterminer, but not both.). For the latter case, we use the UNION operator of the Turtle syntax. To retrieve the conjuncts of a UNION in a SPARQL query, one can either just query for the first OLiA

[18]The GNC tagset is currently under revision and will be converted to UD v.2 with some extension (personal communication with Paul Meurer in Nov. 2017). Thus, in the future, our own GNC annotation model will be replaced by the existing annotation model for UD.

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

```
@prefix : <file:///C:/Users/chiarcos/Desktop/corpus/armenian/EANC_sentences_sample///fiction.tsv#> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix powla: <http://purl.org/powla/powla.owl#> .
:s42_0 nif:nextSentence :s43_0;
       powla:nextNode :s43_0 .
:s43_0 a powla:Root , nif:Sentence .
:s43_1 a nif:Word, nif:nextWord :s43_2;
       powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "Կերուխումս"; powla:nextNode :s43_2;
       conll:WORD "Կերուխումս"; conll:GRAM "sg nom"; conll:HEAD :s43_0; conll:LEM "Կերուխումս"; conll:LEX "N inanim" .
:s43_2 a nif:Word, nif:nextWord :s43_3;
       powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "եմ"; powla:nextNode :s43_3;
       conll:WORD "եմ"; conll:GRAM "pres pl 3"; conll:HEAD :s43_0; conll:LEM "է"; conll:LEX "V intr"; conll:TRANS "be" .
:s43_3 a nif:Word, nif:nextWord :s43_4;
       powla:Terminal; powla:hasParent:s43_0; powla:hasStringValue "պոել"; powla:nextNode :s43_4;
       conll:WORD "պոել"; conll:GRAM "cvb pfv"; conll:HEAD :s43_0; conll:LEM "պոել"; conll:LEX "V tr"; conll:PUNCTR ","; conll:TRANS "do, make, make up" .
:s43_4 a nif:Word,
       powla:Terminal; powla:hasParent :s43_0; powla:hasStringValue "ուրախացել";
       conll:WORD "ուրախացել";conll:GRAM "cvb pfv"; conll:HEAD :s43_0; conll:LEM "ուրախանալ"; conll:PUNCTL ","; conll:TRANS "be glad/happy, enjoy oneself" .
```

Figure 5: Extract of POWLA annotated CoNLL-RDF data

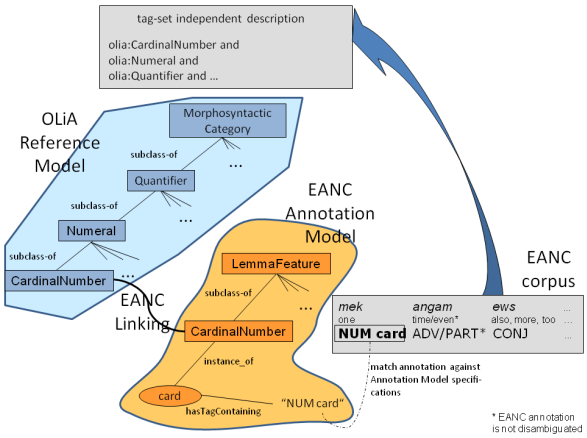

Figure 6: Visualization of the Linking of EANC and OLiA

conjunct (using *rdf:first*), for cases in which a hierarchy is defined stating that the first conjunct is the most probable, or one can extract all the conjuncts (as in the query in Figure 7). Extracting all OLiA conjuncts in order to link an annotation to all of them results, however, in the loss of the information about the conjuncts being mutually exclusive.

6. Experimental Setup

We conducted a case study on word order within auxiliary and main verb constructions. This was first applied to a part of the EANC corpus. In the future, we will replicate it on different Georgian corpora, i.e., the Georgian National Corpus (GNC) and interlinear glossed data (see Section 3.).

6.1. Pipeline

As described above, we first convert the corpora to shallow RDF-representations (CoNLL-RDF / FLEx-RDF). Then, we harmonize the data structures by transforming them to POWLA (Section 5.). This is followed by bringing the different annotation schemes of each of the corpora together through the concept linking with the OLiA tagset (Section 5.2.).

Through the harmonization of the data formats and the linking of the language specific annotations to OLiA, we are able to combine all our resources. The resulting RDF data for each of our corpora can then be queried in a unified manner. We can also add triples containing intermediate query results in order to execute advanced queries faster by using these intermediate results. In our research, we added triples containing the information about a word being an auxiliary or a main verb (according to the language specific definitions) and in a following query, we use this information to analyze the word order. The full pipeline for converting, unifying and getting experimental data is illustrated in Fig. 8.

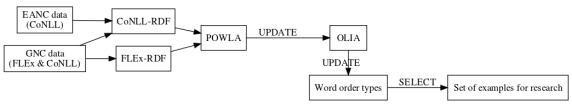

Figure 8: Pipeline of the experiment

The enriched annotation can be used for further qualitative linguistic analysis. We extract candidate sentences with a SPARQL SELECT and then study the distribution of different auxiliary / main verb ordering types manually as a preparation for a future analysis of the word order / focus implications (see Section 2.).

6.2. Scope of the Experiment

We restrict this experiment to the extraction and classification of structurally / morphologically unambiguous cases. In a future research, however, we plan to extend it to more complex sentence structures. Conceptual difficulties of our experiment are the comparability of the types of auxiliaries in the two languages and common complications in the annotation of the corpora for both languages, such as the absence of syntactic annotation and non-disambiguation on the morphological level. The problem of the (natural) shortage in the OLiA-terminology (i.e. absence of the concept converb) was solved by the extension described in Section 5.2.. The linguistic outcomes of the research are preliminary and serve only to exploring a hypothesis. A full-fledged linguistic investigation requires additional annotation efforts.

In the following, we illustrate these steps for Armenian. The pipeline scripts will be published via our GitHub repository[20] under an open license.

6.3. Filtering Clauses

We only considered sentences containing no further tokens tagged as verb beside the auxiliary and the main verb. There

[20] https://github.com/acoli-repo

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

INSERT {
        ?x a ?super
} WHERE {
        ?x a ?eanc_annot .
        GRAPH <http://example.org/eanc.owl> {
                ?eanc_annot a owl:Class .
                FILTER(contains(str(?eanc_annot), "http://purl.org/olia/eanc.owl"))
        }
        GRAPH <http://example.org/eanc_link.ttl> {
                ?c rdfs:subClassOf+/(rdfs:subClassOf*/(owl:unionOf/(rdf:first|(rdf:rest*/rdf:first)))/rdfs:subClassOf*)? ?super .
                FILTER(contains(str(?super),'http://purl.org/olia/olia.owl'))
        }
        FILTER(?c = ?eanc_annot)
}
```

Figure 7: SPARQL UPDATE for OLiA mapping (EANC)

```
PREFIX conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX olia:  <http://purl.org/olia/olia.owl#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

INSERT {
        ?a rdfs:comment "auxiliary" .
        ?b rdfs:comment "main verb" .
} WHERE {

        #get auxiliary according to definition
        ?a powla:hasParent ?s .
        ?a conll:LEM  't'; a olia:Verb .

        { ?b powla:nextNode+ ?a . } UNION { ?a powla:nextNode+ ?b . }

        #get main verb according to definition
        ?b a ?e1 , ?e2 .
        FILTER(?e1 = olia:Gerund)
        FILTER(?e2 = olia:ImperfectiveAspect || ?e2 = olia:PerfectiveAspect

        #get only simple sentences
        FILTER NOT EXISTS {
                ?another_verb powla:hasParent ?s; ?another_verb a olia:Verb
                FILTER (?another_verb != ?a && ?another_verb != ?b)
        }
        MINUS {
                ?q powla:hasParent ?s;  powla:hasStringValue ?question .
                FILTER (contains(str(?question),""))
        }
        MINUS {
                ?neg powla:hasParent ?s; conll:LEM  't'; a olia:Negation.
        }
}
```

Figure 9: Example SPARQL update for auxiliary and main verb annotation for the EANC data

```
PREFIX conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

INSERT {
        ?s rdfs:comment ?wo_comment .
} WHERE {
        ?a powla:hasParent ?s ; ?a rdfs:comment "auxiliary".

        {
        #main verb directly after auxiliary
        ?a powla:nextNode ?b .
        BIND("auxiliary directly before main verb" as ?wo_comment)
        } UNION {
        #main verb NOT directly after auxiliary
        ?a powla:nextNode/powla:nextNode+ ?b .
        BIND("auxiliary NOT directly before main verb" as ?wo_comment)
        } UNION {
        #main verb directly before auxiliary
        ?b powla:nextNode ?a .
        BIND("auxiliary directly after main verb" as ?wo_comment)
        } UNION {
        #main verb NOT directly before auxiliary
        ?b powla:nextNode/powla:nextNode+ ?a .
        BIND("auxiliary NOT directly after main verb" as ?wo_comment)
        }

        ?b rdfs:comment "main verb" .
}
```

Figure 10: SPARQL UPDATE for word order annotation

are some language specific filters to be taken into consideration in order to extract correct examples (see Section 2.3.); e.g. for the auxiliary to be in Armenian, we can only consider sentences in which this auxiliary (recognizable by its lemma (*conll:LEM*)) is combined with a main verb in certain tenses, in which it is not negated etc. A simplified SPARQL UPDATE to mark auxiliary and main verb with a *rdfs:comment* according to these filters is given in Fig. 9.

6.4. Classifying Clauses

Having added the *rdfs:comment* triples to the auxiliary and main verbs language-specifically for the EANC, GNC and FLEx data, the classification of the sentences with respect to the word order of these verbs can be done language-independently by the SPARQL UPDATE script shown in Figure 10: The word order information is also added by inserting *rdfs:comment* triples.

After annotating the selected sentences with their word order features (auxiliary directly/not directly before/after main verb) as a *rdfs:comment*, we export them to a CSV file (using a SPARQL SELECT query which filters out all sentences not annotated with a word order feature) containing the sentences themselves, their genre and their word order type including the position of the auxiliary and main verb. In such a restricted table format, a qualitative analysis of the relation between word order type and focus marking is facilitated and can be done more efficiently than in the underlying RDF format containing triples which are only relevant for the comparability/harmonization of the data and constitute redundant information in the perspective of a qualitative analyser.

7. Discussion

So far, we described the general setup of our approach, its technological components, and the data sets. While a full-fledged linguistic interpretation of our findings is beyond the scope of this paper and will be forthcoming , an evaluation in quantitative terms has been conducted.

7.1. Quantitative Evaluation

Applying the limitation and filters mentioned in Section 6.2. and 6.3., we get 20 043 classified sentences corresponding to 8.13% of the entire EANC subcorpus on hand (246,678 sentences in total). The manual evaluation consisted in examining a subset of classified sentences in order to determine the ratio of false positives and what technical and/or filter shortages caused their occurrence, if any. The distribution of the word order types among the classified sentences as well as the results of the manual evaluation are shown in Table 2.

The occurrence of false positives is mostly due to the non-disambiguated annotation of the EANC. This is especially

Word Order Type	Number of sentences	Manually evaluated sentences	Precision %
AUX V	4,993	303	95.38
V AUX	14,494	300	99.67
AUX * V	540	152	36.18
V * AUX	16	16	0
total	20,043	771	83.40

Table 2: The distribution of word order types in predicative constructions with a *to-be*-AUX in EANC and the results of manual evaluation. The * means, that there is at least one element between V(erb) and AUX(iliary)

the case with the AUX * V word order type[21]. The number of false positives is almost evenly distributed in all of the three genres (fiction, non-fiction, press)[22]. We do not calculate the recall, as it would require to manually search the remaining 226 635 sentences (91.87%) of the subcorpus. However, an analysis of 250 non-classified sentences with 23 false negatives shows that the latter are likewise in most cases due to the non-disambiguated annotation of the EANC. Less than the half of the false negatives also include further non-finite verbs. To exclude these, further filter restrictions must be considered to refine the search later on.

7.2. Conclusion and Outlook

We demonstrated how to employ LLOD formalisms to develop extraction pipelines for features and examples from diverse and heterogeneous corpora and collections of interlinear glossed text. Originally available in different formats, RDF, SPARQL and LLOD vocabularies facilitate unified access, enrichment and exploitation of such data.

After conversion from the original formats to an isomorphic, and semantically shallow RDF representation, SPARQL UPDATE can be applied to conveniently transform the original data to a common data model (here, POWLA). Similarly, SPARQL UPDATE allows to load external ontologies, and with the annotation models for EANC, IGT and GNC that we contribute to OLiA, we can follow their links with SPARQL property paths and render linguistic annotations in terms of ontological concepts.

As a result, extraction and transformation pipelines can be developed for this data, and to the extent that annotations are comparable both in terms of their hierarchical organization and in terms of their linguistic expressiveness, extraction (or transformation) scripts can be applied to other corpora in other languages.

Even after POWLA conversion, however, interpreting the original data structures is not without complications: The hierarchical nesting of *powla:Nodes* in different corpora (e.g. on the level of morphs in FLEx, but on the level of words in CoNLL-RDF) poses difficulties in following *powla:next* immediately. However, as long as we are dealing with trees, and as long as siblings (and siblings only) are always connected by a *powla:next* property, this generalized precedence operator between two variables `?x` and `?y` can be defined by the following SPARQL property path:

```
?x powla:hasParent*|powla:next+|powla:hasChild* ?y.
```

Immediate adjacency is slightly more complicated, and can be implemented by requiring that no intermediate variables exist:

```
MINUS {
?x powla:hasParent*|powla:next+|powla:hasChild* ?t.
?t powla:hasParent*|powla:next+|powla:hasChild* ?y.
}
```

As these property paths can be time-consuming, we can use SPARQL UPDATE to add a triple, say `?x my:next ?y`, for all immediately adjacent *powla:Node*s, and then use this as a shorthand in subsequent queries. This is, indeed, a key advantage of RDF, which allows to use SPARQL UPDATE to pre-compile costly expressions, thereby speeding up the eventual search process.

The impact of this functionality can only be assessed in comparison with state-of-the-art approaches in corpus linguistics: In order to generalize over different source formats, standoff XML formats (Ide and Suderman, 2007) are still considered the state of the art, but their support with off-the-shelf database technology and APIs is known to be limited (Eckart, 2008). Accordingly, corpus management systems with standoff functionality convert standoff XML to an internal, relational database scheme (Zeldes et al., 2009). However, this means that search in such systems and the retrieval of examples is constrained by a static data model and by pre-defined optimizations for a particular type of query (or lack thereof). Using RDF and SPARQL, shorthands can be introduced during the search at any point in time (if the database permits).

Our approach has been successfully implemented and described here for the study of syntactic convergence phenomena in genetically unrelated, but neighboring languages from the Caucasus area, Armenian and Georgian. It is, however, not limited to this task, and can be applied to other linguistic research questions, as well.

While this demonstrates the functionality and the technological appeal of our approach, it must be noted that SPARQL and RDF are not *a priori* linguist-friendly formalisms and technologies. One goal of our project is to facilitate the accessibility and usability of LOD technology for linguists. By demonstrating that these are viable technologies for linguistic problems, and that they allow to overcome technical barriers that currently limit the joint evaluation of available linguistic data sets in an unprecedented way, we can now motivate increased efforts in developing LOD-based infrastructures for linguistic research questions.

[21] The ambiguity is due to the fact that imperfective has the same suffix as the locative case, and infinitive and perfective of some verbs concur in form.

[22] Fiction: 32/252; non-fiction: 51/255; press: 45/256.

8. Bibliographical References

Apronidze, S. (1986). *sit̨q̇vatganlageba axal kartulši. marṭivi c̣inadadeba (Modern Georgian word order. The simple clause)*. Tbilisi Academic Press.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, et al., editors, *Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol 10318*, pages 74–88, Cham, Switzerland, June. Springer.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., and Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *TAL (Traitement automatique des langues)*, 49(2):217–246.

Chiarcos, C., Ritz, J., and Stede, M. (2011). Querying and visualizing coreference annotation in multi-layer corpora. In Iris Hendrickx, et al., editors, *Proceedings of the Eigth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 80–92, Faro, Portugal, October. Edições Colibri.

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

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Wichers Schreur, J., and Nevskaya, I. (2017). LLODifying Linguistic Glosses. In *Language, Data, and Knowledge. LDK 2017. Lecture Notes in Computer Science, vol 10318*, pages 89–103, Cham, Switzerland, June. Springer.

Chiarcos, C. (2010). Towards Robust Multi-Tool Tagging. An OWL/DL-Based Approach. In Jan Hajič, et al., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670, Uppsala, Sweden, July. Association for Computational Linguistics.

Chiarcos, C. (2012). POWLA: Modeling Linguistic Corpora in OWL/DL. In Elena Simperl, et al., editors, *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 225–239. Springer, Berlin, Heidelberg, Germany.

Comrie, B. (1984). Some formal properties of focus in Modern Eastern Armenian. *Annual of Armenian Linguistics*, 5:1–21.

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In Rainer Eckstein et al., editors, *Berliner XML Tage 2005*, pages 39–50, Berlin, Germany, September. Humboldt-Universität zu Berlin.

Eckart, R. (2008). Choosing an XML database for linguistically annotated corpora. *Sprache und Datenverarbeitung, International Journal for Language Data Processing (SDV)*, 32(1):7–22.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, USA.

Harris, A. C. (1981). *Georgian syntax: a study in relational grammar*. Cambridge University Press.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP Using Linked Data. In David Hutchison, et al., editors, *The Semantic Web – ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer, Berlin, Heidelberg, Germany.

Ide, N. and Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *1st Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic, June.

Kahnemuyipour, A. and Megerdoomian, K. (2017). On the positional distribution of an armenian auxiliary: Second-position clisis, focus, and phases. *Syntax*, 20:77–97.

Ramat, P. (1987). Introductory paper. In Martin Harris et al., editors, *Historical development of auxiliaries*, pages 3–19. Mouton de Gruyter, Berlin, New York, Amsterdam.

Tamrazian, A. (1991). Focus and wh-movement in Armenian. *University College London Working Papers in Linguistics*, 3:101–121.

Vogt, H. (1974). L'ordre des mots en géorgien moderne. In Even Hovdhaugen et al., editors, *Linguistique caucasienne et arménienne*. Norwegian University Press, Oslo.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In Michaela Mahlberg, et al., editors, *Proceedings of the Corpus Linguistics Conference*, pages 20–23, Liverpool, UK, July.

Zuo, Y. and Zuo, W. (2001). *The computing of discourse focus*. Lincom Europa, Munich, Germany.

Towards a Linked Lexical Data Cloud based on OntoLex-Lemon

Thierry Declerck[1,2]
[1]DFKI GmbH, Multilingual Technologies
Stuhlsatzenhausweg 3, D-66123, Saarbrücken, Germany
[2]Austrian Academy of Sciences, Austrian Centre for Digital Humanities (ACDH)
Sonnenfelsgasse 19, A-1010 Vienna, Austria
declerck@dfki.de

Abstract

In this paper, we present some considerations on the current state of both the Linguistic Linked Open Data (LLOD) cloud and the core module of the OntoLex-Lemon model. It is our perception that the LLOD is lacking a representation and interlinking of lexical data outside of the context of lexicons or dictionaries, which have been ported to Linked Data compliant formats. And while the OntoLex-Lemon model and its predecessor *lemon* have originally been developed to support the formal representation of language data used in ontologies, the models have been increasingly used for representing lexical entries of dictionaries and lexicons, as this can be seen in corresponding data sets included in the LLOD. As a consequence of that, we are proposing slight modifications of the core module of OntoLex-Lemon, its ontology-lexicon interface, in order to support the representation and linking of lexical data that are not necessarily included in a lexicon, a dictionary or in the terminology used in a knowledge base.

Keywords: lexical data, Linguistic Linked Open Data, OntoLex-Lemon

1. Introduction

The rapid development of the Linguistic Linked Open Data (LLOD) cloud[1] is a success story that is also based on the development of the Lexicon Model for Ontologies (*lemon*)[2] and its successor, OntoLex-Lemon[3], and experience has shown that *lemon* or OntoLex-Lemon can indeed be used for a variety of applications that are not explicitly related to ontologies, like the modelling of lexicographic data[4] or specific lexical phenomena[5].

As the possibility to develop new modules for OntoLex-Lemon is currently under discussion[6], certain aspects dealing with its core module, the "ontology-lexicon interface", seem to require some clarifications and adaptations. In this paper, we present some slight modifications to the core module of OntoLex-Lemon in order to support the deployment of a Linked Lexical Data cloud.

The suggestions we present in this context are also influenced and guided by (Gracia et al., 2017), in whose abstract we can read: "[...] future dictionaries could be LD-native and, as such, graph-based. Their nodes do not depend on any internal hierarchy and are uniquely identified at a Web scale". We can clearly see how OntoLex- Lemon is at the core of such a development, not only in the context of LD-native dictionaries, but also for Linked (stand-alone) Lexical Data.

At the same time, we are perfectly aware of the fact that *lemon*, which stands for "LExicon Model for ONtologies", was originally developed in order to model language data used in ontologies[7]. In this original context, our interpretation of "lexicon" describes the collection of language data that are included in labels or comments in ontologies, aiming to give a human-readable description of the knowledge source's content. For modelling this particular language data *lemon* is using the same formal representation language as the one deployed for the knowledge objects they describe. This approach is ultimately supporting the bridging of the knowledge of the world (or of a domain) and the knowledge of the words that are used in the same ontological environment.

However, it rapidly turned out that *lemon* and its successor, the OntoLex-Lemon model, are being used more and more for modelling digital (versions of) lexicons or dictionaries per se[8]. While this constitutes to a highly positive development, we think that a Linked Data (LD)-based lexicographic network could be independent of specific dictionaries or lexicons containing the lexical data to be represented. Quoting again from (Gracia et al., 2017): In a native LD environment "every lexical element (headword, sense, written form, grammatical attribute, etc.) is treated as a first-class citizen, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements". While the authors still anchor this view in the context of developing an "LD-based dictionary", we argue that specific dictionaries or lexicons are not necessary as container for representing lexical data in a Linked Data environment.

We consider OntoLex-Lemon as an excellent basis for reaching this goal of a Linked Lexical Data cloud, and in the next sections, we will suggest some slight modifications

[1] `http://linguistic-lod.org/llod-cloud`. See also (Chiarcos et al., 2012)

[2] See (McCrae et al., 2012)

[3] `https://www.w3.org/2016/05/ontolex/`. See also for a kind of historical view on the development of *lemon* towards OntoLex-Lemon (McCrae et al., 2017).

[4] See for example (Declerck et al., 2017), (Khan et al., 2017) or (Tiberius and Declerck, 2017).

[5] See (Declerck and Lendvai, 2016).

[6] For example describing a lexicography module for OntoLex-Lemon. See (Bosque-Gil et al., 2017) and `https://www.w3.org/community/ontolex/wiki/Lexicography`.

[7] See again (McCrae et al., 2012).

[8] See again (McCrae et al., 2017).

to be applied to its core module, the ontology-lexicon interface, in order to potentially realise our goal. However, we will first discuss some observations made regarding the current status of the LLOD.

2. Observations on the current State of the Linguistic Linked Data Cloud

When looking at the current state of the Linguistic Linked Open Data (LLOD) in detail, which is displayed in Figure 1[9], it can be noticed that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

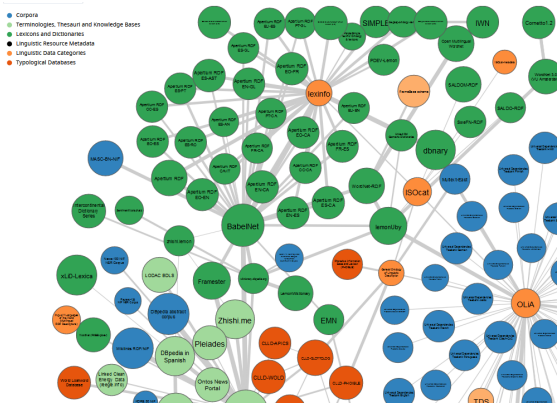

Figure 1: A partial view on the Linguistic Linked Open Data cloud, as of July 2017.

To access lexical items in the LLOD, it is easier thus to enter a lexicon or dictionary data set first and this probably reflects the meaning of the term (or ontology class) `LexicalEntry` that is used by the dictionaries or lexicons in the LLOD, which are making use of *lemon* or OntoLex-Lemon.

Here, we adopt the Wikipedia definition of "lexical entry", which states: "In lexicography, a lexical item (or lexical unit/ LU, lexical entry) is a single word, a part of a word, or a chain of words (=catena) that forms the basic elements of a language's lexicon (=vocabulary)."[10].

The question now is if the term (or ontology class) `ontolex:LexicalEntry` in OntoLex-Lemon only has a "lexicographic" acceptance (i.e an entry has to be part of a dictionary or a lexicon), and this applies even more if we consider that the modelling of language data that occur in the labels of a taxonomy or an ontology is done without taking into consideration any dictionary or lexicon.

We think that in this respect, the core module of OntoLex-Lemon should be clearly distinguished from the definition of a lexical entry that is to be provided by the upcoming lexicography module, which is currently being discussed within the W3C Ontology-Lexica community[11]. The participants in this discussion are perfectly aware of this issue, as they are suggesting the name "DictionaryEntry" to represent the structure of an entry in a (mostly non-LD-native) dictionary, and thus differentiating it from a "LexicalEntry", which is modelling a lexical item that is not necessarily included in a lexicographic work. This is in fact the view supported by OntoLex-Lemon, as the information about the naming of a collection (possibly a lexicon) of lexical items is left to the "LInguistic MEtadata" (*lime*) module, which describes metadata as related to the lexicon-ontology interface.[12].

Now turning our attention back to the analysis of the LLOD again, we consider the example of the aggregated RDF Apertium bi-lingual dictionaries[13] in greater detail. For the RDF version of Apertium, Spanish lexical data that were originally contained in different bi-lingual dictionaries have been merged into one data set and lexical entries of the source and the target languages are pointing to the same BabelNet synset[14]. BabelNet is developing a hub for references to senses and encyclopaedic sources in the LLOD. Having a source language word and a target language word pointing to the same BabelNet meaning (or sense) can therefore be considered a good way to indicate their appropriateness for a translation relation. This is a good case where we can see the benefit of the LLOD approach to the modelling and linking of language data.

At the same time, the Italian lexical data included in RDF Apertium, in its bi-lingual Catalan-Italian dictionary, does not have any link to the Italian data included in the SIMPLE lexicon[15]. Furthermore, SIMPLE is not linking to BabelNet, but to another source containing senses. The direct question is here: why do we have two "entries" for one and the same (Italian) word, in SIMPLE and in Apertium[16]?

When looking at the corresponding RDF Apertium and

[9] The full LLOD cloud can be accessed at `http://linguistic-lod.org/llod-cloud`. There, one can click on the various nodes and get more details about the data sets represented by the "bubbles".

[10] See `https://en.wikipedia.org/wiki/Lexical_item`.

[11] See (Bosque-Gil et al., 2017) and the on-line discussion at `https://www.w3.org/community/ontolex/wiki/Lexicography` for more details.

[12] See (Fiorelli et al., 2015) and `https://www.w3.org/2016/05/ontolex/\#metadata-lime` for more details.

[13] Apertium is an open-source machine translation platform (see `https://www.apertium.org/index.eng.html`).For the porting of Apertium resources to *lemon* and their publication on the LLOD, see `https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/`.

[14] See `http://babelnet.org/`.

[15] See `http://catalog.elra.info/product_info.php?products_id=881` for the original SIMPLE lexicon.

[16] In Apertium, information related to the word "bocca" (mouth)

SIMPLE data in the LLOD, the reader can observe that there are enough similar elements in each representation of the same lexical element. The main difference resides in the (way of) linking to a source representing the corresponding sense(s). One can ask then if, similar to the successful merging exercise done in the case of the monolingual Spanish lexicon in RDF Apertium, it would not be possible to merge all the triples into the LLOD dealing with the Italian form "bocca". In doing so, the merging would not lead to a specific lexicon, but to a data set itself, containing or linking to all related data or information/knowledge. This is basically what we would understand by a Linked Lexical Data cloud.
In the following section, we propose a short analysis of the current version of OntoLex-Lemon.

3. Observations on the current State of OntoLex-Lemon

The graphical view presented in Figure 2 demonstrates the organisation of the core module of OntoLex-Lemon: the "ontology-lexicon interface" (ontolex).

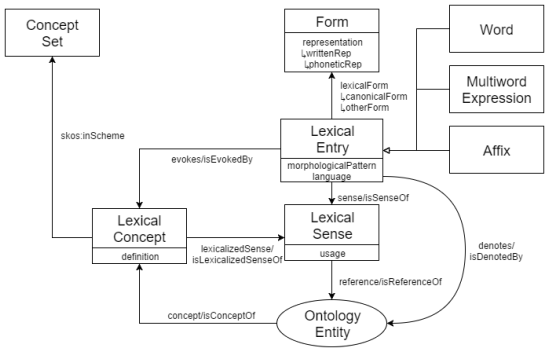

Figure 2: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

Looking now at the `LexicalEntry` class, it can be noticed that this class has kind of a pre-eminent position, which is not due to its central position in the graphic. The pre-eminence I see is the fact that none of the other elements in the field of morphosyntax information has a relation to sense, conceptual or referential resources. Therefore, they all have to "communicate" first with the class `LexicalEntry`. But as quoted before from (Gracia et al., 2017), we would prefer to see all elements of the model being first-class citizens. As a consequence of that, the resulting question is why an instance of a `ontolex:Form`, for example, cannot have a property linking to a sense or to an ontological reference.
One example which has recently been discussed[17] was the Spanish word "cura", which in English can mean "cure", when used in feminine, or "priest" (or similar), when used in masculine. One option for this would be to introduce two separate lexical entries with their corresponding canonical form and sense(s). Like this, the introduction of an instance of `LexicalEntry` would not only be motivated by the part of speech of the word to be represented, but also by its gender. And in addition to that, the sense would play a role in the decision on adding an entry or more for one word. I see in this a weakening principle of the modularity principle existing between the fields of lexical entries and lexical senses. An alternative solution would be to have only one "entry" for the Spanish noun "cura" and to allow the different canonical forms (one in feminine, one in masculine) to have a direct link to the corresponding sense(s). This way, we do not duplicate the number of entries, while keeping the same number of forms, and the OntoLex-Lemon elements (or classes) `LexicalEntry` and `Form` are being treated equally.
We extent this question to elements of the "Decomposition" module[18], which is displayed in Figure 3. This module supports the representation of components of a decomposed compound word or the components of a multi words expression.

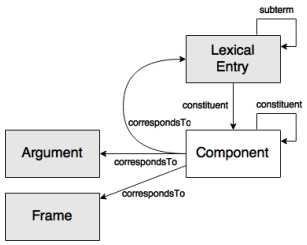

Figure 3: The Decomposition Module of OntoLex-Lemon. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

The cases we were dealing with are the German words "Erdöl" (*oil*) or "Erdgas" (*gas*) on the one hand and "Erdbeer" (*strawberry*) on the other. After decomposition, we have the components "Erd", which can be linked via the property `correspondsTo` to the OntoLex-Lemon entry "Erde" (*earth*), but it can be observed that "Erd" on its own is not a correct word or form of German. In one case, we now need to link the meaning of the component "Erd" to the sense representing a geological surface that needs to be drilled in order to extract oil (or gas), and in the other case to an agricultural sense of "Erd". We do not see how to do this if one has to link to the corresponding "Erde" entry first and we do not want to augment the number of "Erde" entries for this. Another option would be to add a restriction pointing to the corresponding component in the sense description, but then, we would have a direct link between a sense and a component (which is not a lexical entry or even a lexical form).

is available at `http://linguistic.linkeddata.es/page/id/apertium/lexiconIT/bocca-n-it` and in SIMPLE at `http://www.languagelibrary.eu/owl/simple/psc/pscLemon.ttl`.

[17] `https://www.w3.org/community/ontolex/wiki/Lexicography`.

[18] `https://www.w3.org/2016/05/ontolex/\#decomposition-decomp`.

There seem to be enough cases that call for a loosening of the current restriction allowing that only a `LexicalEntry` can be linked by a property to a `LexicalSense`, a `LexicalConcept` or even an ontological reference.
In doing so, the model would be very close to the already quoted statement that "every lexical element (headword, sense, written form, grammatical at- tribute, etc.) is treated as a first-class citizen, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements" (Gracia et al., 2017).

4. About the Status of `LexicalEntry` in OntoLex-Lemon

The discussion about the problematic cases resulting from the fact that the class `LexicalEntry` is playing a central (or pivotal) role as an intermediate between morpho-syntactic and semantic descriptions of lexical data leads to the fundamental question about its status within the model. Looking at many examples of encoding of entries with *lemon* or OntoLex-Lemon, one gets the impression that an instance of the `LexicalEntry` class is in fact a grouping of related word forms, based on their shared Part-of-Speech information. Is this the case, the labelling of the class in term of `LexicalEntry` would be misleading. I am wondering if in such a graph-based model, in which all nodes are to be considered "first-class citizens" (Gracia et al., 2017), such a class as `LexicalEntry` is still needed. In fact, the labelling of this class seems to be a reminiscence of non LD-native dictionaries, in which the access to lexical data was guided by lexical entries, that were organized by extra-linguistic principles, as this is for example the case for the alphabetic ordering of entries, which is "an arbitrary system which brings together completely unrelated words in sequences like: *redneck*, *redness*, *redo*, *redolent*, *redoubtable*" (Rundell, 2015).

5. Towards a Linked Lexical Data cloud

As certain professional lexicographers are aiming at an e-lexicography beyond dictionaries[19], is it not appropriate to also consider an e-lexicography beyond lexical entries, but dealing only with lexical data that can be directly linked to each other in a huge network, which we would like to call the Linked Lexical Data cloud. In this cloud the different lexical data could be linked not only to each other but also to other types of data, and be directly integrated in LLOD-based applications. One could also aim at merging lexical data and so to reduce redundancies of data descriptions.
In this Linked Data Lexical cloud, both the users and Natural Language Processing applications would have direct access to the needed lexical information, responding thus to a certain extend to the needs formulated by publishers and other professionals in the e-lexicographic field. (Køhler Simonsen, 2017) for example stresses the fact that "[...] the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people in fact do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead. So dictionaries are in fact not being used as much as we want them to be. The most important question is: why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks". In order to be able to implement business models for the modern e-lexicography, (Køhler Simonsen, 2017) requires among others that lexicographic products are moving to lexicographic services, the integration of lexicographic data in lexicographic platform and distribution, and to take increasingly into account the lexicographic users and their needs, and in general a move "from dictionary to lexicographic data in software [and] artificial intelligence". The intended Linked Lexical Data cloud could be instrumental in reaching those goals.

[19]We borrow this expression from the title of a talk on "post-dictionary lexicography" given by Ilan Kernerman at eLex 2017, available at `https://www.youtube.com/watch?v=yA3yg6wO5M8`.

6. Conclusion

We presented some considerations about the current state of the Linguistic Linked Data (LLOD) cloud and the OntoLex-Lemon model, which is a core component of the LLOD. As we argue that within the LLOD it would be beneficial to have a formal representation and a dense linking of lexical data that are not necessarily included in a lexicon or in a dictionary-based data-set, we end up in suggesting slight modifications of the OntoLex-Lemon model, also on the base of the discussion of some examples that are difficult, if not impossible, to model with the current version of OntoLex-Lemon. While the suggested modifications of OntoLex-Lemon are minimal, they lead to a fundamental question on the status of the textttLexicalEntry class, which ultimately could be made optional or disappear, at least in the context of the intended Linked Lexical Data cloud. We presented also briefly some views proposed by professionals in the field of lexicography publishing, and which are in line with our consideration that dictionaries are no longer needed as container of lexical data in a Linked Data-based framework.

Acknowledgement

The DFKI contribution to this work has been partially funded by the German Federal Ministry of Education and Research under the funding code 01-W17001 for the project "DeepLee - Tiefes Lernen für End-to-End-Anwendungen in der Sprachtechnologie". Responsibility for the content of this publication is with the author. The ACDH contribution to this work has been partially funded by the H2020 project "ELEXIS" with Grant Agreement number 731015. We would like to thank the anonymous reviewers for their very helpful comments on the original submission of this paper. we also thank Eileen Schnur for her proof-reading on the first three sections of this paper.

7. Bibliographical References

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a module for lexicography in ontolex. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 74–84.

Chiarcos, C., Hellmann, S., and Nordhoff, S., (2012). *Linking Linguistic Resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer Berlin Heidelberg, Berlin, Heidelberg.

Declerck, T. and Lendvai, P. (2016). Towards a formal representation of components of german compounds. In Micha Elsner et al., editors, *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Humboldt University, ACL, 8.

Declerck, T., Tiberius, C., and Wandl-Vogt, E. (2017). Encoding lexicographic data in lemon: Lessons learned. In John P. McCrae, et al., editors, *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. CEURS, 8.

Fiorelli, M., Stellato, A., McCrae, J., Cimiano, P., and Pazienza, M. T. (2015). LIME: the Metadata Module for OntoLex. In *Proceedings of 12th Extended Semantic Web Conference*.

Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 550–559. INT, TrojÃna and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Khan, F., Bellandi, A., Boschetti, F., and Monachini, M. (2017). The challenges of converting legacy lexical resources to linked open data using ontolex-lemon: The case of the intermediate liddell-scott lexicon. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 43–50.

Køhler Simonsen, H. (2017). Lexicography: What is the business model? In Iztok Kosem, et al., editors, *Electronic Lexicography in the 21st Century*, pages 395–415. Lexical Computing CZ s.r.o.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, 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):701–719.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In Iztok Kosem, et al., editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos*, 25(1).

Tiberius, C. and Declerck, T. (2017). A lemon model for the anw dictionary. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 237–251. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.