Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse

A. Parvizi, M. Kohl, M. Gonzàlez, R. Saurí
Oxford University Press
Oxford, United Kingdom
{artemis.parvizi, matt.kohl, meritxell.gonzalezbermudez, roser.sauri}@oup.com

Abstract

The Dictionaries division at Oxford University Press (OUP) is aiming to model, integrate, and publish lexical content for 100 languages focussing on digitally under-represented languages. While there are multiple ontologies designed for linguistic resources, none had adequate features for meeting our requirements, chief of which was the capability to losslessly capture diverse features of many different languages in a dictionary format, while supplying a framework for inferring relations like translation, derivation, etc., between the data. Building on valuable features of existing models, and working with OUP monolingual and bilingual dictionary datasets, we have designed and implemented a new linguistic ontology. The ontology has been reviewed by a number of computational linguists, and we are working to move more dictionary data into it. We have also developed APIs to surface the linked data to dictionary websites.

Keywords: lexical ontology, knowledge reuse, linked data, under resourced languages

1. Introduction

For years, dictionaries, thesauri, and morphological resources were only intended for human consumption and, therefore, inaccessible to machines. Many Natural Language Processing (NLP) tasks such as machine translation, natural language generation, and word sense disambiguation require machine readable lexical resources. These lexical resources must be able to provide some level of interlinking between the data; WordNet[1] is one such example. The structure and the combination of resources (dictionary and thesaurus) that a semantic network like WordNet presents have enabled many NLP tasks to be done at scale. The existence of WordNet as a successful resource encouraged the creation of WordNet in many languages other than English, as well as resources such as FrameNet[2], VerbNet[3], and BabelNet[4] with different functionalities to WordNet.

Re-structuring these lexical resources using Semantic Web standards such as `RDF` and `OWL` offers even more opportunities. An example would be BabelNet, which uses the lemon model (Ehrmann et al., 2014) as a lexical ontology. As a result, the linked data version of BabelNet builds on its many multilingual resources by linking together various datasets into a single resource, with added emphasis on supporting the Semantic Web community and linked data-based NLP applications like word sense disambiguation.

A number of specific benefits of modelling linguistic resources as linked data have been identified (Chiarcos et al., 2013), such as: (i) structural interoperability, (ii) federation, (iii) enhanced conceptual interoperability, (iv) a rich ecosystem of formalisms and technologies, and (v) dynamic import (the possibility of creating resolvable links between resources that are maintained by various data providers). As described in (Chiarcos et al., 2013), the challenge mainly arises from *information integration*; how different types of resources can be combined in an efficient way. An *upper ontology* is one means for various natural languages to be able to communicate, but upper ontologies are often sparse (Hirst, 2009), and they tend to favour one language over others in schema construction, which can diminish their utility in multilingual applications.

2. Motivation

The Dictionaries division at Oxford University Press (OUP) recently launched the Oxford Global Languages (OGL) initiative to create and publish linguistic resources for many languages around the world focussing on digitally under-represented languages. The aim is to help language communities around the world create, maintain, and use digital language resources for their language, while developing digital-ready content formats to support the growing language needs of technology companies worldwide.

OGL data is diverse. It consists of lexical datasets (dictionaries, thesauri, morphologies, etc.) as well as crowd-sourced content from communities of users participating in language games, forums, and submissions. With an eye to integrating this content, we have built our infrastructure using Semantic Web technologies. A key factor for the project's success is having a model that can accommodate complexities across these many languages as well as enable linking between them. For this we need a *lexical ontology*.

Some of the main requirements for the lexical ontology are

(i) extract inflected forms and their grammatical features
(ii) extract translations of a headword[5] in a given language
(iii) extract main and alternative spellings of headwords
(iv) extract alternatives to a headword such as synonyms and antonyms
(v) extract the most common sense of a headword
(vi) extract domain, register, and other semantic content
(vii) extract derivation and historical information of headwords
(viii) extract regional vocabulary
(ix) extract senses and homographs in a certain order
(x) differentiate among direct and near translations (e.g., for idioms)

[1] `https://wordnet.princeton.edu`
[2] `https://framenet.icsi.berkeley.edu`
[3] `http://verbs.colorado.edu/verb-index`
[4] `http://babelnet.org`
[5] The word beginning each separate entry in a dataset.

3. Background

A number of lexical ontologies have been published over the years. These include:

- DOLCE, the developers of which had the vision of creating a foundational ontology that does not intend to be a universal ontology. DOLCE "aims at capturing the ontological categories underlying natural language and human common-sense." (Gangemi et al., 2002)
- GOLD is the first ontology specifically built to represent linguistic knowledge on the Semantic Web, and is thus the first ontology to demonstrate the power of reasoning in linguistics (Farrar and Langendoen, 2003).
- SUMO is an upper ontology created from merging publicly available ontological content into a single ontology (Niles and Pease, 2003). It contains multilingual content and may be used as a lexical ontology, as WordNet has done.
- GUM is a linguistic ontology intended for natural language generation tasks, and as much as is possible, remains multilingual. Although it has portions in common with GOLD, it is not as powerful in reasoning as GOLD. In short, "GUM is an attempt at an intermediate level of abstraction bridging the gap between linguistic and non-linguistic knowledge" (Farrar and Langendoen, 2003).
- OLiA ontologies are based on the OLiA reference model, which is rich in morphology, morpho-syntactic features, and syntax. This model has incorporated tagsets to annotate various linguistic dimensions from around 70 different languages (Chiarcos, 2010).
- lemon was developed to be able to model rich linguistic resources such as WordNet and to allow these resources to be shared and extended. A few of its crucial features are (McCrae et al., 2011): it is based on `RDF(S)`; it does not prescribe the usage of particular categories and properties in linguistics; and it supports the usage of other linguistic ontologies such as GOLD. For modelling grammatical features, lemon relies on grammatical framework[6] (GF). As GF is a sophisticated system incorporating theories of grammar and language, it might not be the most efficient way to represent simple grammatical features.
- LexInfo is a declarative and application-independent attempt to map languages to ontologies (Cimiano et al., 2011). The aim of this project has been to support more than `RDF(S)` and incorporate `OWL` and `SKOS`. LexInfo has simplified many branches of OLiA such as `Morpho-syntactic Feature`. In addition, some of the modelling elements of LexInfo have been incorporated into lemon.
- OntoLex[7] is a community attempt at creating a lexical ontology for presenting machine readable dictionaries. By and large, OntoLex re-purposes lemon and LexInfo, as well as incorporates a few other extensions to lemon.

4. Why a New Lexical Ontology?

The lexical ontologies listed in Section 3. are all valuable resources, and we have learnt from and reused some of the concepts and properties described in them. DOLCE, GOLD, SUMO, and GUM have been excellent starting points. GOLD was developed to capture a linguist's knowledge; thus, it offers depth in places that is not really needed for modelling lexical resources (e.g., the `LinguisticDataStructure` branch). Other resources such as DOLCE and GUM have a focus on modelling spatial or temporal aspects that again were not necessary for us. The complexity that these added features offer was not negligible; therefore, we selectively isolated concepts and properties that were fit for our task.

In short, we discovered that none of the existing linguistic ontologies were developed to *only* support modelling dictionary, morphology and thesaurus data. Since the scope of these ontologies was greater than what we needed, the requirements that we had gathered were not fully covered, and finally due to the structure and the content of the legacy data, the decision was made to develop a new dictionary ontology.

To satisfy our particular use case, we needed to design a new lexical ontology. The first draft was developed based on the following assumptions:

(i) the ontology must support resolving IRIs across various *multilingual* lexical resources such as dictionaries, thesauri, and morphologies, in order to facilitate knowledge discovery;

(ii) the lexical resources must be able to communicate on an abstract level, thus a need for an upper ontology;

(iii) the lexical resources must be as granular as possible, thus a need for a more detailed ontology for each resource that inherits from the upper ontology;

(iv) user requirements must be expressed in SPARQL queries in order to function as competency questions;

(v) the ontology should facilitate reuse by adding cross-ontology annotations (e.g., to GOLD, lemon, and etc.);

(vi) the ontology must be able of accommodating external resources like WordNet, FrameNet, and BabelNet;

(vii) the ontology, if possible, must use standardised vocabulary such as SKOS; and

(viii) the ontology must allow progressive addition of new classifications.

The first draft of the ontology, along with sample data, was reviewed by three computational linguists with a background in logic. The feedback received was incorporated in the second draft of the ontology.

Our next step was to run experiments on the ontology with various datasets to assess its performance. We loaded the ontology as well as monolingual and bilingual dictionaries,[8] morphology datasets, and the English WordNet 3.0 into a triplestore. Based on the performance of the triplestore and the structure and performance of the data, some adjustments were made to the model.

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

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

[8] Datasets added: English, Spanish, English-Spanish, Spanish-English, English-isiZulu, isiZulu-English, English-Northern Sotho, and Northern Sotho-English.

The model, as well as re-structuring dictionary content, is capable of (i) generating cross-language translations, (ii) generating cross-language classical semantic relations (e.g., synonymy and hypernymy), (iii) providing domain and register information, (iv) linking etymologies to headwords and dates of origin, (v) providing pronunciations, (vi) displaying morphological features, (vii) automatically identifying homograph and homophone words, and (viii) linking common examples across languages.

5. The OGL Ontology

This section describes key aspects of the OGL ontology. First, the way the ontology models cross-lingual grammatical information while, at the same time, retaining the particularities of each language (Section 5.1.). Then, we explain how we model sense-level translations and how we enable translation among all available languages using English as an interlingua (Section 5.2.). And finally, we explain to we have leveraged the power of OWL, SKOS and PROV to reason over our data (Section 5.4.).

5.1. Grammatical Information

A core part of the ontology is the area modelling part of speech (POS) and other grammatical features, which needs to be able to convey grammatical features in languages belonging to very different families; it must support distinctions not only specific to dictionary contents but also applicable to data generated in other contexts, for example by language processing tools. Next we detail the design criteria followed (Section 5.1.1.), then we present the main features of this part of the ontology (Section 5.1.2.), and finally we will discuss other grammatical distinctions (Section 5.1.3.).

5.1.1. Design Criteria

Cross-linguistic validity. A necessary requirement is to be able to function across different languages; however, languages can diverge greatly in the way they encode grammatical distinctions. Features expressed in some languages via morphological mechanisms (for example, verbal tense in Romance languages), may be encoded in others by means of independent particles, and yet in others may not be expressed at all (e.g., Chinese). As a result, dictionary-based classifications of POS classes and associated grammatical information tend to be modelled based on the language (or languages) targeted in each project. A POS tagset for English, for example, may present phrasal verbs as independent POS, or may not differentiate between reflexive and reciprocal pronouns as this is not a relevant distinction in that language.

Due to the context of use of the OGL Ontology (at the service of a truly multilingual, wide-scoping linguistic repository), it was crucial that no language was a stronger driver than any other. Thus, we put forward a minimal set of POS tags that were as universally valid as possible and general enough to serve languages of very different typology. Language-specific features are left to be encoded by means of additional grammatical classifications, complementary (and therefore orthogonal) to the basic POS categorization.

A fully compositional approach. Common approaches to modelling grammatical information tend to make grammatical features subsidiary to particular POS tags. In other words, each POS tag is further subclassified based on the morpho-syntactic distinctions that it bears in the language inspiring that classification. For example, nouns in German will be divided into masculine, feminine and neuter, whereas adjective forms in English can be classified into positive, comparative, and superlative.

Nevertheless, a system where grammatical features are tied to particular POS tags will fail the purpose of being valid across languages. First, each POS will have to subdivide into as many subclassifications as found across the different languages covered. For example, pronouns would split at least in the following subcategories, some of which may in addition intersect:

> *absolute, exclusive, inclusive, expletive, reciprocal, reflexive, demonstrative, exclamative, existential, indefinite, interrogative, personal, possessive, relative, clitic.*

Second, the system will introduce redundancy since some grammatical distinctions are shared across POS classes. For instance, the distinction in degree (positive, comparative, superlative) can be found in adjectives and adverbs.

Alternatively, one can think of a minimal set of POS tags as universally valid as possible (very much along the lines of current work like the Universal Dependencies[9] proposal), together with a set of classifications for grammatical distinctions that can potentially apply to different POS classes depending on the language. The OGL ontology classification system follows this compositional approach (see Table 1).

Respectful to the grammatical tradition for each language. Each language is supported by its own grammatical tradition, which is reflected in the way it is explained and taught in grammar books, dictionaries, etc. A central aspect of work at OUP is precisely producing dictionaries for different languages, and therefore the linguistic classifications used should be in agreement with those commonly assumed in the grammatical tradition of each language.

Grammatical classifications in each tradition may however be too constrained to the language (or group of languages) they aim to explain, therefore precluding a wider, cross-linguistic view of grammar distinctions. For example, the marker for *concord*, commonly considered as an independent POS in Bantu languages, can actually be classified as *affix* (more specifically, *prefix*).

In order to respect both perspectives (namely, the cross-linguistic one and the one associated to grammar traditions in each language), the ontology put forward here is to be used for the overall classification and linking of linguistic information across languages. On the other hand, each specific application resorting to such content (e.g., dictionaries, computational lexicons, automatic tools for language processing, etc.) can map the present classifications to whatever tagset is deemed most suitable for its purpose[10].

[9] `http://universaldependencies.org/`

[10] As a matter of fact, we have already developed mappings be-

	POS	Explanation
1	*adjective*	
2	*adposition*	Closed class of words that express spatial or temporal relations, or mark various semantic roles. Typically combining with one complement, generally a noun phrase. Divided into the following 3 subclasses based on the position they take with respect to the complement: before (*preposition*), after (*postposition*) or surrounding it (*circumposition*).
3	*adverb*	
4	*affix*	Morphemes attached to a stem to form a new word. In some cases it is written as part of the same word whereas in others it appears as an independent element. Divided into subclasses: *circumfix*, *combining_form*, *infix*, *prefix*, *suffix*.
5	*article*	
6	*conjunction*	
7	*contraction*	Combination of two or more words belonging to a different POS into a single lexical unit. For example, the combinations of preposition + article in Spanish: *al* (*a*+*el*), *del* (*de*+*el*).
8	*determiner*	
9	*ideophone*	Lexical units that evoke a vivid impression of certain sensations or sensory perceptions (e.g., sound *meow* for a cat), movement, color (e.g., English *bling*, describing the glinting of light on things like gold), shape, action (*ta-da!*), etc. It is a lexical class based on the special relation between form and meaning. In some languages, ideophones correspond to common POS classes (e.g., adjectives, adverbs, etc.), but in others, like English, they are an independent POS.
10	*idiomatic*	Multiword, phrasal or clausal expressions, generally with no compositional interpretation.
11	*interjection*	
12	*noun*	
13	*numeral*	
14	*particle*	Typically encoding grammatical distinctions like negation, mood, tense, or case, etc. They must be associated with another word or phrase to impart meaning, and cannot be classified as other main POS, including functional ones, such as prepositions, conjunctions, etc.
15	*predeterminer*	
16	*pronoun*	
17	*punctuation*	Subclasses: *left_parenth_punc*, *right_parenth_punc*, *sentence_final_punc*, *sentence_medial_punc*.
18	*residual*	Cover class for non-standard forms, such as acronyms and abbreviations.
19	*verb*	

Table 1: POS classification

Satisfying requirements from both lexicography and technology teams. The OGL Ontology was developed to support the representation and storage of information from different sources, ranging from dictionary content to data obtained by automatic means from processing naturally occurring text. Because of that, the ontology contains grammatical features typically used in dictionaries (e.g., POS, subcategorization pattern, gender) but also distinctions necessary to tag content by automatic procedures; for example, a classification of punctuation marks, employed by POS-taggers and parsers.

5.1.2. POS Classification

To satisfy our multilingual requirements, we opted for a minimal set of high-level POS classes that would enable a smooth integration of new languages to the model. For that, again, we followed the spirit of the Universal Dependencies initiative and the proposal of Google universal POS tags (Petrov et al., 2012).

Overall, the OGL Ontology shares 13 POS categories with the POS tagset from the Universal Dependencies (UD) project. There are 6 tags that are only present in our ontology: *affix*, *contraction*, and *idiomatic* (all of them needed here because they are elements present in dictionary and grammar content); *article* and *predeterminer* (both subsumed under *determiner* in UD); and *ideophone*, not accounted for in that ontology. On the other hand, the UD POS tagset has categories not available in our proposal: *auxiliary verb*, *proper noun*, and *subordinating conjunction* (which are respectively classified as *verb*, *noun* and *conjunction* in the OGL Ontology, and subclassified via further grammatical distinctions—see next section), and *other*[11].

5.1.3. Other Grammatical Distinctions

Complementary to the high-level POS classification, we introduced an extensive categorisation for grammatical features. Thus, beyond typical morpho-syntactic categories such as gender, number, case, or tense, we expanded on additional linguistic distinctions that can manifest at the lexical level (e.g., event modality, marker type, aspect).

The rich morpho-syntactic branch in OLiA was a solid base to model this kind of information. However, some of the languages that we intend to model have complex morpho-syntactic features, such as Northern Sotho, wherein a single orthographic word may contain a number of morphemes; and some might have extensive noun systems, such as isiZulu, which has 17 different classes. Neither OLiA nor lemon were able to fully accommodate modelling of such features cf. (Chavula and Keet, 2014).

Our proposal was also informed by the lexical and grammatical labels used in OUP monolingual and bilingual dictionaries, in this way addressing the requirement to be able to represent grammatical information typically expressed in dictionaries, e.g., subcategorisation patterns (transitive, intransitive, reflexive), particle types (infinitive, emphatic,

tween the OGL Ontology and each of our bilingual dictionaries.

[11] The aim is to tag inputs that users and editors are unsure about as other and ask a native editor to correctly tag the POS.

interrogative, etc.), phrase types, etc. Finally, a further source of information were other well-tested classifications for morpho-syntactic knowledge developed with multiple languages in mind or as a collaborative effort among teams in different countries (e.g., EAGLES[12] and, MULTEXT[13]). The OGL Ontology distinguishes a total of 46 classifications for grammatical information, which range from the basic nominal, verbal and adjectival morphological features present in many languages (e.g., *number, gender, case, degree, person, mood, tense*, etc.) to elements codifying syntax (e.g., *subcategorization* patterns), lexical semantic distinctions (e.g., *aspect, telicity, countability*), or pragmatic information (e.g., *definiteness, referentiality, evidentiality, sentence modality*). Some of these categorizations are present in multiple POS classes, whereas others are particular to only one. Furthermore, most of them are shared across several languages, although a few cases had to be tailored to specific ones, such as the classifications on *diptoticity* or *verb form* type for Arabic.

5.2. Cross-Language Translation

Since lemon and LexInfo have been used to model BabelNet, we analysed them carefully and reused some of their concepts and properties (e.g., class `Translation`). However, the modelling decisions made in lemon and LexInfo did not provide enough scope to enable us to perform cross-language classical relation (e.g., synonymy) mapping. Although some of these issues have been addressed in OntoLex's translation module [14], based on our requirements, and the in-house legacy data, OntoLex's translation module is over engineered for our purpose, therefore, we aimed at simplifying this model.

Much of the dictionary data we are dealing with has been digitized from books and therefore requires some gap-filling due to print conventions. For example, a printed bilingual dictionary will give sense [15] translations by supplying the relevant headword from the target language. In order for this to become maximally useful data, though, the URI of the translated sense is preferable, so as to avoid having to disambiguate in the case of polysemous target entries. In other words, a link between two sense level concepts is better than a link among sense level and word level concepts.

To mitigate this issue, we considered translations as yet another headword in a dictionary;

1. if the translation was editorially already at sense level, having a headword structure in place allows us to extract every bit of information available for that particular sense of the headword;
2. if the translation was not at sense level, we automatically generate a headword-like structure for the translation (see Figure 1). This approach results in the creation of a number of synthetic headword-like structures that need to be mapped to actual senses of headwords.

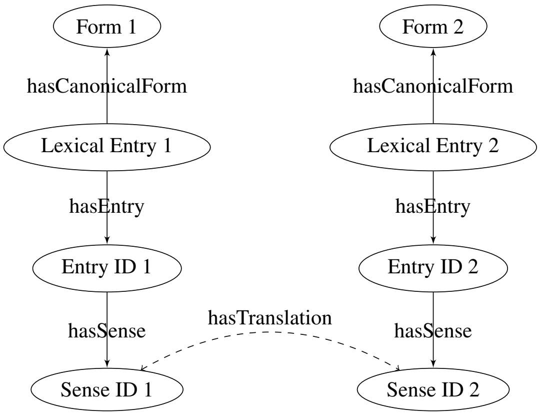

Figure 1: Sense to sense translation

The scenario in (1) will result in accurate and straightforward translation extraction. In contrast, the scenario in (2), needs a considerable amount of work to reconcile the synthetic sense to an editorially curated sense; as some language pairs, such as English–Spanish datasets, may have multiple senses for a given headword, such as *book*.

This strategy, with the aid of a common interlingua (in our case English), results in making cross-language translations available. Cross-language translation is crucial for creating new language resources in languages with limited digital resources. For example, assuming English to isiZulu and an English to Northern Sotho datasets exist, we are able to extract isiZulu to Northern Sotho translations. However, this strategy might not always be accurate; as, a word might have multiple senses, thus isolating a sense to sense translation without some post-processing on data may not be possible; or even in the case of having only one sense in both mentioned dictionaries, the senses might not represent the same meaning. We are in the process of introducing new methods to *improve and evaluate* the accuracy of this cross-language relation discovery. We believe that we not only need to enhance the model to be capable of tackling this issue but also, might need to add other functions to the NLP processes.

5.3. Etymology

The Oxford English Dictionary[16] (OED) is a dictionary of the English language that covers the meaning, history, and pronunciation of over $600,000$ words drawn from over $1,000$ years of English as used in a variety of countries. Indeed, OED is a rich diachronic resource used as a fruitful collection of historical data in several linguistic studies.

Another avenue that we would like to explore in the near future is modelling etymological information. As existing linguistic ontologies do not extensively cover etymological information, as a first step, we have relied on OED to help us identify the most important features. From an etymological point of view, OED contains a large amount of detailed

[12] http://www.ilc.cnr.it/EAGLES/annotate/annotate.html

[13] http://www.tei-c.org/Activities/Projects/mu03.xml, http://nl.ijs.si/ME/

[14] `http://www.w3.org/community/ontolex/wiki/Translation_Module`

[15] A unit of meaning in a dictionary or thesaurus, which can consist of a definition, translation, or set of synonyms, and further information like example sentences and markers for region or register.

[16] `http://www.oed.com`

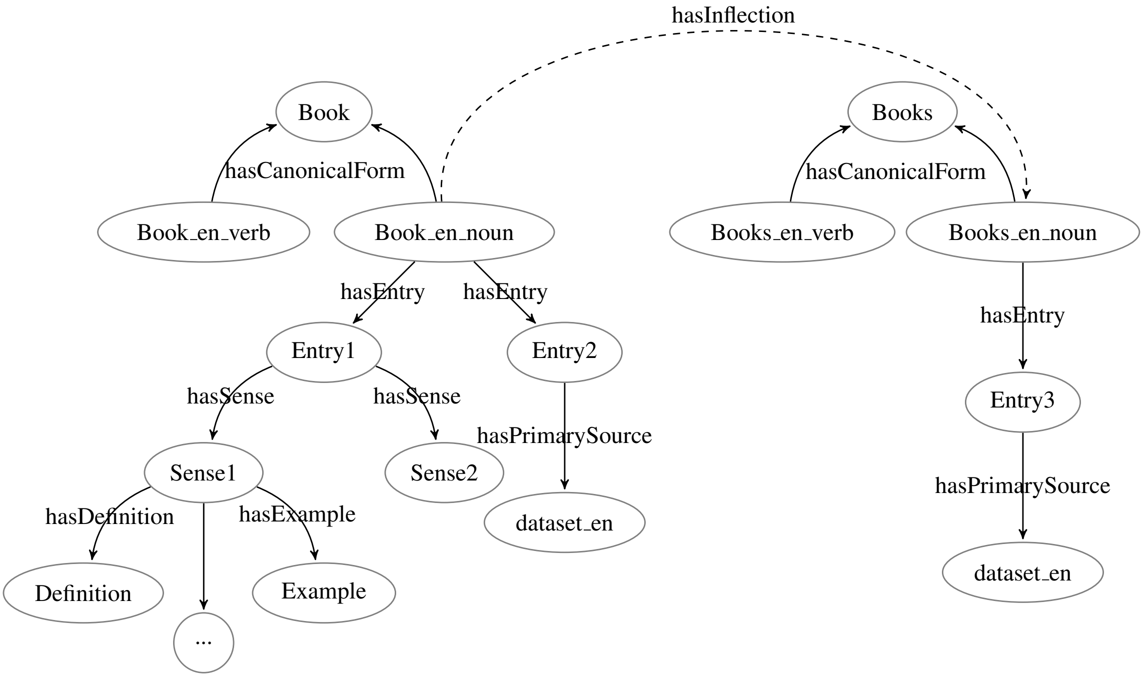

Figure 2: An example of headword *Book*

information showing the word's origin, such as: (i) language(s) of origin (the language(s) from which a word came) (ii) etymology (a string literal with a language tag) (iii) etymology type (such as compound) (iv) date of the current first known use of the word in English (v) cross-references to other headwords in OED.

5.4. Reasoning

Similar to LexInfo, we wanted to leverage the power of `OWL`, `SKOS`[17], and `PROV`[18]; not only to increase the reasoning power, but also to follow standards. Keeping in mind the trade-off between reasoning power and scalability, we were faced with the question of *which `OWL` profile to use?* Namely, how much reasoning could we build in considering the massive scope of 100 languages' worth of data? We also had other practical considerations, such as an eventual requirement to expose an editorial API allowing Oxford lexicographers to edit the content directly. An editorial API needs to be capable of interacting with a triplestore in real-time and able to check the consistency and satisfiability of the data before ingesting new resources. As a result, there must be adequate restrictions in the model, and the triplestore must contain the right `OWL` profile.

Another important consideration was the triplestore performance. Most of the available software doesn't have built-in support for reasoning, but those that do tend to fall into one of two categories: (i) reasoning takes place upon loading the data, or (ii) reasoning happens at query time. The former increased the ingestion time, but responded to queries faster. The latter had an opposite effect; while ingestion was quite quick, the query time would be slowed by the reasoning. We also discovered that some triplestores which reason upon ingestion falter in the case of retractions that include inferred statements. For those, query performance tended to diminish as the number of triples increased as well. All these considerations had some bearing on the decisions made for defining restrictions.

6. Example: The OGL Ontology in Action

Let us consider we have an English dictionary and some morphological data (i.e., inflected forms and grammatical categories), and we would like to represent that data in the OGL ontology model. We first identify a headword (e.g., book) and we treat it as a *form*. In essence, a form is a place-holder for a label comprising a headword with a language tag (e.g $\texttt{book}^{@en}$). A form links to one or more *lexical entries*; a lexical entry is a concept that can identify a headword form, its lexical category, and a language. It may contain pointers to inflections (e.g $\texttt{books}^{@en}$) and various *grammatical features*.

In Figure 3, the form `books` is linked to two different lexical entries in *English* (the language code of English

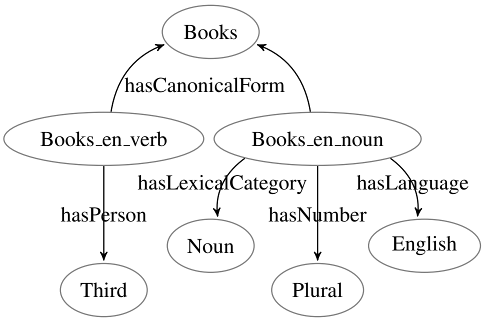

Figure 3: The lexical entry `books` and its relations to other concepts that describe its properties.

[17] https://www.w3.org/TR/skos-reference/

[18] https://www.w3.org/TR/prov-o/

is en): books_en_verb and books_en_noun. Both are identified as English words by means of the relation hasLanguage. The former has the lexical category (POS) verb and a grammatical feature described by means of the relation hasPerson that identifies the form as a Third Person of a verb. The latter has the lexical category noun.

A lexical entry may have multiple *entries*[19], and every entry may have multiple *senses*[20]. An important aspect of the entry class is its relation with the primary source (see Figure 2). Every entry belongs to one and only one source e.g., the English–isiZulu dictionary or the English thesaurus. This restriction enables us to track the origin of each sense of an entry and it also provides an effective filter for speeding up queries. As English is OUP's largest dataset and it is used as a pivot language for translations, querying response times tend to be greater, so the source filter was particularly useful for performance optimisation and content separation.

Sense is the richest lexical entity in this model. As shown in Figure 2, each sense could have the following main relations: definitions, example sentences, translations, subject domain, register, region, and pronunciation information as well as synonyms, antonyms, derivatives, ancestors and other cross-references to other entries or senses[21].

Another important aspect of our data on which we have experimented is the support for right to left, or generally non-Latin script languages. An issue arises in the inconsistency of the way the lexical entries are generated. Since the lexical entry is generated from headword label, language, and lexical category, the structure would benefit greatly from having translations of all the lexical categories to be able to make an IRI in the language. In other words, the aim is to replace بینک_ur_noun with a more readable and consistent بینک_اُر_اِسم as shown in Figure 4.

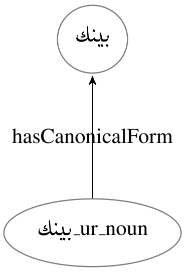

Figure 4: Current representation of the lexical entry for the headword *bank* in Urdu

7. Utilising the Model

7.1. Powering Dictionary Websites

As part of the OGL initiative, OUP has launched a set of dictionary websites that publish lexical resources for various digitally under-represented languages.[22] These websites are powered by a web service, especially designed for this purpose, that serves JSON objects holding the dictionary data following the schema required to display the websites surface. Nonetheless, the web service builds up the JSON output out of the RDF data in the triplestore modelled using the OGL ontology.

The web service consists of two main modules: a cache of JSON objects and a REST API. Due to the poor querying performance mentioned in previous sections, we decided to pre-compute most of the JSON content served by the API and store them in a cache. The pre-computed data is built by means of a background process that retrieves from the triplestore every RDF triple related to each headword (i.e. the list of lexical entries, senses, definitions, examples, translations, synonyms, inflections, etc.). Then, RDF data is serialised as a collection of JSON objects, and stored in the cache. The API, in turn, exposes a set of endpoints that allow search and retrieval of lexical data across the content of the cache. According to the invoked endpoint and the request parameters, the API picks up from the cache the JSON excerpts, and blends a response output from them.

For instance, let's consider that the API endpoint that provides the data for a dictionary headword is invoked to retrieve the headword Book in English. In such case, the collection of triples that represent the *Book* example in Figure 2 are serialised as two different JSON objects: Book and Books. Each of them consists of an array of two lexical entries: Book(s)_en_verb and Book(s)_en_noun; the latter having an array of *Entry* elements, and so on for the *sense* elements of Entry1. The API retrieves the JSON for Book and searches for its list of inflections. Having found the link to Books_en_noun, it retrieves the JSON for Books as well, and blends a new JSON output that contains both entries.

The OGL websites allow also the collection of *user generated content*, that is, new lexical content such as translations and examples gathered from native speakers. Using the OGL ontology model and the above web service has enabled us to collect this new data, combine it with data already present in the triplestore, and display it in realtime. At the same time, this set up endows a language community with some tools that may help them develop and enhance digital resources for their own language.

7.2. A Rich Lexical Web Service

Having developed the web service infrastructure mentioned in the Section 7.1., we are currently pursuing the opportunity to design and implement a set of general purpose APIs. The ultimate goal is to grant access to OUP's lexical content and empower developers to build their applications on top of it. These APIs will enable search and retrieve functionality across our datasets and provide our lexical data in a flexible and customisable manner. We would like to fulfil a broader range of requirements for other applications that are not necessarily constrained to publishing dictionary content.

The first version of this web service will offer access to the English, Spanish, English–Spanish, Spanish–English dictionaries and morphology datasets, and other datasets will follow. These datasets contain rich lexical annotations,

[19] An entry describes the information about a word

[20] A sense is one of the meanings of the word

[21] A cross-reference is a link between various dictionaries that identify concepts that are related in some way.

[22] At of March 2016, OUP has successfully launched 4 websites for Northern Sotho, isiZulu, Urdu, and Malay languages.

such as domain, region, register, and dialect; and thesaurus information, such as synonyms and antonyms.

In short, the public APIs will not only provide access to OUP's valuable language resources, but also facilitate discoverability of new lexical content by means of the power endowed by the ontology.

8. Evaluation

Most of the requirements that were gathered before developing the OGL ontology have been converted into SPARQL queries that were used as competency questions (CQ) (Ren et al., 2014) for assessing the scope of the OGL ontology. A pipeline containing these CQs has been designed and monitored regularly; this pipeline assesses the satisfiability and consistency of data before inserting it into the triplestore.

Performance. Although the OGL ontology has been satisfying all the modelling requirements, while designing the OGL ontology, we encountered some performance issues. Search queries were generally quite slow due, generally, to the large amount of data and the sparseness of the data related to each headword. Therefore, some extra constraints and concepts were added to the ontology to account for this poor performance. For example, as the number of ingested datasets was growing, a need was discovered for a filtering mechanism both for speeding up search, and for distinguishing between these datasets.

Generalisation. Another criterion for assessing and validating the flexibility of the OGL ontology is to be able to ingest various lexical datasets easily. So far, apart from inhouse datasets, we have only converted WordNet and evaluated the quality of the generated links. More investigation into other lexical datasets is necessary to assess the strength of the model to support various data.

In a nutshell, the OGL ontology shown great potential in handling various datasets in a variety of languages. However, new language resources will offer new challenges that may result in the extension of this model. Evaluating this ontology thoroughly requires larger and more varied datasets. Due to the lack of maturity of the triplestore technology, scaling up has proved to be a challenging experience; this has occasionally moved us towards rethinking some of the modelling decisions.

9. Next Steps

As our first tranche of experiments only included a few Indo-European, Austronesian, and Niger-Congo languages, further tests with datasets from other language families are required to prove the model's multilingual capability. OUP have a number of datasets from Sino-Tibetan and Afroasiatic languages, many of which we hope to incorporate in the near term. In the longer term, we expect that data coming from the OGL initiative will make possible a further variety of testing. The first phase of the OGL initiative is a set-up and proving phase, and is scheduled to run to the end of September 2016, by which point we plan to have launched ten languages). These are all languages that we have identified as having large numbers of speakers, but a relatively small number of high-quality digital resources.

Another big milestone for us is to design and implement an editorial system for directly manipulating the datasets in `RDF/XML` format. This system must be capable of assessing the validity of the new or updated input against the OGL ontology and the existing data in the triplestore.

OUP continues to have business requirements for more digitally established languages like English, Spanish, and Chinese as well, so we will test the ontology against these too. Ultimately, we intend to move all our lexical datasets into this model and manage them as linked data. Content pulled from this knowledge graph will power various applications, including Oxford Dictionaries Online[23] and the aforementioned public APIs.

10. Bibliographical References

Chavula, C. and Keet, C. M. (2014). Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In *Proceedings of the 11th OWL: Experiences and Directions Workshop*, pages 61–72. Riva del Garda, Italy.

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

Chiarcos, C. (2010). Grounding an Ontology of Linguistic Annotations in the Data Category Registry. In *Proceedings of the LREC 2010 Workshop on Language Resource and Language Technology Standards (LT&LTS)*, pages 37–40.

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.

Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J., Cimiano, P., and Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, volume 14, pages 401–408, Reykjavik, Iceland, May.

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

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening Ontologies with DOLCE. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 166–181. Springer.

Hirst, G. (2009). Ontology and the Lexicon. In *Handbook on Ontologies*, pages 269–292. Springer.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications (ESWC'11) - Volume Part I*, pages 245–259, Heraklion, Crete, Greece. Springer.

Niles, I. and Pease, A. (2003). Mapping WordNet to the SUMO Ontology. In *Proceedings of the IEEE International Knowledge Engineering Conference*, pages 23–26.

Petrov, S., Das, D., and McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of LREC*.

Ren, Y., Parvizi, A., Mellish, C., Pan, J. Z., Van Deemter, K., and Stevens, R. (2014). Towards Competency Question-Driven Ontology Authoring. In *The Semantic Web: Trends and Challenges*, pages 752–767. Springer.

[23] `http://www.oxforddictionaries.com/`