Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3281–3290
Marseille, 11–16 May 2020

The ACoLi Dictionary Graph

Christian Chiarcos, Christian Fäth, Maxim Ionov
Applied Computational Linguistics Lab (ACoLi)
Goethe University Frankfurt, Germany
{chiarcos, faeth, ionov}@em.uni-frankfurt.de

Abstract

In this paper, we report the release of the ACoLi Dictionary Graph, a large-scale collection of multilingual open source dictionaries available in two machine-readable formats: a graph representation in RDF, using the OntoLex-Lemon vocabulary, and a simple tabular data format to facilitate their use in NLP tasks, such as translation inference across dictionaries. We describe the mapping and harmonization of the underlying data structures into a unified representation, its serialization in RDF and TSV, and the release of a massive and coherent amount of lexical data under open licenses.

Keywords: lexical resources, OntoLex-Lemon, RDF

1. Motivation

Since the confusion of tongues, multilinguality has been both a challenge for and a treasure of mankind, with a number of challenges in the modern age, in particular, the technical challenge to overcome language barriers, and the socio-cultural challenge to preserve linguistic and cultural identities under the pressure of globalization. For both aspects, the publication of lexical data in machine-readable forms is of utmost importance, as this enables technical solutions to provide services across language barriers, in particular for speakers of languages that are underresourced with respect to NLP tools and language resources.

A challenge for the vast amount of lexical information available in digital form, is, however, that it comes in various forms. Several widely-used standards for the representation of lexical information in interoperable form do exist for a long time, e.g., in the context of the Text Encoding Initiative (Burnard and Sperberg-McQueen, 2012), the Lexical Markup Framework (Francopoulo et al., 2006) or tool-specific formats such as the StarDict format,[1] but they are limited with respect to their interoperability with each other,[2] and – more importantly – they are focusing on the standardization of dictionaries as independent, machine-readable entities, whereas many technical applications require an additional focus on the capability of integrating information across different lexical-conceptional resources. Both problems have been driving the more recent development of the OntoLex-Lemon vocabulary (McCrae et al., 2017) into the most important data model for the publication of lexical resources on the web of data, and the publication of lexical-conceptual data as Linguistic Linked (Open) Data. OntoLex-Lemon provides a coherent machine-readable representation that is designed to facilitate its integration across dictionaries and with sources of conceptual knowledge such as ontologies and knowledge bases.

With this paper, we announce the release of a massive collection of multilingual dictionaries in RDF, using the OntoLex-Lemon vocabulary. To facilitate the usage of this data in NLP tasks, we also provide an export into TIAD-TSV, a tab-separated format designed for a series of shared tasks on translation inference across dictionaries.[3]

We describe the mapping and harmonization of the underlying data structures of several substantial dictionary collections into a unified RDF representation that provides lexical data for more than 242 languages and 1,756 language pairs, alongside its export to the TIAD-TSV format.

2. Background

Formalisms to represent lexical resources are manifold and have been a topic of discussions within the language resource community for decades. Important vocabularies that gained considerable popularity include, for example, the Lexical Markup Framework (Francopoulo et al., 2006, LMF), or the dictionary specifications of the Text Encoding Initiative (TEI),[4] but also tool-specific formats such as used by the StarDict dictionary collection mentioned above.

These solutions are designed for the electronic editions and/or search in individual dictionaries. Lexical data does not, however, exist in isolation, and enormous synergies can be unleashed if information from different dictionaries is combined, e.g., for bootstrapping new bilingual dictionaries for languages X and Z by using another language Y and existing dictionaries for $X \mapsto Y$ and $Y \mapsto Z$ as a pivot.

Information integration beyond the level of individual dictionaries has thus become an important concern in the language resource community, and the most promising technology to achieve this goal is to adopt the linked (open) data paradigm for publishing lexical resources, i.e.,

[1] `http://stardict.sourceforge.net/`

[2] Aside from using different formats, the current lack of interoperability between existing representation formalisms is because of conceptual differences, e.g., the varying level of detail they provide in a machine-readable fashion.

[3] TIAD-2017, held at the First Conference on Language, Data and Knowledge (LDK-2017), Galway, Ireland, `https://tiad2017.wordpress.com`.
TIAD-2019, held at the Second Conference on Language, Data and Knowledge (LDK-2019), Leipzig, Germany, `https://tiad2019.unizar.es/`.
TIAD-2020, to be held in conjunction with GLOBALEX-2020 at LREC-2020.

[4] `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`

- to use URIs for unambiguously identifying lexical entries, their components and their relations in the web of data,
- to make lexical datasets accessible via http(s),
- to publish them in accordance with W3C-standards such as RDF and SPARQL, and
- to provide links between lexical data sets and with other LOD resources.

The primary community standard for publishing lexical resources as linked data is the OntoLex-Lemon vocabulary. Originally, *lemon* has been designed as a model for complementing ontologies with lexical information in the Monnet project (McCrae et al., 2011), but with its further development in the context of the W3C OntoLex Community Group, its scope was broadened and it developed towards the primary RDF vocabulary for lexical information. In 2016, the OntoLex vocabulary was published as a W3C Report (McCrae et al., 2017) and is now accessible via a W3C namespace.

The primary element in the model is the *lexical entry* (see Fig. 1), which represents a single word with a single part-of-speech and set of grammatical properties. This entry is composed of a number of forms and a number of senses which enumerate its meanings. The sense can be defined formally, with a reference to an ontology or informally, with a lexical concept, which defines a concept in a non-linguistic and hence cross-lingual manner.

Lexical entries (roughly corresponding to head words in the lexicon) group all forms of a word together into a single element, e.g. including inflected forms for a given part-of-speech. The entry for the verb *'(to) lead'* would include inflected forms such as 'lead', 'leads', 'led'. The word *'lead'* as the metal, being of a different part of speech and having different etymology, would constitute a separate lexical entry. Lexical entries are further grouped into three classes: (single) words, multiword expressions and affixes (such as 'anti-'). A lexical entry is composed of a set of lexical forms, each of which can be represented in different scripts. One of these lexical forms is specified to be the canonical form (i.e., 'lemma').

The semantics of a lexical entry can be given by indicating that it *denotes* an element in the ontology. The element in the ontology can be a class, property or individual that represents the denotation of the lexical entry in question. In many cases, this link to the ontology may need to be described in more detail. For this reason, the model provides the class *lexical sense*, representing the connection between a single lexical entry and its meaning in the ontology. Most lexicons require this to represent links between senses or pragmatic information, so that it is recommended to include a lexical sense for all links between lexical entries and ontology entities.

There has been much discussion of all aspects of the model, however the issue of semantics was of particular interest to the group and led to the introduction of conceptual models in OntoLex-Lemon: Whereas (word) senses are specific to a particular lexical entry (word), they can be grounded in lexical concepts that exist independently from a particular lexicalization. These lexical concepts resemble (external) ontological concepts in their function, they do, however, not require any formalization in terms of an ontology, but are created on the basis of independent (lexicographic or linguistic) considerations.

For example, the Princeton WordNet (Fellbaum, 2010) includes its own conceptual model in the form of hierarchically and relationally organized synsets, defines as representing the common semantic concept expressed by all the lexicalizations associated with it. Lexical concepts can thus be used to represent synonymity between lexical entries (resp., their senses), and if applied to lexical entries in multiple languages, this naturally extends to a formal model of translation equivalence and is thus of particular importance for multilingual lexical data.

This mechanism is complemented by additional vocabulary elements that allow to express more fine-grained translation relations in a more compact fashion: The OntoLex module for variation and translation (VarTrans) allows relations to be defined at three levels:

- *Lexical relations* relate the surface forms of a word, e.g. to represent etymology and derivation
- *Sense relations* relate the meanings of two words, e.g. to express that two senses are translations, synonyms or antonyms of each other
- *Conceptual relations* relate concepts regardless of their lexicalization. Examples of such conceptual relations are the hypernymy or meronymy relations.

As an example, consider the case of relating two lexical entries across languages. For this, the module considers three types of relations:

- **Interlingual Synonymy** is a relation between lexical concepts, claiming equivalence in meaning abstracting from the specific lexical meanings involved.
- **Translation** is a relation between senses claiming that a word with a given sense can be translated into another word with a given sense.
- **Translatable As:** At the lexical level, the `translatableAs` property relates two lexical entries that, in some context, might be translated into each other. Specifically, the property says that there is some meaning of the word in the source language that can be translated into some meaning of the word in the source language.

Implicit interlingual synonymity and the explicit translatableAs relation do not allow to provide explicit information about the relation (e.g., provenance, confidence, etc.), OntoLex-Lemon thus also provides reified Translation elements that take their respective source and target as arguments and can be further enriched with RDF statements.

With an increasing number of applications, the OntoLex-Lemon model has continued to expand in its use cases and has been adopted in a variety of online dictionaries and

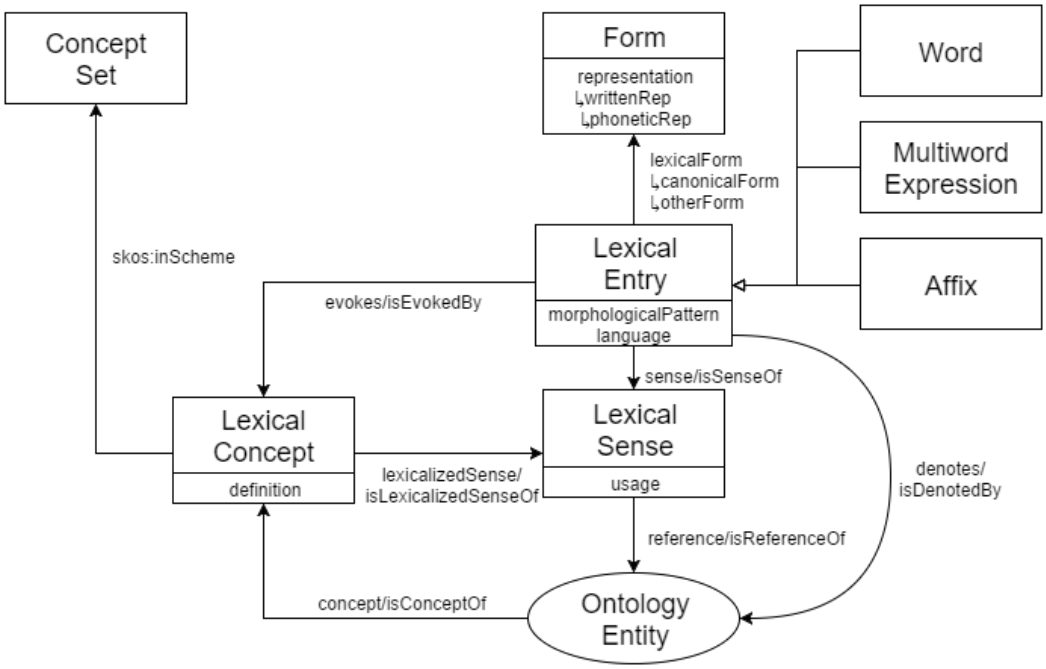

Figure 1: OntoLex-Lemon core model

this has provided a common interface to these dictionaries. Further, OntoLex-Lemon is employed in the context of the WordNet Collaborative Interlingual Index (Bond et al., 2016), where the model is being used to provide a single interlingual identifier for every concept in every language.
The ACoLi Dictionary Graph contributes to the growing amount of lexical-conceptual resources that provide OntoLex-Lemon compliant lexical data. It is created by aggregating over several sources with heterogeneous data models and differences in depth of representation. Our contribution is to provide a two-level normalization of this data:

(a) **RDF**: A coherent OntoLex-Lemon representation that preserves content and structure of the original data in RDF. This includes the preservation of characteristics and features that are not currently within the scope of the OntoLex-Lemon vocabulary. For these, properties and concepts are created that reside in the namespace of the respective source format. Also, we preserve the original data structure insofar as the modelling choices within OntoLex are being explored to follow the original data modelling as closely as possible. The data produced as the result of our OntoLex-Lemon conversion is thus *conceptually interoperable*, but not uniform.

(b) **TIAD-TSV**: From the rich and diverse OntoLex-Lemon representations, we generate a simplified data structure, where different structures are reduced to a simple tabular data format, i.e., a bilingual word list, together with the path of concepts connecting both lexical entries. While this approach leads to a substantial loss of conceptual information, it facilitates subsequent processing in technical tasks such as translation inference across dictionaries. The export to TIAD-TSV is implemented by means of a SPARQL SELECT query.

We describe the OntoLex conversion of three collections of dictionaries: Apertium (Tyers et al., 2010)[5], FreeDict[6], and PanLex (Kamholz et al., 2014)[7] as well as their export to TIAD TSV. In addition to the dictionaries mentioned above, we also provide a TIAD TSV export for DBnary (Sérasset, 2015), a lexical database that comes natively in OntoLex-Lemon, resp., RDF.
Note that these resources are of very different character which entail different modelling strategies in OntoLex-Lemon:

- Bi-dictionaries such as the Apertium dictionaries provide translation information together with grammatical information or pronunciation data about *both* source and target language.
- Bilingual dictionaries such as most PanLex and FreeDict dictionaries provide information about the source language, using target language expressions. Here, the target language expression is typically either a translation or a definition. However, these cannot be reliably distinguished in an automated way, so that target language information is best represented as a definition rather than as a translation.
- Bilingual word lists (most other FreeDict and PanLex dictionaries) provide translation information only, neither grammatical information nor definitions.
- DBnary is an aggregator for all three kinds of lexical information, and best described as a lexical database rather than a dictionary, in that it provides source language information (lexical entry, grammatical information, a definition, sometimes in another language,

[5] https://www.apertium.org/
[6] http://www.freedict.org/
[7] https://panlex.org/

like a bilingual dictionary) and cross-lingual translation links (which, for every language pair, constitute a bilingual dictionary in their entirety) for *multiple* languages.

In consequence, the resulting OntoLex representation establishes interoperability, but only the TIAD TSV export into a more rigid (and less expressive) tabular data structure guarantees uniform data structures.

3. Apertium ↦ OntoLex-Lemon

Apertium[8] is an open source system designed for the machine translation between closely related language varieties, mostly using symbolic methods. Apertium provides NLP components for many low-resource languages, as well as transfer rules and bi-dictionaries for their respective translation.

An RDF representation of this data for the translation between Romance languages and English has been developed at UP Madrid and released as Linguistic Linked Open Data.[9] This LLOD edition of Apertium is based on Monnet-Lemon, the predecessor of OntoLex-Lemon, and it has represented the training data for the TIAD shared tasks since 2017 (TIAD was evaluated against a blind test set provided by a commercial dictionary provider, KDictionaries). We refer to this earlier LLOD edition of Apertium as UPM Apertium.

Unfortunately, this was not a direct conversion, but using an intermediate LMF representation that was derived from the original Apertium data, and the mapping scripts for creating this LMF representation are no longer available. It was thus not possible to update the RDF data against the revised Apertium sources, nor to use the existing pipeline to provide LLOD data for language pairs not covered by the original conversion.

In preparation for our participation in the 2nd Shared Task in Translation Inference Across Dictionaries (TIAD-2019), we created such a converter, so that Apertium language pairs not contained in the provided training data (from the earlier Apertium conversion) could be used to perform an independent evaluation of and to optimize our system.

The converter was implemented in XSLT and designed to mirror the data modelling of UPM Apertium, down to the level of the URI schema so that Apertium URIs resolve against UPM Apertium URIs (for lexical entries that are contained in UPM Apertium).

We deviate from UPM Apertium in two aspects:

- We use OntoLex-Lemon instead of the outdated Monnet-Lemon.
- The converter is generic and designed to be lossless. We thus do not map Apertium tags against LexInfo,[10] because (a) we do not have a LexInfo mapping of Apertium tags for languages that are not contained in UPM Apertium, (b) the Apertium tags are not properly documented, and, thus, the mapping needs to be performed by a specialist in the languages under consideration, and (c) we cannot provide a mapping for future language pairs whose tags are still unknown to us. Things are complicated by the fact that Apertium tags are non-normalized strings, specific to translation pairs (not languages), and occasionally contain typos. Instead, we generate a URI in the Apertium namespace, using string value of every tag as its local name (say, `apertium:vblex` for the tag `vblex`), we complement it with label information from the header of Apertium XML files (e.g., `apertium:vblex rdfs:label "Lexical verb"`), and we assign it as morphosyntactic property to the lexical entry, e.g., `... lexinfo:morphosyntacticProperty apertium:vblex`

The data model of the resulting OntoLex-RDF is shown in Fig. 2.

We converted the full set of 55 Apertium bi-dictionaries, covering 46 languages in total, and provide it under GPL (like the original data) via `https://github.com/acoli-repo/acoli-dicts`. The release contains the build scripts, such that the data can be locally re-built if new Apertium dictionaries are being published or existing dictionaries are being updated. The build scripts provide an implicit versioning via the time-stamp provided with every RDF dump they create.

4. FreeDict ↦ OntoLex-Lemon

FreeDict[11] 'strives to be the most comprehensive source of truly free bilingual dictionaries', it provides over 140 dictionaries in about 45 languages under GPL, and thanks to its members, grows continuously. The primary application of FreeDict dictionaries is for human consultation by means of mobile apps, but it is also used for developing spell-checkers. To a large extent, their content is provided by laymen users. For both reasons, the level of quality and the coverage of FreeDict are more heterogeneous that that of the Apertium dictionaries. Furthermore, many FreeDict dictionaries provide explicit sense information, whereas word senses in Apertium are inferred from translation pairs (i.e., the assumption that every translation pertains to exactly one other word sense). Because of the complementary information they provide, both with respect to word senses and with respect to languages covered, the FreeDict dictionaries collection represents an important complementary resource, even though some of its dictionaries are of moderate quality or size.

FreeDict data are natively modeled in TEI. As no FreeDict-specific schema is provided and the TEI vocabulary is extremely rich (in total, TEI P5 contains 569 elements and 231 attributes) and extensible (TEI customizations allow to introduce novel attributes and to redefine existing ones), we provide a converter that focuses on `entry`, `form` and `sense`. These elements correspond directly to OntoLex-Lemon data structures, and only these are being converted (see Fig. 3). As FreeDict does not provide grammatical information about target language expressions, we do not

[8] `https://www.apertium.org/`

[9] `http://linguistic.linkeddata.es/resource/id/apertium`

[10] `https://www.lexinfo.net/`

[11] `https://freedict.org/`

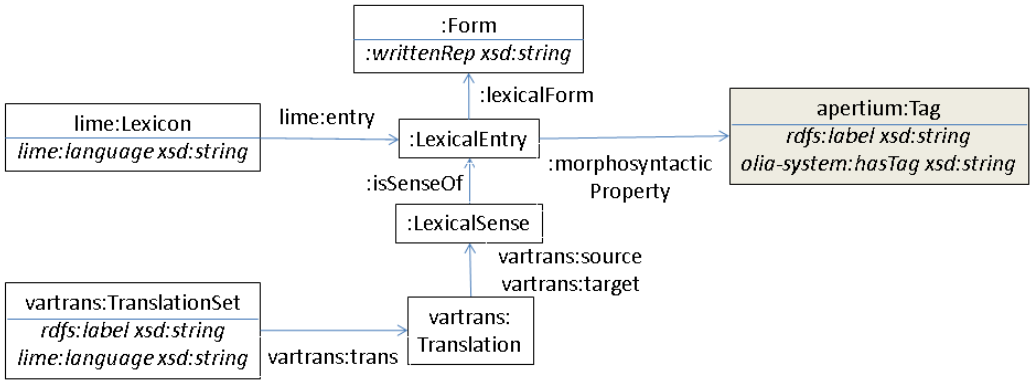

Figure 2: Apertium data model, non-OntoLex concepts shown in grey

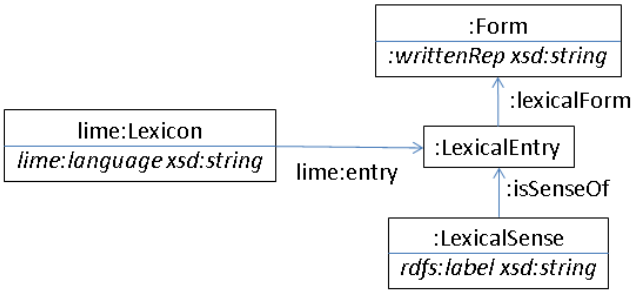

Figure 3: OntoLex-Lemon data model for FreeDict

model them as lexical entries in their own right, but in a more compact form as labels with an associated language tag, attached to the respective senses. Accordingly, we do not provide explicit translation links, see the snippet below:

```
:acquire-en
   a ontolex:LexicalEntry ;
   ontolex:lexicalForm :acquire-en-acquire-form .
:acquire-en-acquire-form a ontolex:Form;
   ontolex:writtenRep """acquire"""@en .
:acquire-en-sense-1 a ontolex:Sense;
   ontolex:isSenseOf :acquire-en;
   rdfs:label """alcanzar"""@es;
   rdfs:label """consequir"""@es;
   rdfs:label """obtener"""@es;
   rdfs:label """procurarse"""@es.
:acquire-en-sense-2 a ontolex:Sense;
   ontolex:isSenseOf :acquire-en;
   rdfs:label """comprar"""@es.
```

It should be noted that the RDF/Turtle output shown above is not only more compact than the native FreeDict format, but also more explicit in that source and target language are made explicit at the respective strings rather than being detached in the header.

5. PanLex ↦ OntoLex-Lemon

PanLex[12] is an effort to create a global lexical translation database, developed by transforming thousands of translation dictionaries into a common database structure. PanLex includes various types of lexical data, including digitized bilingual dictionaries, word lists, but also lexical-conceptual resources such as WordNets. According to the website, the database currently covers 2,500 dictionaries, 5,700 languages, 25,000,000 words and 1,300,000,000 translations. Our own counts over the PanLex database indicate a substantially larger number of dictionaries ('sources', currently 6,743), although many of these are small and their content is partially accessible only. PanLex data is available over a web interface and as CC0-licensed database dumps in CSV, resp. JSON formats. An RDF edition of PanLex has been previously described by Westphal et al. (2015), but the data is no longer available and its vocabulary pre-dates the OntoLex-Lemon model. The PanLex database structure is shown in Fig. 4:

- The source table contains document-level metadata and pointers to the original source of the data. Every entry in a source table corresponds to a single OntoLex dictionary (`lime:Lexicon`).
- The meaning table corresponds to lexical concepts in OntoLex, as, here, different translations of the same 'meaning' are aggregated.
- The definition table provides information about individual meanings, i.e., a description of the lexical concept. This information is optional.
- The meaning_prop table is used to provide a particular meaning with an externally provided identifier, e.g., a WordNet synset ID. The expr relation points to the entry of the expr table that represents the type of relation, the txt attribute is the actual identifier. This information is optional.
- The meaning_class table provides concept-level annotations, e.g., about the register of a concept, using elements from the controlled PanLex vocabulary. This information is optional.
- The expr table holds all lexical expressions as well as all elements from the controlled PanLex vocabularies. Expressions are linked with language codes and language metadata provided by the langvar table.

[12] `https://panlex.org/`

These are distinguished by their language tags: Lexical expressions carry an ISO 639-3 language code, controlled vocabulary is marked by the pseudo-code `art`. Expressions with language tags correspond to lexical forms in OntoLex.

- The denotation table holds information about individual lexical entries, the expr relation points to the corresponding (written representation of the) canonical form.
- The denotation_prop table represents free-text annotations for lexical entries: The relation is represented by a controlled term from the entry table, the value by the txt attribute. This information is optional, if provided, the expr argument (e.g., `translationQualityAssessment`) is mapped to a datatype property of the same name in the panlex namespace: `... panlex:translationQualityAssessment "+".`
- The denotation_class table provides annotations for lexical entries against a controlled vocabulary. This information is optional, if provided, the e1 argument (e.g., `PartOfSpeechProperty`) is mapped to an object property of the same name in the panlex namespace; if the e2 argument is an entry with the language code `art`, it is mapped analogously to an individual in the panlex namespace: `... panlex:PartOfSpeechProperty panlex:Noun`. If the e2 argument is an entry with another language code, we point to the URI of the corresponding expression, i.e., a lexical form.[13]

From the PanLex dump, we first create one XML document per row of the source table, subsequently populated with information directly or indirectly pertaining to this source element. In this way, all information about a particular lexical concept, its lexical entries and the respective forms are provided in a localized fashion via the lexical concept, and in the subsequent OntoLex-Lemon RDF/XML dump, this local structure is preserved.

Each of these XML files is then transformed with an XSLT script to an RDF/XML representation in OntoLex-Lemon, with the concept mapping as described above. It should be noted that PanLex differs from traditional dictionaries in that it provides lexical concepts, but no lexical senses. During the conversion, the controlled PanLex vocabulary (expressions with the "language tag" `art`) is preserved in a separate panlex namespace, but not normalized against LexInfo, so far. Translation relations are expressed implicitly in PanLex: Elements that pertain to the same meaning (lexical concept) are considered translation equivalent. This is a recommended design pattern for OntoLex-Lemon, too, but for the sake of explicitness, we add `vartrans:translatableAs` relations between all lexical entries (denotations) that are grouped under the same meaning.

As for the URI schema, we use `https://panlex.org/snapshot/` as base URI. These URIs resolve. Every PanLex dump carries a time stamp, and this is used to identify the respective release: `https://panlex.org/snapshot/panlex-20191001-csv`. As dumps are not directly addressable, these URIs do not resolve. Within the dump, the sources are identified by their PanLex key, i.e., an integer: `https://panlex.org/snapshot/panlex-20191001-csv/1782`. Within each source, lexical entries, lexical concepts and lexical forms are identified by their PanLex key (`https://panlex.org/snapshot/panlex-20191001-csv/1782#146067`), they are thus duplicated for every PanLex source; however, lexical entries are *not* duplicated for every lexical concept they are assigned to, so that the same lexical entry can relate to different lexical concepts (within the same source).

As for the PanLex vocabulary, this is not ontologically formalized during the export, we merely create URIs for expressions with language code `art`. The most important groups of object properties of lexical entries are grammatical features (panlex:PartOfSpeechProperty, panlex:GenderProperty, panlex:VoiceProperty, etc.), other forms than written representations (panlex:phoneticRep, panlex:phonemicRep), and morphosemantic relations (panlex:Causative_of, panlex:Inchoative_of). Because of the restrictions of the PanLex data model, the latter point to lexical forms rather than to the lexical entries for which these expressions are the corresponding canonical forms. It is not clear, however, whether the PanLex vocabulary is formalized or, in fact, controlled, or whether duplicates or redundancies exist, as many PanLex vocabulary elements appear to be somewhat cryptical (especially those using numerical schemes, e.g., the meaning_class `8.3`).

The resulting OntoLex-Lemon model for PanLex is illustrated in Fig. 5.

6. OntoLex-Lemon ↦ TIAD-TSV

So far, we described the conversion of various dictionary collections to RDF, using the OntoLex-Lemon vocabulary. The RDF format establishes structural interoperability between these data, and OntoLex-Lemon establishes conceptual interoperability in the sense that all data sets are modelled in accordance with the same reference data model. The OntoLex-Lemon vocabulary provides a consistent and lossless[14] view on these dataset, but this also means that it

[13] Note that this entails conceptual relations intended to connect lexical entries (or, for meaning_prop, lexical concepts) with other elements are restricted to links to either the elements to the panlex controlled vocabulary or lexical forms (as produced from the expr table). A property such as `panlex:Inchoative_of` should point to the corresponding lexical entry, instead, but this is not expressible in the PanLex data structure, and it is not corrected in our RDF representation of it.

[14] In terms of semantic coverage, OntoLex-Lemon allows modelling a resource in a lossless fashion because missing information can be preserved by incorporating it in a separate, resource-specific namespace. This was implemented for Apertium and PanLex, but not formally demonstrated (and for the case of FreeDict, also not attempted). Lossless conversion also means that the original data structures are adequately reflected in the application of different vocabulary elements from OntoLex-Lemon. This mo-

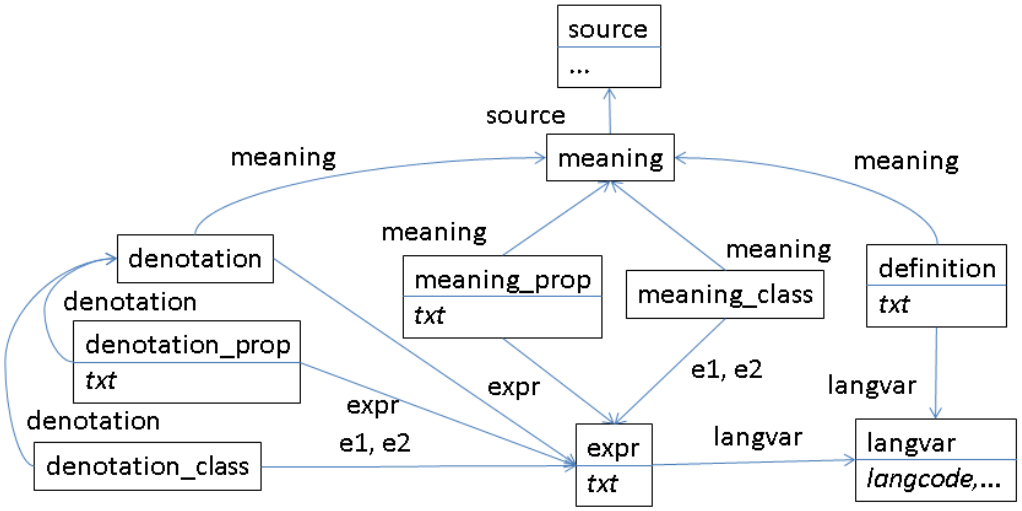

Figure 4: Original PanLex data model

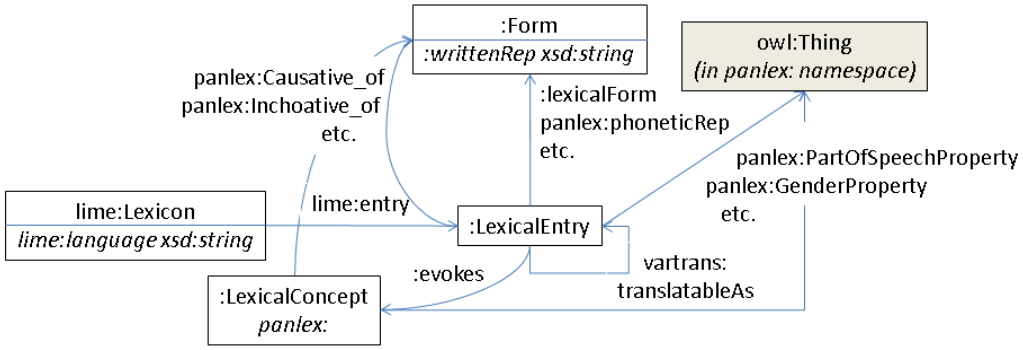

Figure 5: OntoLex-Lemon representation for PanLex data

preserves certain modelling choices that reflect the granularity, the intended usage and some arbitrary simplifications (e.g., the restriction of the range of morphosemantic relations to lexical forms) of the original data structure. The RDF representations of the ACoLi dictionaries are thus interoperable, but they are *not uniform*.

A truly uniform format for lexical data is necessarily reductionistic, and any reductions must be justified by their intended use. Here, we provide the resulting data in a format that is used in a series of shared tasks for Translation Inference Across Dictionaries (TIAD-2017, TIAD-2019, TIAD-2020) for which UPM Apertium served as training data. Aside from the Lemon representation of Apertium data, the training data was also provided in a simplified TSV format which columns reflect the path between a source word and its translation in the UPM Apertium data model:

1. source written representation (string, in *our* data also including the language tag in RDF/Turtle notation, e.g., `"acquire"@en`)
2. source word (URI of the lexical entry)
3. source sense (URI)
4. translation (URI)
5. target sense (URI)
6. target lexical entry (URI)
7. target written representation (string with a language tag, e.g., `"adquirer"@es`)
8. part of speech (LexInfo 2.0 URI, for the source word)

For every translation set (every language pair), one TSV file was provided.

The conversion of our OntoLex-Lemon data to this data structure is trivial, and performed with off-the-shelf RDF technology. A simple SPARQL SELECT statement is sufficient, e.g., for Apertium:

```
SELECT ?srep ?slex ?ssense ?trans
       ?tsense ?tlex ?turi ?trep ?pos
WHERE {
?slex ontolex:lexicalForm/ontolex:writtenRep ?srep.
?slex (^ontolex:isSenseOf) ?ssense.
?ssense (^vartrans:source) ?trans.
    ?trans vartrans:target ?tsense.
?tsense ontolex:isSenseOf ?tlex.
?tlex ontolex:lexicalForm/ontolex:writtenRep ?trep.
OPTIONAL {
        ?slex lexinfo:morphosyntacticProperty ?pos.
}
}
```

tivates the different OntoLex representations for the dictionaries under consideration here.

For other dictionaries, this query must be slightly adjusted, albeit not all URIs are available. Accordingly, the corresponding columns remain empty (we do not insert a placeholder symbol).
For FreeDict, for example, we take the sense labels to represent translations. Lines 5-8 from the listing above must thus be replaced by the following triple:

```
?ssense rdfs:label ?trep.
```

For PanLex, lines 4-7 must be replaced by `vartrans:translatableAs`:

```
  ?slex vartrans:translatableAs ?tlex
```

For all dictionaries, we create one TSV file per language pair.

7. Summary and Outlook

We described the creation of a novel, large-scale lexical-conceptual resource consisting of two components: A rich lexical graph with well-defined semantics in OntoLex-Lemon, and a uniform, but shallow representation of translation data derived from it in a simple TSV format. Although the TIAD-TSV export lacks metadata and information about the original modelling, it allows to recover the full information by means of the URIs of lexical entries, senses and translations it is generated from.
The generated RDF data is packaged such each source dictionary is represented in one corresponding RDF file.[15] The TIAD TSV export is re-packaged such that one file per language pair is generated. For resources that include information from several source dictionaries for the same language pair (here, PanLex), these can be disentangled by means of the URIs of lexical entries that pertain to the original source dictionary. All data compiled is published as open source, adopting the same license as the original data (GPL or CC licenses). In addition to the datasets described above, we also performed a conversion of the XDXF dictionary collection,[16] however, this conversion has been conducted in a shallow fashion and covers 88% of the XDXF dictionaries only, so that it is considered experimental. Data and build scripts are available from our GitHub repository.[17]
Figure 6 illustrates the languages covered by the ACoLi Dictionary Graph, visualized as nodes and represented by their BCP47 codes, and the bilingual (TIAD-TSV) dictionaries connecting them, visualized as edges. Table 1 provides the extraction statistics for the ACoLi Dictionary Graph.
Related research is to be seen in a large number of conversion projects aiming to provide OntoLex-Lemon compliant lexical data. Notable larger-scale efforts include DBnary (Sérasset, 2015) and UPM Apertium (Gracia et al., 2016),

	languages	language pairs	dictionaries	million lexical entries	license
Apertium	46	55	55	1.3m	GPL
FreeDict	45	145	145	1.4m	GPL
PanLex*	194	1,651	2,411	57.2m	CC0
XDXF	51	107	147	2.7m	GPL
DBnary*	119	275	20	0.5m	CC-BY-SA 3.0

* For DBnary and PanLex languages and language pairs, only TIAD TSV files with more than 10,000 translation pairs have been considered. For dictionaries and lexical entries, numbers for the full data are reported.

Table 1: Conversion statistics for the ACoLi Dictionary Graph

as well as the earlier PanLex conversion described by Westphal et al. (2015). Our collection builds on these efforts but exceeds them in terms of scale, and complements the resulting RDF graph with an additional, shallower representation as tab-separated values – designed to facilitate use and reuse of RDF lexical data in language technology. Furthermore, these resources are complementary in the information they provide: FreeDict provides explicit senses that are lacking in Apertium and PanLex (Apertium word senses are induced from translation pairs); because of its designated use case in MT, the grammatical information provided by Apertium is particularly rich and better curated than any of the other resources; DBnary has a good coverage, but its data is crowd-sourced and less well-curated than the data from Apertium or PanLex print dictionaries; PanLex has excellent coverage wrt. languages, but it is partially based on OCRed texts and applies a number internal simplifications (e.g., lowercasing) that lead to corruptions in its data. Providing this data in a consistent and flexible representation, as a graph of lexical data, thus allows us to explore synergies between these data sets.

At the moment, we provide a generic RDF conversion for each of the dictionary collections, we do not harmonize external vocabularies beyond the application of the OntoLex-Lemon vocabulary. In particular, linguistic categories are not yet normalized against the LexInfo ontology. For Apertium, we currently work in this direction in the context of the H2020 Research and Innovation Action Pret-a-LLOD, this is, however, a major undertaking as the tag inventories comprise hundreds of terms. Likewise, we made no attempts so far to establish links between different datasets pertaining to the same language variety. For many cases, this mapping is trivial, however, the dictionaries we provide do not all provide part of speech information, nor do they agree on a particular classification scheme for these. Only if these have been synchronized, we can exploit the formal identity of written representations and common morphosyntactic characteristics to infer `owl:sameAs` relations between lexical entries from different dictionaries. The TIAD export format is designed to apply such techniques, so that we expect that results of the automated linking to be integrated with a future update to the dictionaries.

[15] For the Apertium bi-dictionaries, we provide one archive per source dictionary, containing source language and target language dictionary, together with the translation set, each in a separate file.

[16] Available from `https://sourceforge.net/projects/xdxf/` under GPL.

[17] `https://github.com/acoli-repo/acoli-dicts`

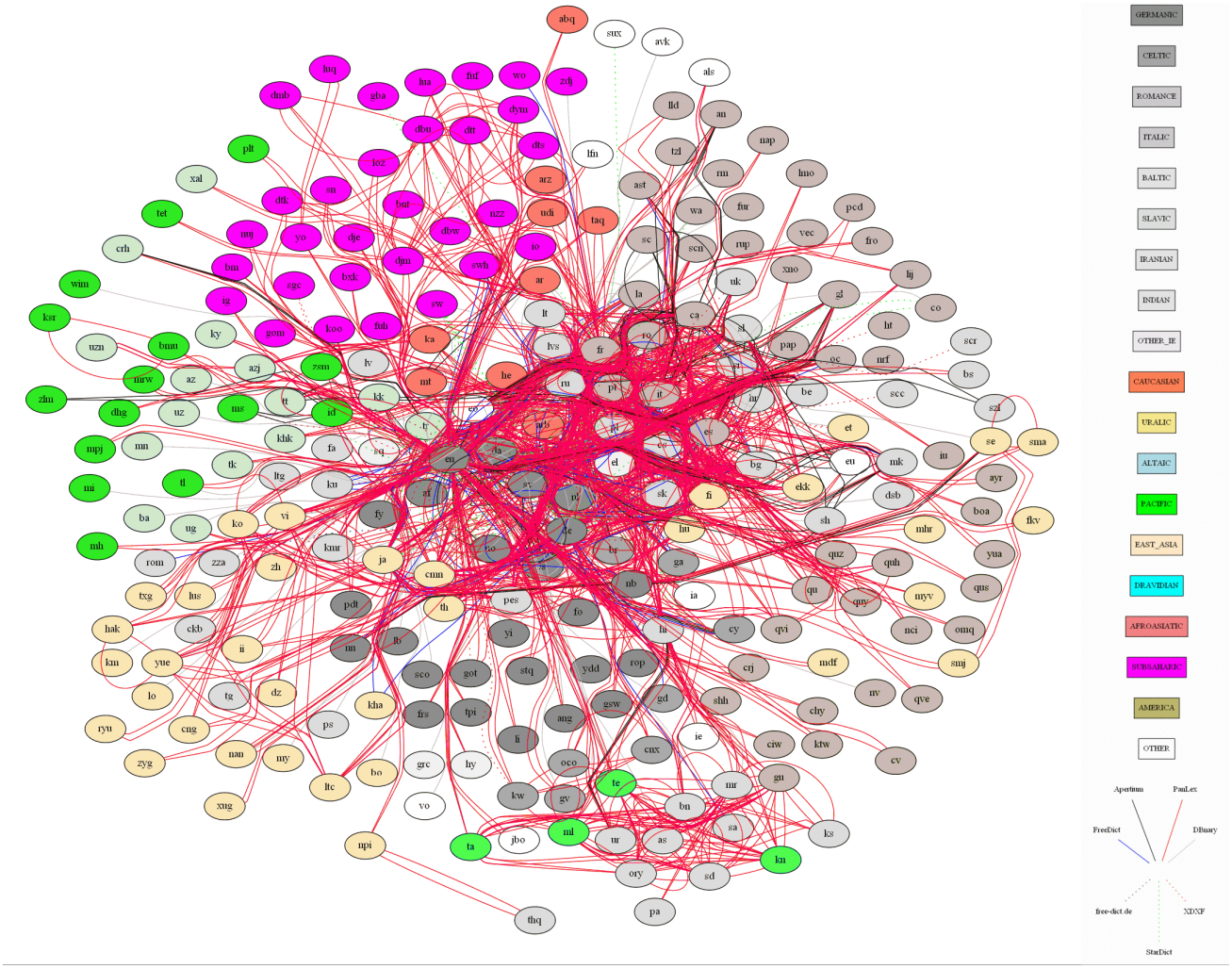

Figure 6: ACoLi Dictionary Graph, TIAD TSV files

Acknowledgments

The research described in this paper has been partially conducted in the context of the BMBF Early Career Research Group 'Linked Open Dictionaries (LiODi)' (initial conversion of Apertium, conversion of FreeDict), and partially in the context of the Horizon 2020 Research and Innovation Action 'Pret-a-LLOD', Grant Agreement number 825182 (revised conversion of Apertium, conversion of PanLex).

8. Bibliographical References

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proc. of the Global WordNet Conference 2016*.

Burnard, L. and Sperberg-McQueen, C. (2012). TEI Lite: Encoding for interchange: an introduction to the TEI. Technical report, Text Encoding Initiative, August. Final revised edition for TEI P5.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., et al. (2006). Lexical markup framework (LMF). In *Proc. of LREC*, volume 6.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2016). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal*, Sep.

Kamholz, D., Pool, J., and Colowick, S. (2014). Panlex: Building a resource for panlingual lexical translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In *Proc. of the 8th Extended Semantic Web Conference*, pages 245–249.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex)*.

Sérasset, G. (2015). DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361.

Tyers, F., Sánchez-Martínez, F., Ortiz-Rojas, S., and Forcada, M. (2010). Free/open-source resources in the apertium platform for machine translation research and

development. *The Prague Bulletin of Mathematical Linguistics*, 93:67–76.

Westphal, P., Stadler, C., and Pool, J. (2015). Countering language attrition with panlex and the web of data. *Semantic Web*, 6(4):347–353.