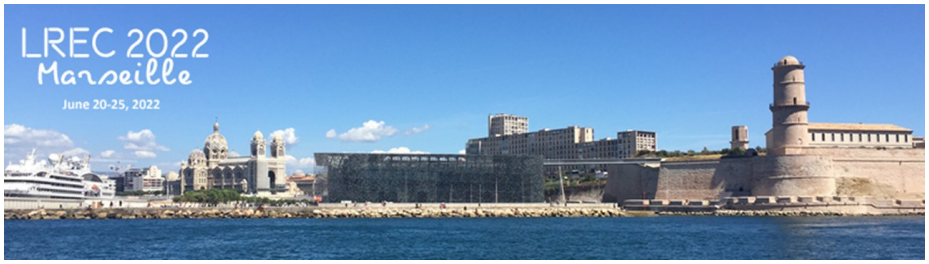


LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

**Proceedings of the 8th Workshop on Linked Data in Linguistics**
**(LDL-2022)**

# PROCEEDINGS

Editors:
Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, Maxim Ionov

# Proceedings of the LREC 2022 workshop on on Linked Data in Linguistics: Revisiting a Decade of Linguistic Linked Open Data (LDL 2022)

Edited by:
Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, Maxim Ionov

The 10th anniversary edition of the LDL can count on the support of the COST Action CA18209 "NexusLinguarum: European Network for Web-centered Linguistic Data Science", as well as two Horizon 2020 projects. Firstly, the Prêt-à-LLOD project (grant agreement no. 825182), which is making linguistic linked open data ready-to-use, and, secondly, the ELEXIS project on building a lexicographic infrastructure (grant agreement No 731015).

**ISBN: 979-10-95546-93-1**
**EAN: 9791095546931**

**For more information:**
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

# Preface

This volume documents the Proceedings of the 8th Workshop on Workshop on Linked Data in Linguistics, held on Friday teh 24th of June as part of the LREC 2022 conference (International Conference on Language Resources and Evaluation).

Since its inception, the workshop series on Linked Data in Linguistics (LDL) established itself as the main venue for discussing how Linked Open Data (LOD) and semantic web technologies can be used for processing, analysing, publishing, and managing linguistic data. This includes the fields of natural language processing (NLP), language resources (LRs), lexicography and digital humanities (DH), and has been leading to the development of linguistic data science as a new area of study. The LDL workshop series has contributed greatly to the development of the Linguistic Linked Open Data (LLOD) cloud and the development of best practices for publishing and accessing language resources and providing language technology services on the web. Most notably, this includes community standards such as the NLP Interchange Format (NIF), the OntoLex-Lemon model of the W3C Community Group Ontology-Lexica, and numerous domain-specific adaptations and extensions that these models have had an influence on.

In addition, there are an increasing number of national, European, and international research projects that build on LLOD technology. These will contribute to its further development and will help ensure the success of this workshop and a high attendance rate. The 10th anniversary edition of the LDL can count on the support of the COST action "NexusLinguarum: European Network for Web-centered Linguistic Data Science", as well as two Horizon 2020 projects. Firstly, the Prêt-à-LLOD project, which is making linguistic linked open data ready-to-use, and, secondly, the ELEXIS project on building a lexicographic infrastructure.

**Organizers**

Thierry Declerck – DFKI GmbH, Saarland Informatics Campus
John P. McCrae – National University of Ireland Galway
Elena Montiel – Universidad Politécnica de Madrid
Christian Chiarcos – Goethe University Frankfurt
Maxim Ionov – Goethe University Frankfurt

**Program Committee:**

Sina Ahmadi (NUI Galway, Ireland)
Paul Buitelaar (Insight, Ireland)
Sara Carvalho (University of Aveiro, Portugal)
Nicoletta Calzolari (ILC-CNR, Italy)
Milan Dojchinovski (Czech Technical University in Prague, Czech Republic)
Agata Filipowska (Poznan University of Economics, Poland)
Francesca Frontini (ILC-CNR, Italy)
Jeff Good (University at Buffalo, USA)
Yoshihiko Hayashi (Waseda University, Tokyo, Japan)
Eero Hyvönen (Aalto University, Finland)
Fahad Khan (ILC-CNR, Italy)
Chaya Liebeskind (Jerusalem College of Technology, Israel)
Gerard de Melo (HPI/Universität Potsdam, Germany)
Steve Moran (University of Neuchâtel, Switzerland)
Verginica Mititelu (Research Institute for Artificial Intelligence of the Romanian Academy, Romania)
Roberto Navigli ("La Sapienza" Università di Roma, Italy)
Sebastian Nordhoff (Language Science Press, Berlin, Germany)
Petya Osenova (IICT-BAS, Bulgaria)
Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia)
Antonio Pareja-Lora (Universidad de Alcalá de Henares)
Felix Sasaki (SAP, Germany)
Andrea Schalley (Karlstad University, Sweden)
Gilles Sérasset (University Grenoble Alpes, France)
Milena Slavcheva (IICT-BAS, Bulgaria)
Ranka Stankovic (University of Belgrade, Serbia)
Armando Stellato (University of Rome, Tor Vergata, Italy)
Andrius Utka (Vytautas Magnus University, Lithuania)

# Table of Contents

# Conference Program

**Friday June 24, 2022**

*The Annohub Web Portal*
Frank Abromeit

*From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)*
Milica Ikonić Nešić, Ranka Stanković, Christof Schöch and Mihailo Skoric

*IMTVault: Extracting and Enriching Low-resource Language Interlinear Glossed Text from Grammatical Descriptions and Typological Survey Articles*
Sebastian Nordhoff and Thomas Krämer

*Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin*
Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti and Paolo Ruffolo

*Use Case: Romanian Language Resources in the LOD Paradigm*
Verginica Barbu Mititelu, Elena Irimia, Vasile Pais, Andrei-Marius Avram and Maria Mitrofan

*Fuzzy Lemon: Making Lexical Semantic Relations More Juicy*
Fernando Bobillo, Julia Bosque-Gil, Jorge Gracia and Marta Lanau-Coronas

*A Cheap and Dirty Cross-Lingual Linking Service in the Cloud*
Christian Chiarcos and Gilles Sérasset

*Spicy Salmon: Converting between 50+ Annotation Formats with Fintan, Pepper, Salt and Powla*
Christian Fäth and Christian Chiarcos

*A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data*
Fahad Khan, Christian Chiarcos, Thierry Declerck, Maria Pia Di Buono, Milan Dojchinovski, Jorge Gracia, Giedre Valunaite Oleskeviciene and Daniela Gifu

*Computational Morphology with OntoLex-Morph*
Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti and Matteo Pellegrini

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 1–6
Marseille, 24 June 2022


# The Annohub Web Portal

**Frank Abromeit**
Applied Computational Linguistics Lab - Goethe University of Frankfurt, Germany
abromeit@em.uni-frankfurt.de

**Abstract**

We introduce the Annohub web portal, specialized on metadata for annotated language resources like corpora, lexica and linguistic terminologies. The portal will provide easy access to our previously released Annohub Linked Data set, by allowing users to explore the annotation metadata in the web browser. In addition, we added features that will allow users to contribute to Annohub by means of uploading language data, in RDF, CoNLL or XML formats, for annotation scheme and language analysis. The generated metadata is finally available for personal use, or for release in Annohub.

**Keywords:** Linguistic Metadata, LOD, LLOD, OLiA

## 1. Introduction

Linguistic metadata has been a research topic for a long time, starting with XML based data formats like TEI[1] and OLAC (Bird and Simons, 2001) and many portals that provide linguistic resource metadata have emerged ever since. For example OLAC[2], the CLARIN infrastructure (Hinrichs and Krauwer, 2014), Meta-Share[3] (Piperidis, 2012), and more recently LingHub[4] (McCrae and Cimiano, 2015). Following the paradigm to distribute data collections as Linked Open Data (LOD) (Bizer et al., 2009)[5], this methodology has been applied to linguistic data[6] (Cimiano et al., 2020), but also to the provenance metadata for linguistic resources. So, for example, LingHub provides linguistic metadata in RDF[7] formats. The RDF framework offers, in contrast to XML based metadata formats, different perspectives, like open data, standardized metadata vocabularies like Meta-Share (McCrae et al., 2015) and DCAT[8], and the ability to process resource metadata along with the actual language data, by means of SPARQL[9] queries. This finally allows tighter integration of NLP-processes that handle corpus, lexicon or terminology language data.

In recent work we have created the Annohub Linked Data set (Abromeit et al., 2020)[10], a metadata collection of annotated language resources, like corpora and lexica. Here, we introduce the Annohub portal (hosted by the Lin|gu|is|tik portal (Chiarcos et al., 2016)) that will provide users with easy access to Annohub's metadata in the web-browser. In addition, NLP-services will enable registered users of the portal to upload annotated language resources in order to perform an analysis on used languages and annotation schemes. The analysis results can then be used to create new entries in the Annohub catalogue. Furthermore, proper editing and commentary functions will help to improve the quality of the gained metadata and to keep the resources listed in Annohub up to date.

## 2. Annohub web portal

One of the goals of the Annohub portal is to bring metadata of prominent lexical resources and corpus data to a broader audience, but also to advertise new language resources that can not be found on other platforms like LingHub, CLARIN centers[11], Meta-Share or elsewhere. Annohub's metadata combines common resource metadata together with detailed language and annotation information. In addition, the provenance metadata is augmented, by linking annotations that have been used in a language resource, to OLiA[12] ontology classes, as well as to the original annotation scheme providers. All metadata is finally provided in a Linked Data representation, that is well suited for its use with other Linked Data applications, such as querying across multiple LLOD datasets, by means of federated SPARQL queries. Possible use cases of the new portal include:

- Search for publicly available annotated language resources like corpora, lexica or terminologies
- Contribute to Annohub by uploading language resources
- Learn about annotation schemes used in language resources

The portal currently encompasses metadata for over 1000 annotated language resources like corpora, lexica and ontologies. These resources are harvested automatically from different locations like LingHub's RDF data

[1] https://tei-c.org/
[2] http://www.language-archives.org/
[3] https://www.meta-share.org/
[4] https://linghub.org
[5] https://lod-cloud.net
[6] http://www.linguistic-lod.org/
[7] https://www.w3.org/RDF/
[8] https://www.w3.org/TR/vocab-dcat/
[9] https://www.w3.org/TR/sparql11-query/
[10] https://annohub.linguistik.de/en/
[11] https://www.clarin.eu/
[12] https://github.com/acoli-repo/olia

dump[13], CLARIN centers[14] (by means of the OAI protocol[15]), but also originate from several selected websites like the OPUS portal[16], the Språkbanken [17] website and a collection of corpora and lexica that have been compiled at the ACoLi Lab, Goethe University of Frankfurt (Chiarcos et al., 2020)[18]. The provenance metadata of each dataset is copied from the original metadata provider (RDF/XML/HTML) or has been added manually. Language and annotation information is extracted from the language data by an automated NLP-pipeline (see (Abromeit et al., 2020)). After the analysis, all language and annotation metadata, as well as the provenance metadata can be edited in the web-browser (see (Abromeit and Chiarcos, 2019) [19]) in order to complement missing information or to correct errors from the automatic analysis steps. The portal is built in Java with Apache Jena[20] and the Apache Tinkerpop framework[21] with two Neo4j[22] databases as backend. One of which is used as a backbone for the web-application, whereas the other database is used to map OLiA ontology classes to annotation tags and URLs found in the language data.

## 3. Ontologies of Linguistic Annotations

The *Ontologies of Linguistic Annotations* (OLiA)[23] provide a formalized, machine-readable view on linguistic annotations for more than 75 different language varieties. They cover morphology, morphosyntax, phrase structure syntax, dependency syntax, aspects of semantics, and recent extensions to discourse, information structure and anaphora, all of these are linked with an overarching reference terminology module. OLiA includes several multi-lingual or cross-linguistically applicable annotation models such as the Universal Dependencies (77 languages), EAGLES (11 European languages) and Multext-East (16 Eastern European and Near Eastern languages). The OLiA core ontology files[24] build the reference terminology module and include over 900 ontology classes. They contain the definitions of fundamental concepts that are commonly used to annotate syntax, morphology and morphosyntax. They are therefore well suited as the basis for an application designed to search linguistic annotations and features in corpora or lexica, independently of used annotation models and languages.

## 4. Looking up language resources

While browsing and searching Linked Data sets like Annohub with the SPARQL query language is reserved to computer scientists only, the new web-interface will allow non-experts to examine Annohub's metadata in detail. Search parameters include:

- Language (as ISO639-3 code)
- Tagset (e.g. PENN)
- Resource type (corpus, lexicon, ontology)
- Annotation (e.g. part-of-speech tag)
- OLiA class
  (e.g. http://purl.org/olia/olia.owl#Verb)
- Resource URL
- Provenance metadata (e.g. author, title)
- Comments made by users

### 4.1. Lookup by language / tagset / type / name / provenance / comment

In order to provide exact results the language information in a query has to be provided as ISO639-3 code. The ISO639-3[25] code table encompasses over 7000 languages. Code guessing from a natural language input may be included in upcoming releases. In order to distinguish unilingual, bilingual and multilingual resources the search form has an option to run a query with AND,OR (exclusive AND/OR) operators.
Currently, Annohub supports 41 annotation schemes[26]. These cover annotations commonly used for annotating corpora, as well as RDF vocabularies like OntoLex-Lemon[27], which is actually not an annotation scheme, but rather a RDF vocabulary that is widely used to model lexical data. A model query can include one or multiple annotation schemes with the above-mentioned logical operators. Available resource types include lexica, corpora and ontologies. Another category are wordnets which will be supported in future releases.[28].

[13] https://linghub.org/linghub.nt.gz
[14] https://centres.clarin.eu/restxml/
[15] https://www.openarchives.org/OAI/openarchivesprotocol.html
[16] https://opus.nlpl.eu/
[17] https://spraakbanken.gu.se/
[18] https://github.com/acoli-repo/
[19] https://annohub.linguistik.de/beta/FID-Documentation.pdf
[20] https://jena.apache.org/
[21] https://tinkerpop.apache.org/
[22] https://neo4j.com/
[23] https://acoli-repo.github.io/olia
[24] http://purl.org/olia/olia.owl, http://purl.org/olia/olia-top.owl, http://purl.org/olia/system.owl
[25] https://iso639-3.sil.org/code_tables/download_tables
[26] Alpino, Ancorra, Brown, Connexor, Dzongkha, Eagles, Emille, Genia, Iiit, Ilposts, Lassyshort, Lexinfo, Mamba, Mamba-Syntax, Morphisto, MULTEXT-East, Ontolex, Penn, Penn-Syntax, Ppcme2, Proiel, Qtag,Russ, Russleeds, Sfb632, Stanford, Stts, Suc, Susa, Tcodex, Tibet, Tiger, Tiger-Syntax, Treetagger, Tueba, Urdu, Ycoe, Ubycat, UBY-POS, UD-POS, UD-Dependencies (Universal Dependencies), located at https://github.com/acoli-repo/olia
[27] https://www.w3.org/2019/09/lexicog/
[28] The resource classification process is described in (Abromeit et al., 2020)

In addition, querying resources by URL, provenance data (e.g. author, title, etc.) or comments made by users, is implemented as a full-text query on all provenance attributes / posted comments.

### 4.2. Lookup by annotation / OLiA class

Words in corpus or lexicon data have tags (strings) or classes (URLs) attached to, that are used to classify them. For example, the tag *Pp3fpi* is used to mark instrumental-case in the Multext-East annotation scheme. Examples[29] for the usage of OLiA annotation classes (URLs) can be found in corpus data that is annotated with the NLP Interchange Format (NIF)[30]. The OLiA ontologies cover over 30.000 annotation tags. By means of the search forms (see Fig.1, 2) resources can be located that explicitly contain an occurrence of a tag or an OLiA annotation class.

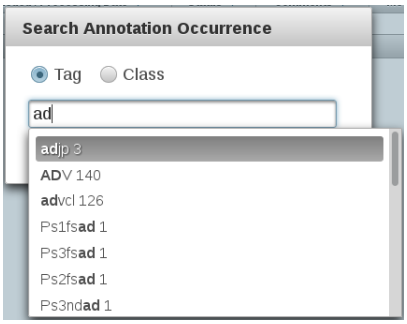


Figure 1: Annotation tag search form

By selecting a tag / OLiA class the number of resources is shown that contain a reference to it.

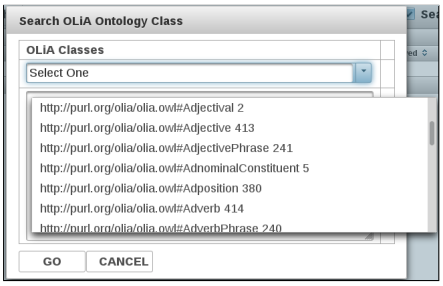


Figure 2: OLiA class search form

## 5. Contributing to Annohub

In order to benefit from the input of the language community, the portal offers an upload-service that allows registered users to analyze language data. Supported data types include Linked Data formats like rdf, nt, n3, etc., CoNLL[31] style data and to some degree XML encoded data[32], also as part of zip, tar and gzip archives. Limits on the size and amount of data files a user can upload are granted individually. Uploading works by providing the download URL of a language resource.[33] Before an upload is started it is checked if a resource is already contained in the catalogue or has been previously unsuccessfully processed. For this purpose the download URL, HTTP header information (e.g. *etag* information [34]) as well as MD5 and SHA256 hashes of already processed resources are kept in a database. Nevertheless, further manual duplicate checking has to be applied since a resource can have different versions and is possibly hosted at multiple locations. Finally, new resources will be queued for processing and progress information as well as the analysis results can be examined in the web-browser. In addition, registered users can comment on individual datasets listed on Annohub. Based on this feedback corrections can be made and it is decided by the reviewers at the linguistic portal[35] which user uploaded datasets will be included in the official Annohub RDF release[36]. General requirements for language resources to be included in the Annohub release are:

- A resource is publicly available via an URL as a downloadable file
- A resource is in RDF, CoNLL or XML format
- A resource includes word annotations from the syntactical or lexical domain. Otherwise only language information will be extracted
- Provenance metadata like a description for a dataset and author, licence, etc. information is provided

Because Annohub does not host the uploaded language resources, but merely the extracted metadata from it, anybody can upload data, despite of any license restrictions. Since the ability to upload content to a website poses a severe risk to fraud, by creating manipulated data packages with the intention to hack services, possible threats have to be carefully investigated.

## 6. Performance analysis

A qualitative analysis of the automatic tagset and language detection for CoNLL data is presented in (Abromeit and Chiarcos, 2019). Here, we focus on the analysis speed for three different data formats used for

[29] https://lider-project.eu/sites/default/files/referencecards/NIF-Corpus-reference-card.pdf

[30] https://persistence.uni-leipzig.org/nlp2rdf/

[31] https://www.signll.org/conll

[32] For a description of the supported XML data formats please see(Abromeit et al., 2020), chapter 6.1

[33] The processing of URL lists is supported as well

[34] https://docs.w3cub.com/http/headers/etag.html

[35] https://linguistik.de

[36] https://annohub.linguistik.de

language data, namely RDF, XML and CoNLL. Runtime is crucial, especially when large numbers of files with unknown content have to be processed in an unsupervised fashion, which is the case for any uploaded content to Annohub, but also applies when processing harvested file lists from CLARIN centers or other language resource metadata providers. A problem that occurs with language data encoded in RDF and XML formats is, that these formats are also widely used for non-linguistic purposes. Therefore, sampling techniques have been implemented in order to rule out unusable data quickly, but also to minimize computation times when processing large files or large collections of files (e.g. in tar archives) by testing a small fraction of a file first and by limiting the total number of data files to be processed.

### 6.1. Processing RDF files

RDF data is processed in a streamline fashion by utilizing the Apache Jena streaming interface[37]. This has the advantage that RDF files do not have to be loaded into a dedicated RDF triple store, which can take long for large datasets. In a first step the RDF data is validated[38] for correct URI specification of the included triples (checking forbidden characters), because this may lead to processing errors later. In case a non-conform URI is found, the RDF data is then converted to an RDF-XML representation by means of the rapper[39] RDF-utility. This has proven to fix any issues reliably. After these prepossessing steps the actual parsing of the RDF data starts. More details about the parsing process can be found in (Abromeit et al., 2020).

### 6.2. Processing CoNLL and XML files

The CoNLL file format is a tabular data format (TSV), where each line contains a word together with lemma, annotation and dependency information (see `https://universaldependencies.org/guidelines.html`). Parsing a CoNLL file works by identifying first the type of data included in the individual columns, because the CoNLL data format is not standardized to a certain order or number of columns (e.g. extra columns can be used to include language specific annotations). Subsequently, the language used in the word and lemma column as well as the annotation schemes used in 'annotation' columns are determined. XML files are treated in the same way as CoNLL files after they have been converted from the XML format to a CoNLL representation.

### 6.3. Evaluation

Table 1 shows the computation times for some well known datasets. Tests were performed on a Xeon server CPU (quad-Core) with 20GB RAM. The processing time in the last column of the table is composed of three parts (a) download time (b) validation time (only RDF) and (c) the time for NLP analysis. For better comparison, (a) and (b) are omitted for the RDF files. Download times for the CoNLL and XML examples could be neglected.

- All triples in a RDF file are examined. Since the runtime scales linear with the number of triples this alone can explain the different runtimes. A second performance factor is the number of database writes which scales linear with the amount of identified tags[40]. Since lexica generally do not contain word annotations, but rather word definitions in different languages (Wiktionary: eng, Wordnet: eng, DBnary (de): 515 languages), this factor is rather small[41]. A substantial part of the computation time is spent for validating a dataset before parsing (Wordnet: 20s, Wiktionary: 110s, DBnary: 120s). However, disabling the validation step could lead to errors while parsing, with finally no results.

- Similarly to RDF files, the runtime for CoNLL files scales linear with the number of words in a dataset. However, the extraction process for annotation data is much simpler than for RDF and XML files, since tags only have to be read from a column of a tsv file. In fact, the runtimes for the two example CoNLL files are nearly identical, although one of them is 3 times larger and also has more database writes.

- For each XML file a sample of 5000 sentences was used. The different runtimes can be explained with the number of database write operations.

| Dataset | Type | Tripples/Lines | Writes | t[s] |
|---|---|---|---|---|
| Wordnet[42] | RDF lexicon | 2637168 | 6 | 53[43] |
| Wiktionary[44] | RDF lexicon | 3501697 | 41 | 123[45] |
| DBnary[46] | RDF lexicon | 11267006 | 79 | 190[47] |
| UD_Hindi-HDTB[48] | CoNLL corpus | 320968 | 385 | 42 |
| UD_Arabic-NYUAD[49] | CoNLL corpus | 90286 | 221 | 44 |
| kubhist-stockholms-posten[50] | XML corpus | 2812692 | 525 | 48 |
| Pride and Prejudice[51] | XML corpus | 339639 | 1683 | 120 |

Table 1: Annohub processing times

[37] `https://jena.apache.org/documentation/javadoc/arq/org/apache/jena/riot/system/StreamRDF.html`

[38] Jena command-line-tool *riot –validate*

[39] `http://librdf.org/raptor/`

[40] The persisted annotation data includes matched, but also unmatched annotations (for CoNLL and XML data only). Storing unmatched annotations ensures that these can later be automatically matched if the database if updated with an appropriate OLiA annotation model description that includes the definition of a formerly unknown tag

[41] Nevertheless, there exist RDF corpora as well

## 7. Summary & outlook

We introduced a new web portal that hosts metadata of publicly available annotated language resources. In addition to automated harvesting processes for such resources, and following the crowd-sourcing idea, registered users of the portal can contribute to Annohub by uploading datasets in order to extend the metadata in the Annohub catalogue which is released as a Linked Data set. The portal (**`https://annohub.linguistik.de/beta`**[52] is currently in the beta testing phase. Guest users (*login=acoli* and *password=guest*) can search all released resources in the Annohub dataset, but can not upload data or post comments. For registration as a beta-tester, please contact us with some information about your research interests. The source code of the project will be available at `https://github.com/ubffm/Annohub` under MPL 2.0 license.

Additional services can be provided in future releases, for example to convert language data listed in Annohub into a different format and make it available for download. For example from XML to CoNLL or CoNLL-RDF[53] format. Furthermore, providing a SPARQL endpoint, in order to query datasets listed in Annohub directly, could ease access to language data for researchers even more. This effort however, would require a considerable powerful technical infrastructure, which is not available right now. Finally, existing OLiA annotation models are steadily refined, but also new OLiA models will be added over time to cover yet unsupported annotation schemes.

[42] `http://wordnet-rdf.princeton.edu/wn31.nt.gz`

[43] 68s, including download and RDF-validation

[44] `https://lemon-model.net/lexica/wiktionary_en/en/en.nt.gz`

[45] 273s, including download and RDF-validation

[46] `https://kaiko.getalp.org/static/ontolex/latest/de_dbnary_ontolex.ttl.bz2`

[47] 275s, including download and RDF-validation

[48] `https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3424/ud-treebanks-v2.7.tgz` UD_Hindi-HDTB/hi_hdtb-ud-train.conllu

[49] `https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-3424/ud-treebanks-v2.7.tgz` UD_Arabic-NYUAD/ar_nyuad-ud-test.conllu

[50] `https://spraakbanken.gu.se/lb/resurser/meningsmangder/kubhist-stockholmsposten-1830.xml.bz2`

[51] `https://opus.nlpl.eu/download.php?f=Books/v1/parsed/en.zip`, Austen_Jane-Pride_and_Prejudice.xml

[52] Not `https://annohub.linguistik.de/de/beta` !

[53] `https://github.com/acoli-repo/conll-rdf`

## 8. Acknowledgements

The research described in this paper was conducted in the context of the Specialized Information Service Linguistics (FID), funded by German Research Foundation (DFG/LIS, 2017-2022). The author would like to thank Christian Chiarcos for providing expert advice throughout the project. We would also like to thank Thorsten Fritze and Yunus Söyleyici for technical support and Vanya Dimitrova for helpful comments.

## 9. Bibliographical References

Abromeit, F. and Chiarcos, C. (2019). Automatic Detection of Language and Annotation Model Information in CoNLL Corpora. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 23:1–23:9, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Abromeit, F., Fäth, C., and Glaser, L. (2020). Annohub – annotation metadata for linked data applications. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 36–44, Marseille, France, May. European Language Resources Association.

Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information Systems*, 5:1–22, 07.

Chiarcos, C., Fäth, C., Renner-Westermann, H., Abromeit, F., and Dimitrova, V. (2016). Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4463–4471, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Chiarcos, C., Fäth, C., and Ionov, M. (2020). The ACoLi dictionary graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France, May. European Language Resources Association.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for eHumanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1525–1531, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

McCrae, J. P. and Cimiano, P. (2015). Linghub: a linked data based portal supporting the discovery of language resources. In *SEMANTiCS*.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., and Cimiano, P. (2015). One ontology to bind them all: The meta-share owl ontology for the interoperability of linguistic datasets on the web. In *MSW@ESWC*.

Piperidis, S. (2012). The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 36–42, Istanbul, Turkey, May. European Language Resources Association (ELRA).

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 7–16
Marseille, 24 June 2022


# From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)

**Milica Ikonić Nešić*, Ranka Stanković†, Christof Schöch‡, Mihailo Škorić†**
*University of Belgrade, Faculty of Philology, Serbia
milica.ikonic.nesic@fil.bg.ac.rs, †University of Belgrade, Faculty of Mining and Geology, Serbia
{ranka.stankovic, mihailo.skoric}@rgf.bg.ac.rs
‡University of Trier, Germany; schoech@uni-trier.de

**Abstract**

In this paper we present the wikification of the ELTeC (European Literary Text Collection), developed within the COST Action "Distant Reading for European Literary History" (CA16204). ELTeC is a multilingual corpus of novels written in the time period 1840—1920, built to apply distant reading methods and tools to explore the European literary history. We present the pipeline that led to the production of the linked dataset, the novels' metadata retrieval and named entity recognition, transformation, mapping and Wikidata population, followed by named entity linking and export to NIF (NLP Interchange Format). The speeding up of the process of data preparation and import to Wikidata is presented on the use case of seven sub-collections of ELTeC (English, Portuguese, French, Slovenian, German, Hungarian and Serbian). Our goal was to automate the process of preparing and importing information, so OpenRefine and QuickStatements were chosen as the best options. The paper also includes examples of SPARQL queries for retrieval of authors, novel titles, publication places and other metadata with different visualisation options as well as statistical overviews.

**Keywords:** Wikidata, linked data, SPARQL, distant reading, literary corpus, named entity linking, ELTeC

## 1. Introduction

The COST Action "Distant Reading for European Literary History"[1] ran from 2017 to 2022 and aimed to use computational methods for the analysis of large collections of literary texts. The main goal of this networking project was to compile and analyse a multilingual open-source collection of novels, named European Literary Text Collection (ELTeC). ELTeC contains corpora of 100 novels per language written between 1840 and 1920 that are encoded in XML, are linguistically annotated and contain detailed metadata (Schöch et al., 2021).

The term *distant reading* (Moretti, 2000) describes an alternative or a complement to *close reading*: Instead of detailed, qualitative interpretations of selected literary texts, the idea is to analyse large collections of literary texts using quantitative methods of text analysis and machine learning. Formal and quantifiable textual features are used as indicators for relevant literary phenomena, with their patterns of occurrence then being related to categories such as authors, genres, or literary periods (Schöch et al., 2020).

This paper presents an approach for publishing the metadata and named entities (NE) from the sub-collections of ELTeC as linked open data. More precisely, the paper presents results for 700 novels from the first seven languages (English, Portuguese, French, Slovenian, German, Hungarian and Serbian) that are morpho–syntactically tagged (Stanković et al., 2022b) and partially annotated with named entities (Stanković et al., 2019; Frontini et al., 2020), as well as the case study on Named Entity Linking (NEL) for the Serbian ELTeC sub-collection.

Linked open data for literary texts is slowly gaining traction, as evidenced by resources such as BookSampo (Mäkelä et al., 2013) or projects like POSTDATA (Bermúdez-Sabel et al., 2021) and Mining and Modeling Text (Schöch et al., 2022). The motivation for the presented activity was to increase the visibility of the ELTeC collection, to connect it to open knowledge bases, as well as to allow searching and analyzing texts using linked open data. The incentive for the presented activity was the successful initial implementation for Serbian (Ikonić Nešić et al., 2021) that was further applied to other six languages with support of the sub-collection coordinators.

We use the term *wikification* not only for entity linking with Wikidata as the target Knowledge base, but also for creating and populating Wikidata items related to novels which will be further used for entity linking.

The crucial point for automation of wikification was the synergy of the powerful open source tools OpenRefine (Huynh, 2012) and QuickStatements (Manske, 2019). This enabled 700 novels from the core collections and 20 from extended sub-collections of ELTeC to be described in Wikidata, including associated items for their first editions, print editions, digital editions and the ELTeC (electronic) editions. This resulted in approximately 20,900 automatically added statements. To the best of our knowledge, this work is the first example of data about literary corpora for seven languages being automatically imported into Wikidata using different open source tools.

Section 2 is dedicated to the ELTeC: in Subsection 2.1 an overview of the text collection is given, in Subsec-

[1]Distant Reading for European Literary History (CA16204), `https://www.distant-reading.net`.

tion 2.2 the XML/TEI encoding of novels is explained, while in Subsection 5.1 the NER approach applied to novels is introduced.
The ELTeC Linked data model is presented in Section 3: in Subsection 3.1, the main data model, automation and the management of ELTeC Wikidata are presented, while the pipeline, from data preparation to Wikidata linking, is presented in Subsection 3.2.
The process of automation of ELTeC Wikidata population is presented in Section 4. The entity linking is described in Section 5.2: entity recognition and linking with Wikidata identifiers.
The development of a user friendly interface with predefined SPARQL queries with visualization is presented in Section 6. A set of web pages was developed with integrated results of SPARQL queries to help literary scholars that are not familiar with SPARQL. Several different visualisation options, based on Wikidata Query Service should allow new aspects of distant reading of the literary data. Section 7 concludes and summarizes our entire research and outlines several possibilities of extensions to this research.

# 2. ELTeC Text Collection

## 2.1. Overview of ELTeC Collection

Within the COST Action "Distant Reading for European Literary History", a research network of more than 200 researchers from more than 30 countries was built to foster digital, cross-lingual research into the history of the European novel. The envisaged activities were to build a multilingual corpus of European novels and develop appropriate, digital methods of analysis. Its main objective was the production of a unified, uniform, multilingual, digital novel collection dubbed the "European Literary Text Collection", or ELTeC for short (Odebrecht et al., 2021), containing novels first published between 1840 and 1920 in Europe.
ELTeC is a multilingual resource that provides learning opportunities regarding collaborative research for the European, multilingual community of researchers in (computational) literary studies. It is also a foundation for the development of cross-lingual methods and a first step towards a history of European literature that would be truly digital, multilingual and diverse (Schöch, 2022).
The novels are selected from the time period 1840-1920 and currently, 10 corpora are complete while seven more are in progress, in addition to several extension collections. The latest release (v1.1.0) was published in April 2021, containing 14 sub-collections and 1,200 novels. Its key characteristics are that each corpus represents the variety of production, that texts are encoded in XML-TEI, that they are linguistically-annotated (morpho-syntactically, NE) and that everything is published under open licences (Schöch et al., 2021; Burnard et al., 2021).
ELTeC is designed to support a wide range of distant reading methods. Such methods cover various computational approaches to literary text analysis, regarding authorship and textuality, time and space, theme and style, or character and plot (more in (Schreibman and Siemens, 2008; Eve, 2022)). Many of them have already been applied to ELTeC, among them stylometric authorship attribution (Škorić et al., 2022; Cinkova and Rybicki, 2020), stylistic analysis (Stanković et al., 2022a; Patras et al., 2021; Krstev, 2021b) or direct speech detection (Byszuk et al., 2020). Linguistic annotation and detailed metadata support many of these methods.

## 2.2. XML/TEI Encoding of Novel's Metadata

The ELTeC coding scheme was produced with no intention to present the original documents in all their original structure or layout complexity, but to make it easier to access the texts that are encoded in a predictable manner. The relevant COST Action working group agreed that the ELTeC should be delivered in a TEI–encoded format, using a schema developed specifically for the project (Burnard et al., 2021).
In order to be compliant with the TEI guidelines, a documents needs to provide metadata in the `<teiHeader>`. Each novel from the ELTeC collection at level-1 (text with structural and layout annotations) is prepared as an XML/TEI document and contains a TEI header with the following required XML elements:

- `<fileDesc>`: description of the electronic edition, which includes the title of the work and the name of the author, as well as the statements of responsibility (scanning, correction, annotation), date of publication, size (measured by the number of words). Identifiers can be assigned to authors and their work, such as VIAF and Wikidata.
- `<sourceDesc>`: brief bibliographic description of the first edition and the edition used as the source for ELTeC (if different from the first edition).
- `<profileDesc>`: description of the text in terms of meeting criteria used for the selection of novels (e.g. author's gender, novel's size, time slot of the first edition, number of recent reprints,...).
- `<revisionDesc>`: review of all changes to the digital edition since its first publication.

An opportunity for speeding up the process of data preparation for Wikidata was seen in using information already encoded in the header of each novel (Krstev, 2021a; Ikonić Nešić et al., 2021). This approach will be elaborated in Section 3.

# 3. ELTeC Linked Data Model

## 3.1. Wikidata Class Selection

Wikidata is an open source knowledge base where the underlying structure in RDF is a collection of triples,

each consisting of a subject (Wikidata item to which the claim refers), a predicate (Wikidata property), and the object (value). A value can be another item, a string, a time, a period, a location, an URL, or a quantity, depending on the property type. Statements can use qualifiers that show the contexts of the validity of the statement and they can include references. Qualifiers and references are also represented in the form of triples, where the subject is the claim.
The items and properties in Wikidata that are used to structure the ontology are:

- classes: class (Q16889133), entity (Q35120) and Wikidata metaclass (Q19361238),
- properties: instance of (P31) and subclass of (P279)

Classes conceptually group together similar items.

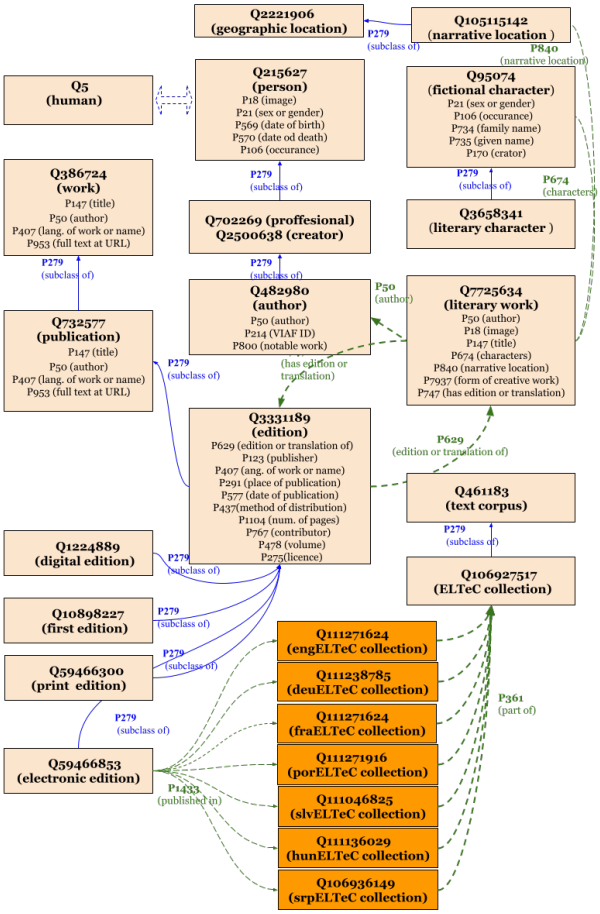


Figure 1: The class diagram of Wikidata used for novels and editions in ELTeC text collection.

Figure 1 presents the class/instance relation of all classes and relations that are used in this research. The blue lines represent "subclass of" relations between classes, while green lines presents other properties. The class person (Q215627) is used as "instance of" (P31) class humans (Q5), as recommended in the Wikidata documentation (Class person). Each item for a novel is connected with an appropriate item that is an instance of electronic edition (Q59466853), first edition (Q10898227), print edition (Q59466300) and digital edition (Q1224889) using property (P747) (has edition or translation), and every item of edition must be connected with a corresponding item for a novel with inverse property (P629) (edition or translation of). Orange boxes represent items for each of the seven corpora of ELTeC that are (P279) subclasses of electronic edition (Q59466853). All seven are published in ELTeC Collection (Q106927517) which is "subclass of" text collection (Q461183). A list of all properties that are used for authors, novels and editions is presented as a part of Wikidata: WikiProject ELTeC (Property overview). It is necessary to emphasize that for now only items for novels in the Serbian part of ELTeC are connected with appropriate items for main characters and narrative places. All items for main characters are created manually and all of them are instances of literary character (Q3658341). Narrative places are instances of class city (Q515).

### 3.2. ELTeC Data Model Aligning with Wikidata Classes

Having consistent TEI headers enabled extraction of metadata and linking with Wikidata. Data extraction was a necessary step to automate the process of importing novels and editions into Wikidata. After careful selection of classes and properties, it was necessary first to find exact mappings between them and elements of the novels' XML documents. Figure 2 shows an example of the mapping for the French novel *Lucingole* (Q111366753) written by Catulle Mendès (Q971215).
A set of metadata of the ELTeC novels was extracted from the element `<teiHeader>`, the part of which is presented in Table 1. The first column of Table 1 represents the TEI XPath to and element or attribute for ELTeC edition (the upper cell) and for different types of editions, where *type* can be first, print or digital (the lower cell). The second column contains information about the class of the instantiated data that is used for mapping. More about mapping and chosen classes can be seen in (Ikonić Nešić et al., 2021).

## 4. Automatization of ELTeC Wikidata Population

An opportunity for speeding up the process of item creation was seen in using the information encoded in the header of each novel, as explained in Subsection 2.2. The main aim of this research was to build Wikidata entities by using the model and mapping presented in Subsection 3.2 for the novels belonging to those ELTeC corpora that already provide a so-called level-2 encoding with morpho-syntactic and NE annotation: English, Portuguese, French, Slovenian, German, Hungarian and Serbian.
As the guideline model for the automation activities,

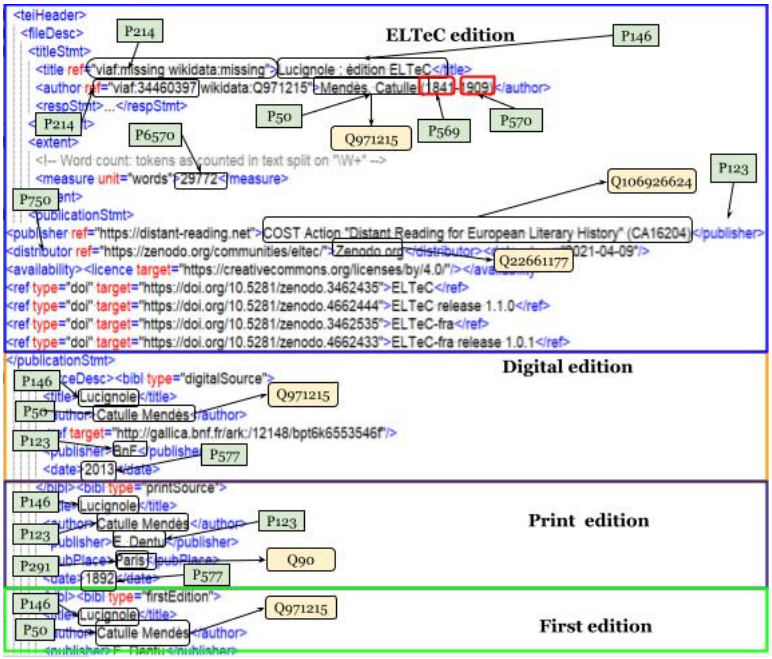


Figure 2: Mapping between metadata header and Wikidata (the novel *Lucignole* (Q111366753))

| TEI XPath to element or attribute for ELTeC edition data | element is instance of |
|---|---|
| /titleStmt/title | Q783521 (title) |
| /titleStmt/author | Q482980 (author) |
| /extent/measure[unit="words"] | Q8034324 (word count) |
| /publicationStmt/publisher | Q105044823 (publisher) |
| /publicationStmt/distributed | Q12540664 (distributor) |
| /publicationStmt/availability/licence@target | Q79719 (licence) |
| /profileDesc/langUsage/language[ident="fr"] | Q34770 (language) |
| **TEI XPath to element or attribute for different type edition data** | **element is instance of** |
| /sourceDesc/bibl[type=@typeSource]/title | Q783521 (title) |
| /sourceDesc/bibl[type=@typeSource]/author | Q482980 (author) |
| /sourceDesc/bibl[type=@typeSource]/publisher | Q105044823 (publisher) |
| /sourceDesc/bibl[type=@typeSource]/pubPlace | Q1361759 (place of pub.) |
| /sourceDesc/bibl[type=@typeSource]/data | Q1361758 (date of pub.) |

Table 1: Mapping between metadata to Wikidata for editions

the use case of SrpELTeC at Wikidata (Ikonić Nešić et al., 2021) was employed.

Data preparation and the import process were done via the synergy of OpenRefine (Verborgh and Wilde, 2013) – a tool for working with messy data, like cleaning, converting from one format to another, with the addition of external data via a web service – and QuickStatements, a Wikidata editor for adding and removing statements, tags, properties, labels and descriptions.

The following processing steps were performed on all novels with level-2 annotations:

- preparation of metadata of ELTeC sub-collections for import into Wikidata,
- import of data into OpenRefine and reconcile data with external source (Wikidata),
- importing data into Wikidata using QuickStatments,
- analysis of imported dataset using a set of SPARQL queries.

The procedure for the extraction of all metadata from the headers into one *CSV* (comma separated values) file, appropriate for further transformations and exploitation of text collections in OpenRefine, was integrated in the already existing tool for creation, management and exploitation of lexical resources *Leximir* (Stanković and Krstev, 2012).

After mapping metadata to Wikidata, OpenRefine was used to automate the data preparation, check existence and perform disambiguation. A process of manually checking of extracted metadata was required to solve some uncertainties. Namely, several instances of wrong date of birth or death of authors or missing VIAF IDs etc. were found and solved in collaboration with members of other teams of the working group for different languages.

Since author-related entries are a precondition for the automatic item creation, the OpenRefine *reconciling*

process was used to check if each entry existed. Reconciliation is the process of matching our dataset with that of an external source – in this case we use this process to identify existing items in Wikidata – a necessary step that enables linking of the file contents to the identifiers (QID) of existing Wikidata items and the creation of new ones for those that do not exist.

For missing authors, items as instances of authors (Q482980), were automatically created, with labels, description and properties such as dates of birth and death and the author's gender, which were extracted from the element `<author>` from metadata `<teiHeader>`, if the information was available.The process of entering authors in Wikidata will not be described, and we will focus on entries for novels and editions in Wikidata. The main entities involved in these tasks are: text collection (Q461183), novels (Q7725634) and version, edition, or translation (Q3331189).

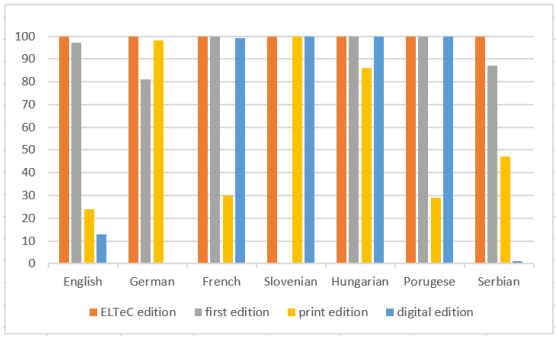


Figure 3: Statistical overview of edition items.

The next step was editing the Wikidata schema using OpenRefine. Creating a Wikidata input set schema defines subjects (items which we create), predicates (properties) that will connect subjects, and objects in RDF triples which are values of extracted metadata. The subject of the statement one or more properties whose value can be a Wikidata item, external URL, or literal (string). The subject of the statement one or more properties whose value can be a Wikidata item, external URL, or literal (string). After editing and saving the Wikidata schema, we exported it as a *QuickStatements* file and automatically added it to Wikidata.

700 novels of the ELTeC level-2 collection with 700 ELTeC (electronic) editions, 565 first editions, 414 print editions and 413 electronic editions were automatically added (totaling in approximately 20,900 statements).The statistical overview of quantities automatically added to Wikidata for each language is presented in Figure 3. More information about the metadata mapping can be found in (Ikonić Nešić et al., 2021).

# 5. NER for ELTeC

## 5.1. Literary Characters and Narrative Locations in Novels

The main goal of named entity recognition, in general, is to indicate in a text names of persons, their roles, locations, organizations, and other entities relevant for specific purposes. The NER team agreed that seven categories of entities should be indicated in the novels: PERS, ROLE, DEMO, ORG, LOC, WORK, and EVENT, which were assessed as being of the greatest importance for further literary studies (Stanković et al., 2019). Developing the NE layer of the ELTeC, testing the automatic NER for Distant Reading in ELTeC and fostering NER results and analysis are presented in (Frontini et al., 2020).

Entities belonging to one of the following NE classes were represented in Wikidata in this phase: PERS entities which correspond to main characters of a novel, ROLE entities used for their titles, professions or positions and LOC entities that designate places where the action of a novel takes place (geopolitical locations). This research was focused on two categories, PERS and LOC. The main characters of the novel can be found in the list of the extracted PERS entities, while in the LOC entity list one expects to find where the narrative of the novel is set. All entities in both categories were sorted by frequency of occurrence in each novel, and the most frequent entities are taken as literary characters (Q3658341) and narrative places, i.e. geographic location (Q2221906). This task cannot be fully automated, since the names of same characters can be mentioned in a text in a number of different ways, such as: *Čedomir Ilić*, *Čedomir*, *Ilić*, *Čeda*, and it is not clear enough if places mentioned in novel are narrative places or places that are mentioned by some characters. Using these extracted named entities we were able to manually add 123 narrative locations and 904 main characters for 69 novels to Wikidata.

The main characters were described with a set of properties: gender, profession, whether the character is fictional or not, relations between characters (husband, wife, parent, child, etc.) and professions of characters related to the main ones. Since the basic information for each novel and its author is already in Wikidata, e.g. the birthplace of an author, his/her residence at the time of writing, the place of novel's first publication, it is now possible to relate the ELTeC geodata the (place of publication and places of narrative) to other time/space coordinates, and consider more detailed mapping visualizations as presented in Section 6.

Using SPARQL query https://w.wiki/5BX7, we produce graph (Figure 4) with the number of novels that are mentioning particular locations (places) based on Wikidata. *Srbija (Serbia)* is mentioned in 39 novels and *Beograd (Belgrade)* is mentioned in 19 novels. The graph with number of characters in novels, generated using https://w.wiki/5BX9, is presented in Figure 5. It can be seen that *Djuradj Branković : istoričeskih ro-*

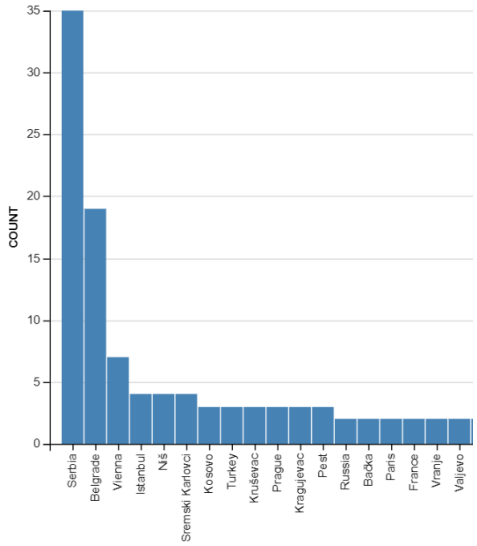


Figure 4: Number of novels mentioning the locations.

*man (Djuradj Branković : a historical novel)* has the largest number of characters (41). Currently, only the

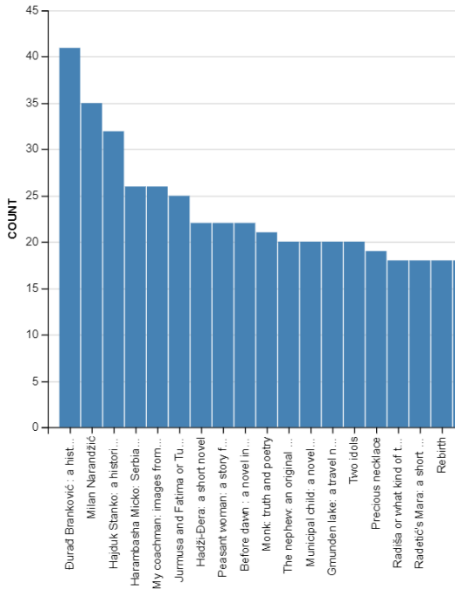


Figure 5: Main characters mentioned in novels.

narrative locations and literary characters in the Serbian part of ELTeC collection are populated, the other languages were not covered with this research.

## 5.2. From NE Extraction to Wikidata across Inception to NIF

After the main characters and narrative places were manually added, in order to validate the viability of our approach in a realistic scenario, we used the tool INCEpTION (Klie et al., 2018) for the Wikidata named entity linking on a subset of SrpELTeC collection. INCEpTION is a web-based environment for interactive text annotation and knowledge management with integrated machine-learning based assistance features and entity linking with Wikidata. The user identifies entity mentions and links them to Wikidata. To link text to an item (a class or instance), the user selects a span of text and searches for the linking item using an autocomplete text with items from Wikidata. (Castilho et al., 2018)

For the purpose of our research, two Serbian novels Ivkova slava : pripovetka (Ivko's patron saint's day: a short story) and Nečista Krv (Impure blood) were imported into INCEptION and linked with main characters and locations. We present the main characters and locations for the novel Impure blood in Table 2.

| **Main characters** | **Narrative locations** |
|---|---|
| Sofka (Q109693861) | Vranje (Q211645) |
| Magda (Q10974671) | Srbija (Q403) |
| Marko (Q109747266) | Beograd (Q3711) |
| Arsa (Q109747507) | Turska (Q43) |
| Mita (Q109747662) | Carigrad (Q16869) |
| Simka (Q109748862) | Solun (Q210176) |
| Todora (Q109748881) | Morava (Q211328) |
| Tone (Q109748906) | |
| Ahmet (Q109748924) | |
| Milenija (Q109748942) | |
| Tomča (Q109748839) | |
| Stana (Q110283369) | |
| baba-Simka (Q110826779) | |

Table 2: Characters and locations in *Impure blood*.

Figure 6 presents an example of linking character *Sofka* from the novel *Impure blood* with Wikidata item *Sofka* (Q109693861).

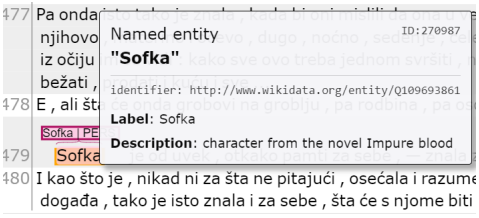


Figure 6: Inception & Wikidata NEL in *Impure blood*.

The workflow of linking characters and narrative places is presented in Figure 7.

The full process of linking entities with knowledge bases using the INCEpTION annotation platform is described in (Klie et al., 2020).

After linking annotations in INCEpTION to the knowledge base, we were able to write queries to find occurrences of all linked entities (e.g. specific persons) or find verbs that precede specific places. First steps towards RDF editions of the ELTeC corpus are publishing two Serbian novels Ivkova slava : pripovetka

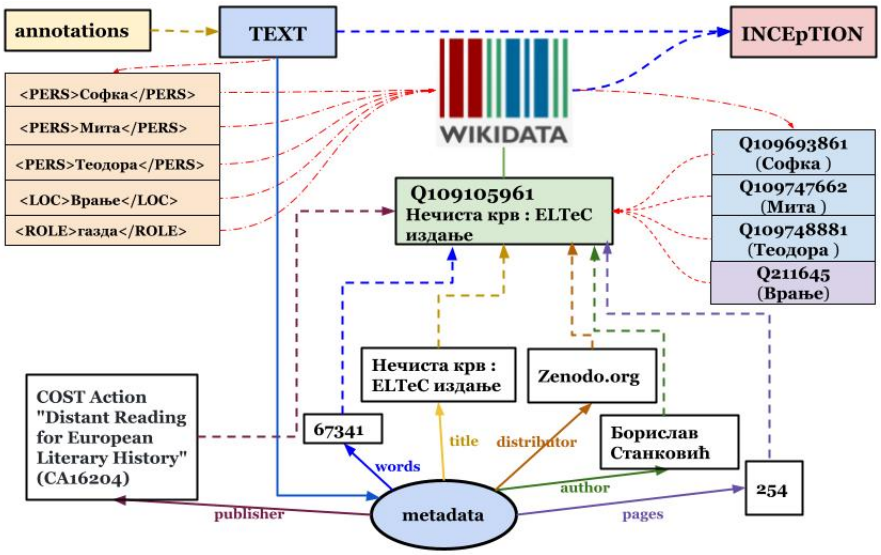


Figure 7: The Workflow: mapping metadata with Wikidata and Inception on the novel *Impure blood.*

(Ivko's patron saint's day: a short story) and Nečista Krv (Impure blood), POS-tagged, lemmatized, with NER and NEL with Wikidata, available in NIF (Ikonić Nešić and Stanković, 2022b). An example of an NIF excerpt of the novel *Nečista krv (Impure blood)* is presented in Figure 8.

## 6. The Overview of ELTeC@Wikidata by SPARQL Queries

In order to facilitate the use of Wikidata about ELTeC, we created a website with a set of predefined SPARQL queries that enable retrieval of authors, novel titles, publication places, characters, family relations of characters, their roles and others, and offer different visualization options (Ikonić Nešić and Stanković, 2022a). Different queries were written that supplied the tables: the title of the novel, the name of the author, the author's pictures, the year of publication, the main characters, and for those with imported narrative places and main characters also the relations between them, as the number of places mentioned by authors, and etc.

Figure 9 represents the timeline visualization of all authors in seven sub-collections URL. Figure 10 represents the map of first publication places.

The query presented below produces https://w.wiki/5BpU, a map of places of birth of authors, colour-coded by the time span.

```
#defaultView:Graph
SELECT DISTINCT ?person ?name ?bplace
?byear ?coord ?layer
WHERE {
?novel  wdt:P747 ?edition;
        wdt:P50  ?person.
?edition wdt:P1433 ?coll.
?coll wdt:P361 wd:Q106927517}
?person wdt:P570 ?dod;
        wdt:P19 ?place
?place wdt:P625 ?coord.
OPTIONAL{?person wdt:P569 ?dob.}
OPTIONAL{?person wdt:P18 ?image.}
BIND(YEAR(?dob) AS ?byear)
BIND(IF(byear<1851,"-1850",
IF(byear<1901,"1851-1900",
IF(byear<1951,"1901-1950",
"after-1950")))) AS ?layer)
?person rdfs:label ?name.
FILTER((LANG(?name)) = "en") ?place
rdfs:label ?bplace.
FILTER((LANG(?bplace))="en") }
ORDER BY (?byear)
```

In Figure 11, blue points represent time spans before 1700, orange between 1751-1800, green between 1801-1850, and red between 1851-1900.

The list of all novels, authors and editions for English, German, French, Portuguese, Slovenian, Hungarian and Serbian collection is presented in WikiProject ELTeC.

## 7. Conclusion and future work

In this paper we presented our recently finished activity of populating Wikidata with 720 novels from the ELTeC for seven languages (English, Portuguese, French, Slovenian, German, Hungarian and Serbian). The presented approach is language independent, so we hope that this can be an inspiration for other ELTeC corpora to expand their visibility using open linked data. The research in the digital humanities has increasingly advanced the importance of linked (open) data and with this activity we try to contribute to the distant reading methods using linked data.

Current activities include manual Named Entity Linking with Wikidata using INCEPTION platform, but future activities will be focused on training a model for automatic Named Entity Linking and exploring the formal data structures for tabular formats in language technology: CoNLL-RDF and CoNLL-RDF ontology (Chiarcos et al., 2021).

The second type of future activities will concern publishing entire annotated corpora as Linguistic Linked

```
<file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91555_91560>
        a                       nif:EntityOccurrence , nif:OffsetBasedString , nif:Word ;
        nif:anchorOf            "Sofku" ;
        nif:beginIndex          "91555"^^xsd:nonNegativeInteger ;
        nif:endIndex            "91560"^^xsd:nonNegativeInteger ;
        nif:lemma               "Sofka" ;
        nif:nextWord
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91561_91562> ;
        nif:posTag              "PROPN" ;
        nif:previousWord
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91542_91554> ;
        nif:referenceContext
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_0_96737> ;
        nif:sentence
                <file:/srv/inception/repository/project/34/document/359/source/SRP19101_1.tsv#offset_91497_91628> ;
        itsrdf:taClassRef       <PERS> ;
        itsrdf:taIdentRef       <http://www.wikidata.org/entity/Q109693861> .
```

Figure 8: SrpELTeC NIF sample


Figure 9: ELTeC sub-collections timeline


Figure 10: Map of first publication places

Open Data. Using NIF or Web Annotation / Open Annotation, the export of all level-2 novels additionally supplied with NEL layer could be published in the RDF store to be available via the SPARQL endpoint.

Following the current and future trends and challenges,

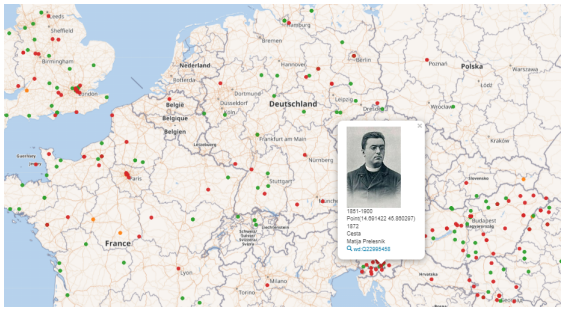


Figure 11: Birthplaces of authors, time span coloured.

as well as a proposal of the use of design patterns (Khan et al., 2021) we will apply the OntoLex-FrAC: Frequency, Attestations, Corpus Information module for complementing dictionary of lesser known, archaic words extracted from the old novels.

## 8. Acknowledgements

The preparation of the text corpora and a virtual mobility were supported by the COST Action "Distant Reading for European Literary History" (CA16204). Linked data development was done in the scope of the project "WikiELTeC–Wikidata about old Serbian novels from collection ELTeC" and supported by the COST Action "NexusLinguarum, European network for Web-centred linguistic data science" (CA18209). Both Actions are funded by COST (European Cooperation in Science and Technology, see www.cost.eu. The authors would like to thank Prof. dr Cvetana Krstev for her valuable comments which helped to improve the manuscript.

## 9. Bibliographical References

Bermúdez-Sabel, H., Díez Platas, M. L., Ros, S., and González-Blanco, E. (2021). Towards a common model for European Poetry: Challenges and solutions. *Digital Scholarship in the Humanities*.

Burnard, L., Schöch, C., and Odebrecht, C. (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*, (14).

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., and Eder, M. (2020). Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 100–104.

Castilho, R. E. D., Klie, J.-C., Kumar, N., Boullosa, B., and Gurevych, I. (2018). Linking Text and Knowledge Using the INCEpTION Annotation Platform. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 327–328.

Chiarcos, C., Ionov, M., Glaser, L., and Fäth, C. (2021). Formal Data Structures for Tabular Formats in Language Technology.

Cinkova, S. and Rybicki, J. (2020). Stylometry in a Bilingual Setup. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 977–984, Marseille, France, May. European Language Resources Association.

Eve, M. P. (2022). *The Digital Humanities and Literary Studies*. The Literary Agenda. Oxford University Press, Oxford, New York, February.

Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D., and Stanković, R. (2020). Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*.

Ikonić Nešić, M., Stanković, R., and Rujević, B. (2021). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca – Journal for Digital Humanities*, 21(2):60–87.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., et al. (2021). When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data. *Semantic Web journal*.

Klie, J.-C., Eckart de Castilho, R., and Gurevych, I. (2020). From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6982–6993, Online, July. Association for Computational Linguistics.

Krstev, C. (2021a). The Serbian Part of the ELTeC Collection through the Magnifying Glass of Metadata. *Infotheca – Journal for Digital Humanities*, 21(2):26–42.

Krstev, C. (2021b). White as Snow, Black as Night – Similes in Old Serbian Literary Texts. *Infotheca – Journal for Digital Humanities*, 21(2):119–136.

Mäkelä, E., Hypén, K., and Hyvönen, E. (2013). Fiction literature as linked open data—The BookSampo dataset. *Semantic Web*, 4(3):299–306.

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, 1 (February):54–68.

Patras, R., Odebrecht, C., Galleron, I., Arias, R., Herrmann, B. J., Krstev, C., Poniž, K. M., and Yesypenko, D. (2021). Thresholds to the "Great Unread": Titling Practices in Eleven ELTeC Collections. *Interférences littéraires/Literaire interferenties*, 25:163–187, October.

Schreibman, S. and Siemens, R. (2008). *Companion to Digital Literary Studies*. Blackwell Companions to Literature and Culture. Blackwell Publishing Professional, Oxford, hardcover edition, December.

Schöch, C., Eder, M., Arias, R., and Pieter Francois, A. P. (2020). Foundations of Distant Reading: Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts. In *Digital Humanities 2020*.

Schöch, C., Patras, R., Erjavec, T., and Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*.

Schöch, C., Hinzmann, M., Röttgermann, J., Dietz, K., and Klee, A. (2022). Smart Modelling for Literary History. *International Journal of Humanities and Arts Computing*, 16(1):78–93.

Schöch, C. (2022). What is ELTeC all about? In *Belgrade Training School 2022: Exploring ELTeC: Use-Cases for Information Extraction and Analysis. Belgrade, March 21-23, 2022*.

Stanković, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. (2019). Named Entity Recognition for Distant Reading in Several European Literatures. In *DH Budapest 2019*.

Stanković, R., Krstev, C., Šandrih Todorović, B., Vitas, D., Škorić, M., and Nešić, M. I. (2022a). Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France, june. European Language Resource Association (ELRA).

Stanković, R., Škorić, M., and Šandrih Todorović, B. (2022b). Parallel bidirectionally pretrained taggers as feature generators. *Applied Sciences*, 12(10).

Verborgh, R. and Wilde, M. D. (2013). *Using OpenRefine*. Packt Publishing, 1st edition.

Škorić, M., Stanković, R., Ikonić Nešić, M., Byszuk, J., and Eder, M. (2022). Parallel stylometric document embeddings with deep learning based language models in literary authorship attribution. *Mathematics*, 10(5).

## 10. Language Resource References

David Huynh. (2012). *OpenRefine*. `https://openrefine.org/`, 3.5.

Milica Ikonić Nešić and Ranka Stanković. (2022a). *SparqlELTeC*. `http://jerteh.rs/resursi/WIKIDATA-SPARL/`.

Milica Ikonić Nešić and Ranka Stanković. (2022b). *srpNIF*. `http://llod.jerteh.rs/ELTEC/srp/NIF/`.

Klie, Jan-Christoph and Bugert, Michael and Boullosa, Beto and Eckart de Castilho, Richard and Gurevych, Iryna. (2018). *The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation*. https://inception-project.github.io.

Magnus Manske. (2019). *QuickStatements*. https://quickstatements.toolforge.org/, 2.0.

Carolin Odebrecht and Lou Burnard and Christof Schöch. (2021). *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.* Zenodo, https://github.com/COST-ELTeC.

Ranka Stanković and Cvetana Krstev. (2012). *LeXimir - Tool for lexical resources management and query expansion*. 1.0.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 17–25
Marseille, 24 June 2022


# IMTVault: Extracting and Enriching Low-resource Language Interlinear Glossed Text from Grammatical Descriptions and Typological Survey Articles

**Sebastian Nordhoff, Thomas Krämer**
Language Science Press, gesis
sebastian.nordhoff@langsci-press.org, thomas.kraemer@gesis.org

**Abstract**

Many NLP resources and programs focus on a handful of major languages. But there are thousands of languages with low or no resources available as structured data. This paper shows the extraction of 40k examples with interlinear morpheme translation in 280 different languages from LaTeX-based publications of the open access publisher Language Science Press. These examples are transformed into Linked Data. We use LIGT for modelling and enrich the data with Wikidata and Glottolog. The data is made available as HTML, JSON, JSON-LD and N-quads, and query facilities for humans (Elasticsearch) and machines (API) are provided.

**Keywords:** Interlinear Glossed Text, Extraction, Linked Data, Low-Resource Languages, FAIR, Linguistic Linked Open Data

## 1. Introduction

There are currently 7616 spoken languages on Earth.[1] Digital resources for these languages are in a very skewed distribution, as surveyed by (Joshi et al., 2020). English has good resources, a few additional languages have satisfactory resources and three other groups of languages can at least list some resources of a certain size or quality. These four groups arrive at 72 languages altogether. The great majority of languages, however, have only minimal resources in both extent and annotation depth, and many languages have no resources available for NLP at all. The latter two groups comprise 93.87% (2 413) of all languages investigated by (Joshi et al., 2020) (Table 1). Beyond that, there are another 5 000 languages which did not even make it into the (Joshi et al., 2020) survey. As we start the International Decade of Indigenous Languages[2] in 2022, this very skewed distribution is concerning.

## 2. Low Resource Languages and Diversity Linguistics

While the NLP community has not produced structured datasets for these low/no resource languages, structured data does indeed exist within the field of Diversity Linguistics. Diversity Linguistics is the field which concerns itself with the variety of languages spoken in the world. This concerns in-depth treatment of a particular language (grammatical description) as well as large-scale comparison of a given phenomenon (e. g. position of the verb before or after the object) in hundreds or thousands of languages. This comparative work can be found in articles in journals or edited volumes, in monographs, or also in databases. We can name AUTOTYP[3] or the CLLD datasets (WALS,[4] APiCS[5]), of which there are 19 as of 2022.

The academic inquiry is complemented by language archives where audiovisual data are stored, some of them transcribed, translated and glossed, in varying percentages. We can name ELAR,[6] AILLA,[7] TLA,[8] Paradisec.[9] See (Nordhoff, 2020a) for a breakdown of their accessible holdings.

These different data sources have been tapped into over time: academic books and articles ((Lewis and Xia, 2010; Xia et al., 2014)), typological databases ((Chiarcos and Ionov, 2019; Ionov, 2021)), and language archives ((Nordhoff, 2020a; Nordhoff, 2020b; von Prince and Nordhoff, 2020)), producing structured data which allows for programmatic and quantitative approaches.

## 3. The Example Sentence

While the field of Diversity Linguistics is actually quite far from NLP in its practices, it produces nevertheless semi-structured texts. This structure can be exploited to retrieve meaningful elements. The most common datatype is the linguistic example with interlinear morpheme translation (IMT). In this kind of element, we have part-whole relations between morphemes, words and sentences, and translational equivalence relations on the word level and the sentence level between the source language (white) and the translation (grey). This is shown in Figure 1.

From examples like this, we can extract morpheme-to-morpheme translations, which can be used to populate

[1] https://glottolog.org/glottolog/glottologinformation

[2] https://en.unesco.org/idil2022-2032

[3] https://github.com/autotyp/autotyp-data/tree/v1.0.0

[4] http://wals.info

[5] http://apics-online.info

[6] https://www.elararchive.org

[7] https://ailla.utexas.org

[8] https://archive.mpi.nl/tla

[9] https://catalog.paradisec.org.au

| | | criteria | | | | |
|---|---|---|---|---|---|---|
| | Class | unlabeled data | labeled data | example | # lgs | % |
| 5 | winners | good | good | Spanish | 7 | 0.28 |
| 4 | underdogs | good | insufficient | Russian | 18 | 1.07 |
| 3 | rising stars | good | none | Indonesian | 28 | 4.42 |
| 2 | hopefuls | ? | smallish sets | Zulu | 19 | 0.36 |
| 1 | scraping-bys | smallish | none | Fijian | 222 | 5.49 |
| 0 | left-behinds | none | none | Warlpiri | 2 191 | 88.38 |

Table 1: Joshi et al's classes

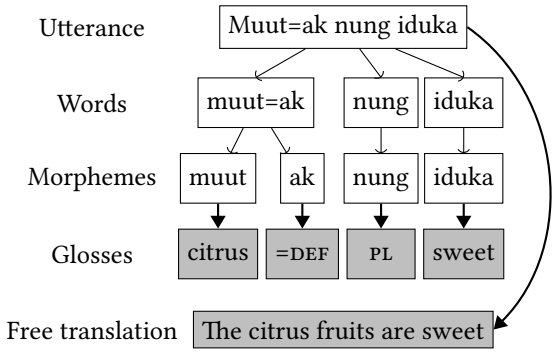


Figure 1: An example of interlinear text (`https://imtvault.org/b/157/ex/wl09-cb9806ea53.htm`, (Klamer et al., 2017)). Light arrows denote part-whole relations; thick arrows denote translational equivalents. Note that there is no translation for the word level.

a dictionary or a word list. The data model used here is discussed in more detail in Section 4. Data sources for interlinearized examples can be found in a variety of places in different formats (see Section 7.4).

## 4. Data Modelling

The interlinear sentence has received quite some theoretical treatment. The first technical approach was the implementation in the program Shoebox, which would later become Toolbox.[10] The representation used therein was actually never intended to be used in a productive environment, but turned out to become the mainstay for language documenters for more than two decades. Shoebox/Toolbox was developed by SIL, who discontinued development in favour of FLEx, an XML based tool.[11] In parallel, ELAN[12] ((Wittenburg et al., 2006)) is another XML-based tool for the representation of correpondences and part-whole relations in glossed texts ((Nordhoff, 2020a)). While XML suggest a good perspectives for programmatic extraction of data, (Nordhoff, 2020a) reports that while syntactically valid XML, the ELAN files retrieved from language archives are semantically wildly heterogeneous, making a principled approach very difficult (also compare (Cimiano et al., 2020, 4)).

On a more theoretical level, (Drude, 2002) proposed a very elaborate model with a multiplicity of tiers. The XML Interlinear Glossed Text (XIGT, (Goodman et al., 2015)) format has a recursive structure instead, allowing for an arbitrary number of tiers ((Xia et al., 2014)). (Chiarcos and Ionov, 2019) and (Ionov, 2021) developed a Linked Data version of XIGT, called LIGT, also used in (Nordhoff, 2020a; Nordhoff, 2020b). For the purposes of this paper, a very simple data model distinguishing the tiers of "utterance" and "word", with respective translations, is sufficient; the level of "morpheme" is disregarded. Basic storage is done in JSON, while transformations into JSON-LD, RDF, and CLDF are also made available. An additional morpheme tier could also have been made available, but it was determined that data consumers could easily create such more granular structures easily themselves should the need arise and that it was not necessary to provide an artificially inflated dataset.

## 5. Data Sources

Extraction of interlinear examples from documents has a comparatively long history. The ODIN project ((Lewis and Xia, 2010; Xia et al., 2014)) [13] crawled the web for pdfs and tried to extract the examples. Copyright problems and the generally poor extraction facilities, however, posed great challenges for this endeavour. While ODIN is still up and running, it uses meanwhile outdated technology (eg HTML framesets), has encoding issues and does not provide dereferenceable URIs for the examples (Figure 2).

Another source for interlinearized texts are cross-linguistic databases. The Atlas of Pidgin and Creole

[10] `https://software.sil.org/toolbox`

[11] `https://software.sil.org/fieldworks`

[12] `https://archive.mpi.nl/tla/elan`

[13] `http://odin.linguistlist.org`

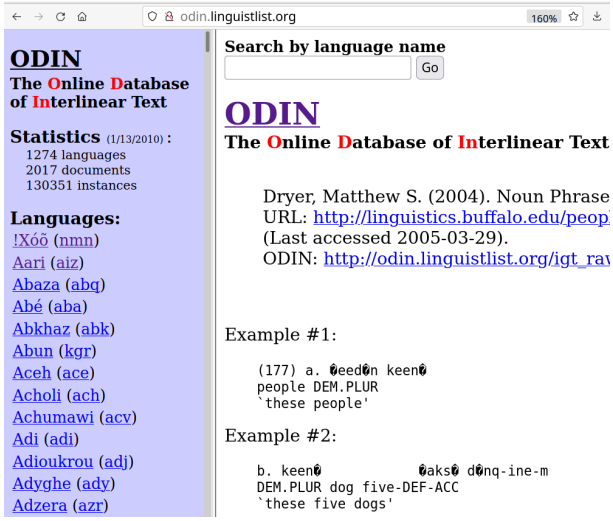


Figure 2: A screenshot of the ODIN website, showing an example of the Aari language. Note the URL, which does not give the ID, and the encoding problems. The example given has the "Verified" rating "highest". There is also "high", "auto" and "low", with presumably worse quality.

Language Structures (APiCS[14], (Michaelis et al., 2013)) offers its example sentences for download in the CLDF format ((Forkel et al., 2018)). These examples were parsed by (Chiarcos and Ionov, 2019), who used them to develop the LIGT format. The APiCS data have the advantage of being available under a free license.
(von Prince and Nordhoff, 2020) and (Nordhoff, 2020a; Nordhoff, 2020b) downloaded data from a variety of language archives, which store ELAN files. ELAN is an XML-format with explicit correspondences between morphemes, words, and sentences. These ELAN files were than converted to the RDF LIGT format, drawing on previous work by (Nordhoff et al., 2016).
Published books, most databases and most of the language archives share the problem of unclear copyright status, which hinders dissemination and reuse. Enter Language Science Press.

## 6. Language Science Press

Language Science Press is an open-access publisher in linguistics which has published over 180 books (monographs and edited volumes) since 2014. All books are released under a CC-BY license, and the LaTeX source code is available on GitHub. The source code is structured in an identical manner for most books as far as naming conventions and directory structure are concerned, so that a given approach can nicely scale. This is different from, say, the ODIN project or the work on language archives, which had to deal with wildly divergent input data.

[14] `https://apics-online.info/`

For the issue at hand, the task was to retrieve a maximum of interlinear example data from Language Science Press, analyze them, enrich them, and make them available for reuse.

## 7. Data Handling

### 7.1. Data Source Identification

For this task, we downloaded the source code of free Language Science Press books. LangSci books have an ID, which corresponds to a GitHub repo. For instance, the book *Attributive constructions in North-Eastern Neo-Aramaic* with the catalog page `https://langsci-press.org/catalog/book/123` has the GitHub repo `https://github.com/langsci/123`.
Not all IDs correspond to published books as some submitted books are rejected. Currently, there are 211 titles listed on the catalog page.

### 7.2. Data Extraction

The highest current ID is 349, so we iterated through the numbers from 1 to 349 and tried to clone the resulting GitHub address. This yielded 173 repositories with usable tex files. For these repositories, we retrieved 3 033 tex files with a total of 25 020 723 words. The content of these tex files was parsed for examples following the gb4e syntax.
This is illustrated in (1) from (Klamer et al., 2017).

(1) Kamang (Schapper, fieldnotes)
Muut=ak nung iduka.
citrus=DEF PL sweet
'The citrus fruits are sweet.'

The source code for this example is

```
\langinfo{Kamang}{}{Schapper, fieldnotes} \\
\gll  Muut=ak nung iduka. \\
  citrus=\textsc{def} \textsc{pl} sweet    \\
\glt `The citrus fruits are sweet.'
```

The LaTeX markup like `\langinfo`, `\gll` and `\glt` allow us to meaningfully identify the language name (first line), the source line (starting with `\gll`), the interlinear morpheme translation (following the source line), and the translation (following `\glt`). All examples must have the `\gll` and `\glt` parts; the `\langinfo` part is optional, as is citation information (not shown in the example).
The extraction is complicated by a variety of intervening TeX markup, such as `\textsc{}` for small capitals and similar. The raw data for source line, interlinear line, and translation line have thus to be stripped of their TeX markup. After that, the words of the source and interlinear line can be tokenized and matched. Examples which differ in the number of words between the source line and the interlinear line are discarded. This yields 39,352 vanilla examples with a unified structure.

## 7.3. Data Linking

For the purposes of this paper, we distinguish a `ligt:Utterance`, which contains a `ligt:WordTier`, which in turn has a number of `ligt:Words`.[15] The relation between those items are given in Figure 3. For a more elaborate representation, see (Chiarcos and Ionov, 2019).

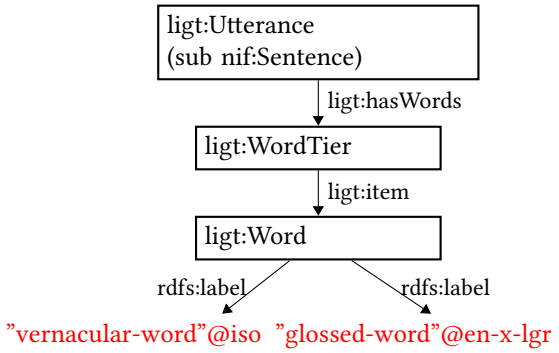


Figure 3: The relevant part of the LIGT model. Note that the predicate `rdfs:label` is assigned twice, but with different language tags. The vernacular label gets the ISO 639-3 code of the language under discussion, while the label containing the glosses gets an RFC 5646 label “en” with a private subtag “-x-lgr” for Leipzig Glossing Rules, following specifications in Section 2.2.7 of the RFC.

We link the extracted examples to Glottolog, the Leipzig Glossing Rules, and Wikidata

### 7.3.1. Glottolog

Glottolog[16] ((Nordhoff and Hammarström, 2012; Hammarström and Forkel, 2021)) is a knowledge base which contains information about 8,155 languages, some with dialects, and the genealogical classification, amounting to a set of language family trees with more than 25,000 nodes. These nodes have a human readable label (such as “Kamang”) and a so-called glottocode with a persistent URL (e.g. `https://glottolog.org/resource/languoid/id/kama1365`). We extracted all language names from the freely available Glottolog dataset ((Hammarström et al., 2021)). and matched the language names we retrieved from our examples to retrieve the corresponding Glottocode. In addition, all examples from books with the title “A grammar of X” were automatically assigned to the language X. If the language name retrieved for a given example could not be matched to a Glottocode in this way, we did a web lookup on glottolog.org with the partial name search. If the result set had the length 1, or if only one result was of type “language” (rather than “dialect” or “family”), this result was retained. Altogether, this yielded 17,425 examples with metadata on source language, for a total of 280 different languages. See Appendix A for a list.

### 7.3.2. Leipzig Glossing Rules

The Leipzig Glossing Rules are a list of standard abbreviations for grammatical categories such as NOMinative or ACCusative, which are followed by most publications in Diversity Linguistics. We extracted these from the interlinear line and linked them to `https://www.eva.mpg.de/lingua/resources/glossing-rules.php#`.

### 7.3.3. Named Entity Extraction

There are close to zero NLP tools available for the languages studied in the field of Diversity Linguistics. But fortunately, we have translational equivalents into English for our example sentences. A translation is a faithful rendering of the meaning of a sentence in a given language in another language. Therefore, we can actually use our English translation as a proxy for named entity extraction, as the entities/-concepts should match between the source and the translation. We ran the translation sentences through `https://cloud.science-miner.com/nerd/service/disambiguate`. Upon inspection, a number of the concepts turned out to be misretrievals. For example, translations with “don’t” in them are linked to the Wikidata Q17646620, which is about an Ed Sheeran song with the same title. A blacklist was created for these cases.
In a second step, the base concepts were matched with their Wikidata superclasses using the predicates p31 ‘instanceOf’ and p279 ‘subclassOf’. This allows us to assert that a goat is a mammal is an animal is an organism, greatly enhancing the querying possibilities. This is relevant for instance when linguists want to test hypotheses about certain verbs being sensitive to [±ANIMATE]. Section 7.6 will discuss querying in more detail.
Unfortunately, Wikidata does not provide a very clean ontology. Five problems were discovered:

1. misunderstandings of the predicate subclassOf (sweat > excrement > biodegradable waste > waste > bad)

2. useless use of upper ontologies (all sounds are acoustic waves are elastic waves are mechanical waves are waves are oscillations are changes are occurrences are temporal entities are spatio-temporal entities are entities)

3. conjunct categories (“inflammable solid”) which needlessly inflate the category count. The categories “inflammable substance” and “solid substance” would have been sufficient.

4. Eurocentrism (housekeeping activities are “activities of households as employers; undifferenti-

[15] Unfortunately, the PURL for LIGT did not resolve while we were writing this paper, so our resources point to a local copy of the LIGT ontology instead.

[16] `https://glottolog.org`

ated goods- and services-producing activities of households for own use" (Q29584238) as part of the Statistical Classification of Economic Activities in the European Community). This is irrelevant in an African context.

5. Other regiocentrisms (all baked items are Bánh; all dairy produce is part of some Russian classification "dairy products and ice cream, as well as services" (Q27149326)).

A blacklist of nearly 1500 entries had to be created to weed out problems caused by the listed shortcomings. Taking into account this second blacklist, we arrive at an augmented count of 28,777 entity tokens (6,773 types). The most frequent concepts are: food (Q2095, 833 instances), organism (Q7239, 788 instances), animal (Q729, 486 instances).

### 7.4. Data Storage and representation

The extracted examples (see 7.2) are further processed in two forms. One leads to csvw[17]-based CLDF representation[18], another pipeline feeds the IMTVault search and API available at `https://imtvault.org`.

For IMTVault, extracted examples are transformed into plain JSON as well as expanded JSON-LD 1.1.[19] Following the w3c best practises[20], we chose the expanded representation, as no explicit context reference is needed in downstream processing. Additionally, we use the robust titanium-json-ld library for JSON-based Serialization for Linked Data,[21] which provides sound support for transformation from JSON to JSON-LD 1.1, and from JSON-LD to RDF N-Quads.

The plain JSON representation is also used to create a search index based on elasticsearch, which serves the faceted user interface for search available at `https://imtvault.org/search`.

This allows us to present linguistic examples in a suitable way to query for both humans and machines (Figure 4) using either static, referable snapshots of the collection or dynamically via http based retrieval services.

### 7.5. Minting / URL Resolution

IMTVault has a built-in URL resolver to refer to books and examples, which can be prompted for various formats. The URL pattern includes two dynamic path elements, the book ID (taken over from Language Science Press) and a generated utterance ID.

**Resolving utterances** Utterances are identified by book ID and an example ID generated as the hexdigest of hashing the sourceline with SHA-256, truncated to 10 digits. The resolver supports four representations: minimal html (appending .htm to the URL pattern), plain JSON (.json), expanded JSON-LD 1.1 (.jsonld) or RDF N-Quads Dataset 1.1 representation (.nq),[22] leading to `https://imtvault.org/b/80/ex/01-9383b907b9.htm`, `https://imtvault.org/b/80/ex/01-9383b907b9.json`, `https://imtvault.org/b/80/ex/01-9383b907b9.jsonld`, and `https://imtvault.org/b/80/ex/01-9383b907b9.nq`, respectively.

**Resolving books** Without a file ending, the resolver will redirect to the original publication as landing page of a book at LangSciPress (`https://imtvault.org/b/157`). With the file endings .htm or .ld provided, the resolver will generate a list of all examples found in the respective book. `https://imtvault.org/b/157.json` will thus return a json list of all 99 examples from book 157, *The Alor-Pantar languages: History and typology. Second edition.*

### 7.6. Data Querying

**Query search index** The elasticsearch index can be queried programmatically. The following curl command executes a query for 'banana' to the IMTVault index of utterances. If not using an API tool such as postman[23] or insomnia,[24] the XSRF-TOKEN value needs to be obtained beforehand.

```
curl 'https://imtvault.org/express/iss/_search'
  -H 'Cookie: XSRF-TOKEN=XXXX'
  --data-raw '{"query": {
                "multi_match": {"query": "banana"}
              }}'
```

The query can be adapted as required, following the elasticsearch query syntax.[25] For users interested in running their queries locally, the CLDF data can be loaded into a SQLite database providing yet another query platform.

## 8. FAIR language examples

We applied the best practises known as the FAIR data principles[26] in the implementation of IMTVault. Findability, accessibility, interoperability, and reusability of linguistic resources are achieved to varying degrees:

- F1. (Meta)data are assigned a globally unique and persistent identifier. See the patterns in Section 7.5. The identifiers are unique, and persistent.

[17] (Tennison, 2014)

[18] `https://github.com/langsci/imtvault/tree/main/cldf`

[19] `https://www.w3.org/TR/json-ld11/`

[20] `https://w3c.github.io/json-ld-bp/#use-json`

[21] `https://github.com/filip26/titanium-json-ld`

[22] `https://www.w3.org/TR/n-quads/#n-quads-language`

[23] `https://www.postman.com`

[24] `https://insomnia.rest`

[25] `https://www.elastic.co/guide/en/elasticsearch/reference/6.8/full-text-queries.html`

[26] `https://www.go-fair.org/go-fair-initiative/`

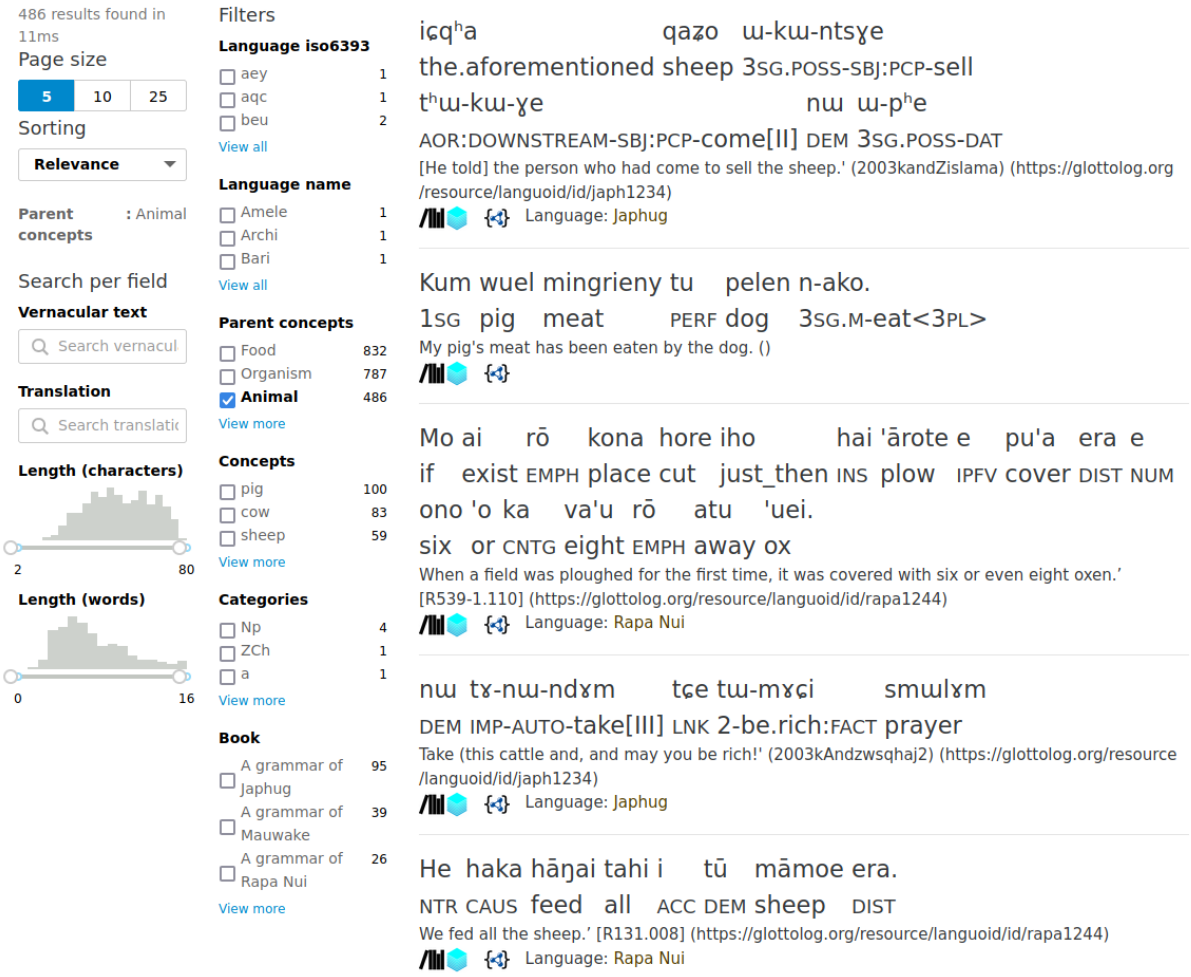


Figure 4: Querying facilities for humans. The screenshot shows a query for the topic “animal”. The screenshot shows the result list including sentences with interlinear morpheme translation from Japhug and Rapa Nui, covering different kinds of animals (sheep, pig, dog, oxen, cattle).

- F2. Data are described with rich metadata (defined by R1 below): The utterances are described using the relevant metadata schemes and referencing the original publication.
- F3. Metadata clearly and explicitly include the identifier of the data they describe: An identifier for each utterance is generated by IMTVault
- F4. (Meta)data are registered or indexed in a searchable resource : IMTVault provides a user interface for search for humans. The backing index can be queried (Section 7.6).
- A1.1 The protocol is open, free, and universally implementable: HTTP and Elasticsearch/Lucene query language are open standards.
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary: IMTVault implements authentication and authorisation. While currently all resources are available without restriction, IMTVault could handle embargoes or other types of access control if required.
- A2. Metadata are accessible, even when the data are no longer available : As data are embedded into the metadata, this does not apply.
- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. JSON and JSON-LD are W3C recommendations.
- I2. (Meta)data use vocabularies that follow FAIR principles: The vocabularies used (RDF Schema, Dublin Core terms/elements, liodi/ligt) themselves comply with the FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data : We reference Wikidata, Glottolog, and the Leipzig Glossing Rules in a qualified manner.

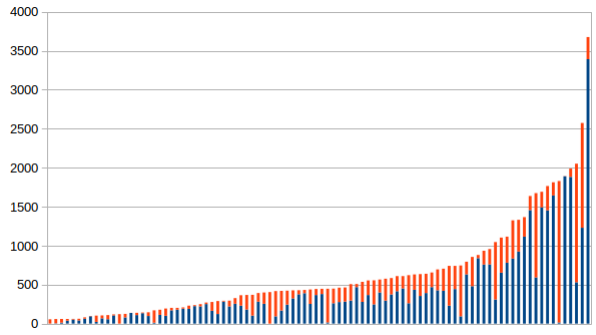


Figure 5: Retrieved examples (blue) vs skipped examples (red) for all books with a quorum of at least 50 examples.

- R1.1. (Meta)data are released with a clear and accessible data usage license : The License is CC BY 4.0 and indicated in both API responses and the user interface for search.
- R1.2. (Meta)data are associated with detailed provenance : The original publication is named and referenced. In addition, the primary citation is given as well if it could be retrieved.
- R1.3. (Meta)data meet domain-relevant community standards, which are Glottolog and LGR in our case.

## 9. Evaluation

All tex files retrieved from GitHub together contain 60 615 LaTeX commands `\gll` signalling interlinear examples. Of these, 39 352 were retrieved for IMTVault. Figure 5 gives the amount of retrieved and non-retrieved examples per book. The average of examples retrieved per `\gll` passage is 62.91%. If we compute this number per book, we arrive at a median value of 66.46%.

Authors often use `\gll` for certain elements which are not interlinear examples in the strict sense, so there can be good reasons to skip them. We investigated how succesful our algorithm was in sorting the relevant (retain) from the irrelevant (discard) examples. We drew a random sample of 100 passages introduced by `\gll` from the 60 615 and inspected manually whether this passage was correctly/incorrectly retained/discarded as an example. This was done in two steps. At first, a book was drawn at random, then, an example was drawn among the ones present in the book. This was repeated until 100 examples were reached. The reason for this two-tiered approach was that otherwise books with many examples such as *A grammar of Japhung* with over 3500 would have completely dominated the set. For the drawn examples, the pdf, the tex code, and the representation on IMTvault.org were compared. Among the 100 examples drawn, 16 were not good interlinear text and should be discarded. This had been done correctly for all of them. Most often, the reason for this was a missing translation. 84 should have been retained, but this was only the case for 72 of them. 12 were missed, or one in seven. The precision was thus 100% while the recall was 85.7%, giving an aggregate F-score of 91.9%.

Turning to concepts, the sample was extremely sparse. May sentences were of the type *Why read the book?*, which is too short and bland to do meaningful Named Entity Recognition. As such, only 6 concepts were correctly attributed to examples of the sample, while a further 6 were misattributions, often of pop songs with banal titles such as *Tender Years* by George Jones or *Live Life* by the Kinks. We conjecture that concept retrieval might have a very skewed distribution: grammatical descriptions in general have longer and more colourful examples, which are better suited for NER, while more theoretical works tend to have very barren examples, which are boiled down to the minimum, eg *John sees Mary*. If this is the case, we should find more named entities in texts from endangered language archives as well, cf. (Nordhoff, 2020b). Further research will test this hypothesis.

## 10. Conclusion and Outlook

We started with the observation by (Joshi et al., 2020) that over 90% of the world's languages have no NLP resources. We now provide 40 000 sentences in 280 languages, most of them no/low resource, as a structured dataset under a free license for reuse. The dataset respects the FAIR principles as well as the Linked Data Principles. We have a clearly defined pipeline, a storage format, a query/dissemination platform and consumers downstream. Language Science Press will continue to produce about 30 books a year, but there are other Open Access publishers whose publications could also be crawled to extract interlinear examples. An obvious candidate would be the Diamond-OA journal *Glossa*.[27]

This resources improves on ODIN or the interlinear text extracted from language archives reported in (Nordhoff, 2020a; Nordhoff, 2020b) in that the data are available under an open license and good facilities for querying and dereferencing are in place. As compared to the APiCS set created by (Chiarcos and Ionov, 2019), IMTVault has added about the double the amount of sentences (40k as compared to 18.5k for APiCS) and a more extensive range of formats and querying possibilities.

Integration of the APiCS data by (Chiarcos and Ionov, 2019) is a logical next step, as is the integration of data from endangered language archives ((von Prince and Nordhoff, 2020)), to the extent that the licenses employed there permit this. Further refinement of Named Entity Recognition will be necessary, as well as better algorithms for the identification of the language an example is in based on the surrounding text.

[27] https://www.glossa-journal.org

## 11. Bibliographical References


Chiarcos, C. and Ionov, M. (2019). Ligt: An LLOD-native vocabulary for representing interlinear glossed text as RDF. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, number 70 in OpenAccess Series in Informatics (OASIcs), pages 3:1–3:15, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

Drude, S. (2002). Advanced glossing: A language documentation format and its implementation with Shoebox. In Peter Austin, et al., editors, *Proceedings of the International LREC workshop on Resources and Tools in Field Linguistics*.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Goodman, M. W., Crowgey, J., Xia, F., and Bender, E. M. (2015). Xigt: extensible interlinear glossed text for natural language processing. *LREC*, 49(2):455–485.

Hammarström, H. and Forkel, R. (2021). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). *Glottolog 4.5*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ionov, M. (2021). APiCS-Ligt: Towards semantic enrichment of interlinear glossed text. In Dagmar Gromann, et al., editors, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 27:1–27:8, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6282–6293.

Klamer, M., Schapper, A., and Corbett, G. (2017). Plural number words in the alor-pantar languages. In Marian Klamer, editor, *The Alor Pantar languages*, number 3 in Studies in Diversity Linguistics, page 365–403. Language Science Press, Berlin.

Lewis, W. D. and Xia, F. (2010). Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing (LLC)*, 25(3):303–319.

Susanne Maria Michaelis, et al., editors. (2013). *Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://apics-online.info.

Nordhoff, S. and Hammarström, H. (2012). Glottolog/langdoc: Increasing the visibility of grey literature for low-density languages. In *Proceedings of LREC 2012*.

Nordhoff, S., Tuttle, S., and Lovick, O. (2016). The Alaskan Athabascan Grammar Database. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 5. European Language Resources Association (ELRA).

Nordhoff, S., Krämer, T., and Forkel, R. (2022). IMT Vault (v1.0). Data set.

Nordhoff, S. (2020a). From the attic to the cloud: mobilization of endangered language resources with linked data. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 10–18, Marseille, France, May. European Language Resources Association.

Nordhoff, S. (2020b). Modelling and annotating interlinear glossed text from 280 different endangered languages as linked data with LIGT. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 93–104, Barcelona. Association for Computational Linguistics.

Tennison, J. (2014). a primer, CSV on the web. W3C working group note. Cambridge: W3C.

von Prince, K. and Nordhoff, S. (2020). An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of LREC 2020*. LREC, Marseille.

Wittenburg, P., Hennie, B., Russel, A., Klassmann, A., and Sloetjes, H. (2006). *ELAN: A Professional Framework for Multimodality Research*.

Xia, F., Lewis, W. D., Goodman, M. W., Crowgey, J., and Bender, E. M. (2014). Enriching ODIN. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, number 2014.

# A. Appendix

This is a list of languages and their glottocodes for which at least one interlinear example could be retrieved.

| | |
|---|---|
| abui1241 | Abui |
| adan1251 | Adang |
| afri1274 | Afrikaans |
| agua1253 | Aguaruna |
| akan1250 | Akan |
| alam1246 | Alamblak |
| alba1267 | Albanian |
| aleu1260 | Aleut |
| amel1241 | Amele |
| amis1246 | Amis |
| anti1246 | Antioch |
| arab1395 | Arabic |
| arch1244 | Archi |
| assa1263 | Assamese |
| awad1243 | Awadhi |
| awji1241 | Awjilah |
| awtu1239 | Awtuw |
| bamu1253 | Bamun |
| bari1284 | Bari |
| bari1286 | Bariai |
| basq1248 | Basque |
| bath1244 | Baṭḥari |
| bava1246 | Bavarian |
| beja1238 | Beja |
| bena1259 | Bena |
| beng1280 | Bengali |
| berb1260 | Berber |
| bero1242 | Berom |
| bezh1248 | Bezhta |
| bilo1248 | Biloxi |
| bium1280 | Biu-Mandara |
| blag1240 | Blagar |
| bong1285 | Bongo |
| bong1298 | Bongor |
| bora1263 | Bora |
| braj1242 | Braj |
| braz1246 | Brazilian Portuguese |
| budu1265 | Buduma |
| bukh1238 | Bukharic |
| buku1249 | Lubukusu |
| buna1278 | Bunaq |
| bund1253 | Bundeli |
| buru1296 | Burushaski |
| cant1236 | Cantonese |
| capp1239 | Pharasiot |
| cayu1261 | Cayuga |
| cent1972 | Central Kurdish |
| chum1261 | Chumburung |
| coos1249 | Hanis Coos |
| copt1239 | Coptic |
| cusc1236 | Cuzco Quechua |
| cypr1249 | Cypriot Greek |
| dadi1249 | Dadiya |
| dani1285 | Danish |
| dido1241 | Tsez |
| digo1243 | Digo |
| dink1262 | Dinka |
| doma1258 | Jerusalem |
| dutc1256 | Dutch |
| dyir1250 | Dyirbal |
| efik1245 | Efik |
| egyp1253 | Egyptian Arabic |
| elem1253 | Eleme |
| enga1252 | Enga |
| ewee1241 | Ewe |
| faro1244 | Faroese |
| fefe1239 | Fe'efe'e |
| fern1234 | Pichi |
| finn1318 | Finnish |
| fore1270 | Fore |
| fuli1240 | Fuliiru |
| furu1242 | Furu |
| fyam1238 | Fyem |
| gaaa1244 | Ga |
| ghod1238 | Godoberi |
| gida1247 | Gidar |
| gree1276 | Greek |
| guja1252 | Old Gujarati |
| gunn1250 | Gungbe |
| guro1248 | Guro |
| gyel1242 | Gyele |
| hait1244 | Haitian Creole |
| hali1245 | Coastal Marind |
| halk1245 | Halkomelem |
| hass1238 | Ḥassāniyya |
| hind1269 | Hindi |
| hung1274 | Hungarian |
| hunz1247 | Hunzib |
| hupd1244 | Hup |
| icel1247 | Icelandic |
| ikkk1242 | Ik |
| inan1242 | Inanwatan |
| indo1316 | Indonesian |
| indo1319 | Indo-European |
| inui1246 | Inuit |
| iraq1241 | Iraqw |
| ital1282 | Italian |
| itza1241 | Itzá |
| japh1234 | Japhug |
| jara1276 | Jarawara |
| jita1239 | Jita |
| kabw1241 | Kabwa |
| kaby1243 | Kabyle |
| kaer1234 | Kaera |
| kagf1238 | Ut-Ma'in |
| kala1372 | Kalasha |
| kama1365 | Kamang |
| kava1241 | Kavalan |
| kelo1247 | Klon |
| keng1240 | Kenga |
| khez1235 | Khezha Naga |
| khuz1234 | Khuzestan |
| kild1236 | Kildin Saami |
| kili1267 | Kilivila |
| kima1244 | Kimaragang |
| kips1239 | Kipsigis |
| klao1243 | Klao |
| kohu1244 | Kohumono |
| komi1268 | Komi-Zyrian |
| kore1280 | Korean |
| kulu1253 | Tibeto-Burman |
| kuma1276 | Nêlêmwa-Nixumwak |
| kumz1235 | Kumzari |
| kway1241 | Kwaya |
| kwom1262 | Kwoma |
| laca1243 | Lacandón |
| lako1247 | Lakota |
| lamm1241 | Western Pantar |
| late1256 | Late Egyptian |
| lati1261 | Latin |
| latv1249 | Latvian |
| lavu1241 | Lavukaleve |
| lele1264 | Lelemi |
| lese1243 | Lese |
| lezg1247 | Lezgian |
| limb1268 | Limbum |
| loni1238 | Loniu |
| lugb1240 | Lugbara |
| lule1254 | Lule Saami |
| mait1250 | Maithili |
| mako1251 | Makonde |
| mala1464 | Malayalam |
| malt1254 | Maltese |
| mamm1241 | Mam |
| mang1381 | Mangarrayi |
| mang1394 | Mangbetu |
| mani1292 | Meithei |
| maoo1244 | Mao |
| maor1246 | Maori |
| mapu1245 | Mapudungun |
| mauw1238 | Mauwake |
| maya1287 | Mayan |
| mayo1261 | Mayogo |
| mege1234 | Megeb |
| mehr1241 | Mehri |
| meje1239 | Meje |
| mesk1242 | Meskwaki |
| mian1256 | Mian |
| midd1317 | Middle English |
| midd1321 | Middle Dutch |
| midd1343 | Middle High German |
| mina1268 | Minangkabau |
| mira1254 | Miraña |
| misk1235 | Mískito |
| mofu1248 | Mofu-Gudur |
| moha1258 | Mohawk |
| molo1266 | Moloko |
| mopa1243 | Mopán Maya |
| moro1292 | Moroccan Arabic |
| mwan1247 | Mwani |
| nalc1240 | Nalca |
| ndem1249 | Ndemli |
| ndut1239 | Ndut-Falor |
| nezp1238 | Nez Perce |
| noma1260 | Nomaande |
| nort2641 | Northern Kurdish |
| nort2671 | North Saami |
| nort3139 | North Levantine Arabic |
| nort3142 | Sason |
| norw1258 | Norwegian |
| nubi1253 | Nubi |
| nucl1301 | Turkish |
| nucl1302 | Georgian |
| nucl1328 | Wambaya |
| nucl1417 | Igbo |
| nucl1622 | Marind |
| nucl1630 | Barai |
| nupe1254 | Nupe-Nupe-Tako |
| nyan1308 | Nyanja |
| oksa1245 | Oksapmin |
| olde1238 | Old English |
| olde1242 | Old Egyptian |
| oldf1239 | Old French |
| oldj1239 | Old Japanese |
| oldr1238 | Old Russian |
| olds1249 | Old Spanish |
| omah1247 | Omaha-Ponca |
| oman1238 | Omani |
| onei1249 | Oneida |
| oroc1248 | Oroch |
| paam1238 | Paamese |
| papu1250 | Papuan Malay |
| paum1247 | Paumari |
| phal1254 | Palula |
| pipi1250 | Pipil |
| pnar1238 | Pnar |
| polc1243 | Polci |
| poli1260 | Polish |
| rapa1244 | Rapanui |
| rash1249 | Rashad |
| roma1327 | Romanian |
| russ1263 | Russian |
| ruul1235 | Ruuli |
| safa1245 | Safaitic |
| sans1269 | Sanskrit |
| sant1410 | Santali |
| sanz1248 | Sanzhi |
| sara1340 | Saramaccan |
| savo1255 | Savosavo |
| scot1245 | Scottish Gaelic |
| shua1254 | Shua |
| siee1239 | Sie |
| sigi1234 | Sigidi |
| sino1245 | Sino-Tibetan |
| siwi1239 | Siwi |
| skol1241 | Skolt Saami |
| sout2674 | South Saami |
| sout2789 | Central Dagaare |
| sout2969 | Southern Paiute |
| sran1240 | Sranan Tongo |
| stan1288 | Spanish |
| stan1289 | Catalan |
| stan1290 | French |
| stan1295 | German |
| suba1252 | Suba-Simbiti |
| suda1236 | Sudanese Arabic |
| surs1245 | Sursilvan-Oberland |
| swah1253 | Swahili |
| swed1254 | Swedish |
| swis1247 | Swiss German |
| taga1270 | Tagalog |
| tago1246 | Tagoi |
| taji1245 | Tajik |
| tama1365 | Tamasheq |
| tari1263 | Tarifiyt |
| taro1263 | Tarok |
| teiw1235 | Teiwa |
| teop1238 | Teop |
| tian1238 | Tianjin Mandarin |
| toab1237 | Toqabaqita |
| tobe1252 | Tobelo |
| tokp1240 | Tok Pisin |
| toto1304 | Totoli |
| tuar1240 | Tuareg |
| tuka1247 | Tukang Besi |
| udih1248 | Udihe |
| uduk1239 | Uduk |
| unaa1239 | Una |
| uppe1455 | Upper Guinea Crioulo |
| viet1252 | Vietnamese |
| vlaa1240 | Western Flemish |
| waim1252 | Waima'a |
| wapp1239 | Wappo |
| wara1294 | Komnzo |
| waya1269 | Wayana |
| weno1238 | Wobé |
| wers1238 | Wersing |
| yace1238 | Yatye |
| yagu1244 | Yagua |
| yima1243 | Yimas |
| yiwo1237 | Yiwom |
| yong1288 | Yongning Na |
| yoru1245 | Yoruba |
| yura1255 | Yurakaré |
| zand1248 | Zande |
| zena1248 | Zenaga |

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 26–34
Marseille, 24 June 2022


# Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

**Margherita Fantoli[1], Marco Passarotti[2], Francesco Mambrini[2], Giovanni Moretti[2], Paolo Ruffolo[2]**
[1]Katholieke Universiteit Leuven, [2]Università Cattolica del Sacro Cuore
[1]Oude Markt 13, 3000 Leuven, Belgium. [2]Largo Gemelli 1, 20123 Milan, Italy
margherita.fantoli@kuleuven.be,
{marco.passarotti, francesco.mambrini, giovanni.moretti}@unicatt.it, paolo.ruffolo@posteo.eu

**Abstract**

This paper describes the process of interlinking the 130 Classical Latin texts provided by an annotated corpus developed at the LASLA laboratory with the LiLa Knowledge Base, which makes linguistic resources for Latin interoperable by following the principles of the Linked Data paradigm and making reference to classes and properties of widely adopted ontologies to model the relevant information. After introducing the overall architecture of the LiLa Knowledge Base and the LASLA corpus, the paper details the phases of the process of linking the corpus with the collection of lemmas of LiLa and presents a federated query to exemplify the added value of interoperability of LASLA's texts with other resources for Latin.

**Keywords:** Linguistic Linked Open Data, Corpora, Latin

## 1. Introduction

Scholars of Latin are particularly lucky when it comes to the availability of online linguistic resources. A long tradition of computational approaches and cutting-edge digital editing projects results today in an abundance of textual and lexical resources scattered on the web. Although high-quality linguistic resources are nowadays freely accessible online, in most cases they are stored in separate silos and enhanced with layers of linguistic annotation following different criteria and tagsets.

Among the several linguistic resources today available for Latin[1], the CIRCSE Research Center in Milan[2] and the LASLA laboratory in Liège[3] (Laboratoire d'Analyse Statistique des Langues Anciennes) have developed a number of manually validated lexical resources and annotated corpora. The CIRCSE has built, among others, the Word Formation Latin (WFL) derivational lexicon (Litta and Passarotti, 2019), a set of sentiment lexicons (Sprugnoli et al., 2020b) and a few syntactically annotated corpora, including the Index Thomisticus Treebank (IT-TB) (Passarotti, 2019) and the UDante Treebank(Cecchini et al., 2020). The LASLA has produced a manually verified lemmatized and morphosyntactically annotated corpus of more than 1.5 million words mainly belonging to Classical Latin literature (see Section 3).

As mentioned, one of the limitations that currently affect linguistic resources is their sparsity and diversity for what concerns data formats, annotation guidelines and sets of tags adopted. In order to overcome such limitation, the CIRCSE Research Center has developed the LiLa Knowledge Base, with the objective of making distributed linguistic resources for Latin interact through the application of the principles of the Linked Data paradigm (see Section 2).

In their work with digital resources for Latin, LASLA and CIRCSE share a large set of common features, but also show a number of differences. Each research center is dedicated to the development of high-quality, manually created or verified linguistic resources for ancient languages. They both endeavor to comply with the high-quality standards of existing – traditional – resources, such as dictionaries. Finally, both CIRCSE and LASLA combine interest for the lexical and the morphological/syntactic information encoded in texts and words.

However, since the Sixties the LASLA has mainly focused on annotating a corpus of Classical Latin and Ancient Greek literature, and has valued consistency and continuity with respect to internal criteria more than fitting the standards de facto built by the research community working on linguistic resources (like, for instance, those adopted by the Universal Dependencies initiative[4]). Moreover, the integration of Natural Language Processing (NLP) tools into the LASLA corpora (like the tagger Collatinus[5]) was always made with reference only to the LASLA schema of annotation.

Through the LiLa Knowledge Base, instead, CIRCSE supports the web-based interoperability between lexical and textual resources for Latin according to standards widely adopted in the Linguistic Linked Open Data community. Furthermore, the resources currently interlinked in LiLa include annotated corpora (like the IT-TB) that feature texts from the Medieval era, which are outside the chronological boundaries of the LASLA collection.

[1]For an overview of the linguistic resources currently available for Latin see (Passarotti et al., 2020).

[2]https://centridiricerca.unicatt.it/circse_index.html

[3]http://web.philo.ulg.ac.be/lasla/

[4]https://universaldependencies.org/

[5]https://outils.biblissima.fr/fr/collatinus-web/

In spite of the different approaches pursued by the two centers in the past, the idea of combining the high-quality textual data annotation of LASLA with the interoperability provided by LiLa's adoption of the Linked Data paradigm appears potentially very fruitful. With its dense network of other lexical and textual resources, LiLa is indeed capable of opening new avenues of research for scholars working on Latin texts, whose everyday work is strictly bound to the possibility of collecting empirical evidence from texts from different eras, genres and places.

As a consequence, LASLA and CIRCSE have decided to join their forces to interlink LASLA's Classical Latin texts with the LiLa Knowledge Base. This paper describes how such interlinking was performed. After introducing the LiLa Knowledge Base (Section 2) and the LASLA corpus (Section 3), the paper details the process of linking the texts into LiLa (Section 4) and presents a query that can be performed on the interlinked data as a way to exemplify the added value of interoperability of LASLA's texts with other resources for Latin (Section 5).

## 2. The LiLa Knowledge Base

The "LiLa - Linking Latin" project[6] aims to reach interoperability between the wealth of existing lexical and textual resources that have been developed in the last decades for Latin. One of the main problems that LiLa intends to solve is the fact that such resources and tools are often characterized by different conceptual and structural models, which makes it difficult for them to interact with one another.

To this goal, LiLa has undertaken the creation of an open-ended Knowledge Base, following the principles of the Linked Data paradigm[7]. All content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning an HTTP Uniform Resource Identifier (URI) to each data point. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as: [a] the RDF data model, which is based on triples: (i) a predicate-property connects (ii) a subject (a resource) with (iii) its object (another resource, or a literal) (Lassila and Swick, 1998); and [b] SPARQL, a query language specifically devised for RDF data.

Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies to model the relevant information. The main ones are POWLA for corpus data (Chiarcos, 2012), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015), and Ontolex-Lemon for lexical data (Buitelaar et al., 2011; McCrae et al., 2017).

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically based (Figure 1), grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words ("tokens"), lexical resources describe properties of words (in "lexical entries"), and NLP tools process words (producing "NLP outputs")[8]. The core of the Knowledge Base is the so-called Lemma Bank,[9] a collection of about 200,000 Latin lemmas – defined as the canonical form of a lexical item, i.e. its citation form – taken from the database of the morphological analyzer LEMLAT (Passarotti, M. et al., 2020) (Passarotti et al., 2017). Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma.

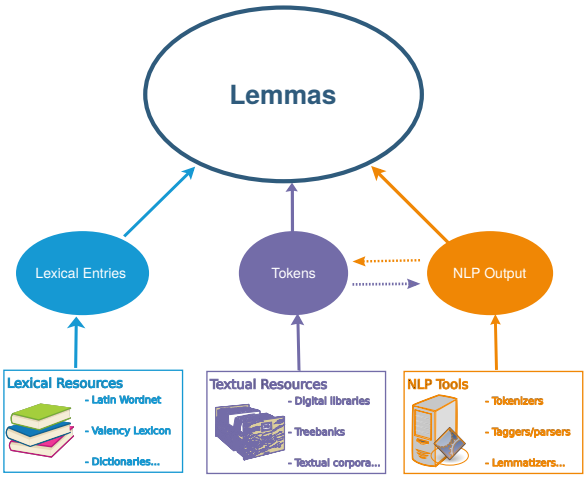


Figure 1: The architecture of LiLa

## 3. LASLA

The Latin section of the LASLA corpus contains nowadays 2,500,000 semi-automatically annotated tokens: for every token of the corpus, the automatic annotation has been manually verified by a Latin scholar. A significant part of the corpus (more than 1.7M tokens) will be soon released for free download.

The LASLA corpus features mainly Classical Latin literary texts, both poetical and in prose[10]. The earliest author in the corpus is Plautus (III-II century BC) and the latest Apuleius (II century AD; to be released soon). The data available for sharing and linked to the LiLa

[6] https://lila-erc.eu/

[7] https://www.w3.org/DesignIssues/LinkedData.html

[8] In Figure 1, the arrows going from and to the node for "NLP Output" represent the fact that tokens that are the output of a specific NLP tool (a tokenizer) become the input of further tools (like, for instance, a syntactic parser).

[9] http://lila-erc.eu/lodview/data/id/lemma/LemmaBank.

[10] http://web.philo.ulg.ac.be/lasla/textes-latins-traites/.

| Lemma | LASLA Index | French |
|---|---|---|
| cubitus | 1 | Le coude (the elbow) |
| cubitus | 2 | L'action d'être couché (the act of lying down) |

Table 1: Example of an homographic lemma from the LASLA dictionary

Knowledge Base include 130 works of 21 different authors.

The linguistic information available in the corpus consists of lemmatization, morphological tagging and an additional syntactic layer for verbs. The choice of the lemma in the LASLA corpus is based on the Forcellini dictionary (Facciolati, J. and Forcellini, E., 1771). A sequence of alphanumerical tags encodes the morphological description of the word form and some syntactic features[11]. The annotation guidelines are those provided by (Philippart de Foy, 2014).

A partial list of the lemmas included in the LASLA texts is available in the so-called LASLA dictionary[12]. The LASLA dictionary is an essential resource to address homographic lemmas, which are distinguished in the dictionary by the use of an index. In particular, the index "N" is assigned to proper nouns and "A" is assigned to adjectives derived from proper nouns (e.g. *Romanus*, "Roman, of Rome"). If one of two homographic lemmas in the LASLA dictionary is a proper noun (or an adjective derived from a proper noun), the index "N" (resp. "A") allows to disambiguate. For instance, the lemma *urbs* meaning the city of Rome is assigned index "N", whereas the lemma *urbs* meaning a generic city is not assigned any index. Similarly, in case one of two homographic lemmas is a proper noun and the other is a derived adjective (e.g. the person's name *Latinus* and the adjective *Latinus* "of the Latium"), they are assigned respectively indices "N" and "A". If none of the homographic lemmas is a proper noun, they are simply distinguished through sequential numbers.

The LASLA dictionary provides also further information, like the French translation of Latin words, to help the annotators of the corpus and its users in choosing the right homographic lemma (see Table 1).

The creation of the corpus started in 1961 with the foundation of the LASLA, and is still going on nowadays. Textual annotation is performed both via an online semi-automatic web-interface where annotators choose, for every word, the correct analysis among those proposed by the software, and through a standalone tagger with a post-correction interface (Verkerk et al., 2020).

The LASLA corpus is searchable along the different linguistic categories on the Opera Latina website[13]. In addition, the HyperbaseWeb portal, developed at the "UMR 7320 : Bases, Corpus, Langage" of the Université de Nice[14], allows to search and perform some statistical analysis of the corpus as a whole, as well as of specific thematic subsections of it (e.g., historiographical, poetic, dramatic texts).

## 4. LASLA in LiLa

This section details the process undertaken to perform the linking of the LASLA corpus in the LiLa Knowledge Base.

As said, the LiLa project adopts the assumption that the lemma, i.e. the form of a word's inflectional paradigm that is used to index a lexical entry in a dictionary or to lemmatize a corpus, is a gateway to connect the different resources. The Ontolex-Lemon ontology provides a convenient model to formalize this assumption and to express most of the relevant properties of lemmas used in standard Latin lexicography or in the practice of corpus annotation (McCrae et al., 2017).

The lemma in LiLa is defined as a subclass of the class `Form` of Ontolex,[15] which includes all forms that are potentially used (or usable) as citation forms for lexical entries or to lemmatize corpus tokens. Each of them is defined by a series of object and data properties. In particular, the Ontolex-Lemon data property 'written representation' (WR)[16] registers the different spellings or graphical variants of one lemma[17]. All forms in Ontolex-Lemon must have at least one WR; lemmas can also have a special type of representation that we define in the LiLa ontology, namely the 'prosodic representation'[18], where we register the quantity (long or short) of the form's vowels. Vowel quantity (which is generally not marked in the corpora) is often crucial to disambiguate words, such as *pŏpulus* ("people", with short *o*) and *pōpulus* ("poplar", with long *o*)[19].

[11] More precisely, the LASLA corpus indicates whether a verb belongs to a main clause or a subordinate clause. For subordinate verbs, it shows which is the subordinating element that introduces the clause.

[12] `http://cipl93.philo.ulg.ac.be/LaslaEncodingInitiative/Files/lasladic.pdf`. The list has not been updated yet with the lemmas found in the latest additions to the LASLA corpus.

[13] `http://web.philo.ulg.ac.be/lasla/opera-latina/`

[14] `http://hyperbase.unice.fr/hyperbase/`

[15] `http://www.w3.org/ns/lemon/ontolex#Form`

[16] `http://www.w3.org/ns/lemon/ontolex#writtenRep`

[17] Note however that, whenever two or more spellings entail also a change in the inflectional paradigm of a word, these are not registered as WRs of the same lemma; instead, we create as many different lemmas as we need to account for all the inflectional paradigms (Passarotti et al., 2020).

[18] `https://lila-erc.eu/lodview/ontologies/lila/prosodicRepresentation`

[19] `http://lila-erc.eu/data/id/lemma/118463`; `http://lila-erc.eu/data/id/lemma/118501`

The Part of Speech (POS) and the inflectional category are other properties that provide decisive contributions to disambiguation. For this reason, all the forms in the Lemma Bank of LiLa are annotated with tags from the Universal POS tagset (Petrov et al., 2012) and are classified according to their inflectional paradigm. The list of inflectional classes is inspired by the traditional grammars of Latin and is the one used by the morphological analyzer LEMLAT (Passarotti et al., 2017).

Like all annotated corpora, LASLA registers lemmatization with a string identifying the canonical form attached to the token. For instance, the token *uiuamus* ('let us live', 1st-person plural subjunctive present) is lemmatized with the string 'uiuo'; the same goes also for POS tagging, where the tag (e.g. VERB) is also encoded as a string. Like several other corpora, the string used for lemmatization in LASLA occasionally includes disambiguation indexes: in the case of *populus* mentioned above, LASLA uses the indexes 1 and 2 to distinguish between "the people" ("populus1") and "the poplar" ("populus2").

'Linking' a corpus to LiLa means converting the string-based annotation recorded for each corpus token into a link to a lemma in the Lemma Bank. In turn, the process entails the identification of the correct lemma corresponding to the lemmatization string registered in the corpus. The POS tag and the inflectional class attached to the tokens, when this information is available as it is the case with LASLA, are features that can help in disambiguating many of the cases where the lemma string is not sufficient. Such workflow implies three steps:

1. to align the POS tagset and the inflectional classes used in the source corpus (LASLA) and in the LiLa Lemma Bank;
2. to align the indexed strings of the homographic lemmas in the source to the correct lemma in the Lemma Bank;
3. to match the lemma and POS tag strings in the source with the WRs and the POS tags in the Lemma Bank to identify the candidates.

### 4.1. Matching POS and Inflectional Classes

Most of the LASLA POS tags, described in the documentation of the LASLA dictionary, show a 1:1 correspondence with those of LiLa, as detailed in Table 2

Although the great majority of the lemmas labeled with these POS tags in the LASLA corpus are assigned the corresponding Universal POS tag in the Lemma Bank, some exceptions do hold, due to the different criteria of application of POS tags in the two resources. For instance, the names of populations are tagged as proper nouns in LASLA, while they are assigned the POS tag for adjectives in the Lemma Bank (see Section 4.3.1 for the treatment of these exceptions).

A particularly compelling case of mismatch between the POS tags of LASLA and those of LiLA is represented by those words that are labeled as pronouns in LASLA and either as Determiners (DET) or as Pronouns (PRON) in LiLa. For instance, words in the category "Indefinite Pronoun" in LASLA can be tagged either as PRON in LiLa (e.g. *aliquis*, "somebody"), or as DET (e.g. *aliquantulus*, "small, little"). The issue is closely related to the fact that the difference between the tags PRON and DET in the Universal POS tagset is still fuzzy. The Universal tag DET is assigned to those words "that modify nouns or noun phrases and express the reference of the noun phrase in context"[20]. Pronouns, instead, are defined as terms that "substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context"[21]. However, the UD guidelines report that it is not simple to draw a line between DETs and PRONs[22].

| LASLA POS | LiLa POS |
|---|---|
| Verb | VERB |
| Adjective | ADJ |
| Adverb: generic, relative, interrogative, negative, int/neg | ADV |
| Preposition | ADP |
| Substantive | NOUN |
| Proper noun (i.e. Noun + Index N) | PROPN |
| Coordinating Conjunction | CCONJ |
| Subordinating Conjunction | SCONJ |
| Interjection | INTJ |
| Numeral | NUM |

Table 2: 1:1 mapping between LASLA and LiLa POS

In Latin, as well as in several other languages, some words can be used both as DET and PRON according to the definitions given above. For instance, the lemma *is* can be used both as PRON (meaning "that person") and as DET (e.g., *eo loco*, "that place"). The LASLA tagset conflates both categories under the label "Pronoun", which covers both usages. Such uncertainty is reflected in the documentation provided by the LASLA dictionary, where the label "Pronoun" alternates with "Pronoun/Adjective". In Lila, instead, the tag DET is assigned when both usages are possible (as with *is*), while PRON is assigned to those words that can be used only as pronouns (like *aliquis*, "somebody", which has a distinct adjectival form: *aliqui*).

To sum up, the tags "Pronoun" and "Pronoun/Adjective" of LASLA were matched with either DET or PRON of the Lemma Bank.

As for the inflectional classes, the tagsets of LASLA

[20] https://universaldependencies.org/u/pos/DET.html

[21] https://universaldependencies.org/u/pos/PRON.html

[22] "It is not always crystal clear where pronouns end and determiners start. [...] Language-specific documentation should list all determiners (it is a closed class) and point out ambiguities, if any" (https://universaldependencies.org/u/pos/DET.html).

and LiLa can be easily aligned, except for the names of Greek origin following an irregular inflection. While LiLa makes use of a separate tag for the "irregular" nouns of each declension (like, for instance, for the second declension irregular nouns), the LASLA tagset includes two broad categories "Anomalous" and "Greek declension" covering the nouns of any declension. As a consequence, in these cases, there is a many-to-many correspondence between the two tagsets. In addition, in the LASLA corpus many words are alternatively tagged as Greek declension and as "regular" declension based on the inflection of single word forms. For instance, the proper noun *Orestes* is assigned in the LASLA corpus alternatively the tag for the third declension in the case of forms that are inflected according to the paradigm used also for any other Latin word (e.g. accusative *Orestem*), and that for the Greek declension in the case where the Greek ending is used (as in the accusative form of Greek origin *Oresten*). While linking the two resources, the lemmas affected by this issue were treated manually (see Section 4.3.2).

## 4.2. Handling Homography

As said, homography is addressed in LASLA by using indices. Information that allows readers to identify the indexed lemmas is provided in the LASLA dictionary. For instance, there are two third declension neuter nouns *tempus* in Latin, respectively meaning "time" and "temple" (the side of the head near the eye), as it is recorded in the LASLA dictionary. These words are identified respectively as "tempus1" and "tempus2" in the LASLA corpus.

The work to link these strings to the correct entry in the Lemma Bank can only be made manually, by matching the lexicographic information in the LASLA dictionary with that provided by the array of lexical resources currently linked to LiLa Knowledge Base.

For instance, information that allows to disambiguate the two nouns with WR "tempus" is found in the WFL lexicon linked to LiLa (Litta et al., 2019), which assigns to each of the two lemmas *tempus* in question its respective derivatives. The information provided by WFL proves particularly helpful when two homographic lemmas formed with the same prefix are derived from two different base verbs. For instance, this is the case of the two verbs *contingo* in the Lemma Bank, both formed with prefix *cum* and respectively meaning "to happen" and "to dye". WFL informs that one verb derives from *tango* ("to touch") and the other from *tingo* ("to wet, moisten, bathe").

A second resource exploited to get the information that leads to correct disambiguation of homographic lemmas is the Latin-English dictionary (Lewis, Ch. and Short, Ch., 1879) (L&S), which is now partially linked to the LiLa Knowledge Base (Mambrini et al., 2021). The definition and translation provided by L&S can be used to distinguish homographic lemmas. One example is given by the two homographic verbs of the third conjugation *sero*. In LiLa, the link with the dictionary provides a translation of the two lemmas ("to sow, to plant", "to join and bind together"). The LASLA dictionary distinguishes them using the verbal paradigm, i.e. by indicating that the perfect indicative is *serui* for one lemma ("sero2"), and *seui* for the other ("sero3"). In total, 2,118 LASLA homographic lemmas were linked manually to the LiLa Lemma Bank by exploiting the linguistic information found in the two resources.

| Type of Match | No of Lemmas |
|---|---|
| 1:1 | 19,543 |
| 1:0 | 3,369 |
| 1:N | 932 |
| TOTAL | 23,844 |

Table 3: Number of lemmas per type of match (LASLA to LiLa)

## 4.3. Linking LASLA to the Lemma Bank

Once that the POS tags used by LASLA and LiLa were aligned and the homographic lemmas were manually matched, we proceeded to link all the other, non-homographic lemmas of LASLA to those of the LiLa Lemma Bank. The linking was based on: [a] the form of the lemma from LASLA and the value(s) of the Ontolex-Lemon data property 'written representation' from LiLa, and [b] their POS. The results of the match are shown in Table 3.

The one-to-one matches were considered validated, as one LASLA lemma matches both the form and the POS of exactly one LiLa lemma. The steps taken to perform the linking of the one-to-zero and the one-to-many matches are described in the following Sections.

### 4.3.1. One-to-zero Matches

First we considered the 3,369 LASLA lemmas where no match for the tuple $(form, POS)$ was found with the $(WR, POS)$ tuples of the LiLa Lemma Bank.

A relevant source of mismatch was the fluctuating distinction between nouns and proper nouns in the two resources. For this reason, we decided to conflate the two categories. After conflation, we were able to match 298 LASLA lemmas to exactly one LiLa lemma, while 25 lemmas showed a one-to-many correspondence and 3,046 lemmas still remained unmatched. For instance the lemma *babylonicum*, "textiles from Babylonia", originally tagged as proper noun in LASLA and noun in LiLa, was matched correctly after conflation.

Out of the 25 one-to-many matches, 14 were once again disambiguated automatically on the basis of their inflectional class. For instance, the third declension neuter noun *bacchanalia* of LASLA matched with two neuter proper nouns *bacchanalia* in the Lemma Bank, respectively of the third and of the second declension[23].

[23] `http://lila-erc.eu/data/id/lemma/405`; `http://lila-erc.eu/data/id/lemma/404`

Based on the correspondence between the tagsets for inflectional classes used by the two resources, the match with the latter was discarded.
For the remaining 11 lemmas, it was necessary either to proceed with manual disambiguation or to add the missing lemmas in the Lemma Bank. The former was the case of e.g. the proper noun *annus* of LASLA, which matched with the two nouns *annus* in LiLa, one meaning "year" and the other, more commonly spelled *anus*, meaning "posteriors"[24].
To handle the remaining one-to-zero 3,046 lemmas, we removed the constraint on the POS, thus extracting the lemmas that matched exclusively on the level of LASLA form and LiLa WR. As a result, 1,031 lemmas were matched automatically with exactly one LiLa lemma and were manually validated. 59 lemmas showing a one-to-many match were manually disambiguated. For instance, the LASLA adverb *attamen* ("but yet") corresponds to the subordinating conjunction *attamen* in the Lemma Bank and not to the noun *attamen*[25], which is a Late Latin term meaning "impurity".
Finally, the 1,956 lemmas still remaining were the ones showing no match between the form of the lemma in LASLA and a WR in the Lemma Bank. These cases were tackled by enriching the LiLa Lemma Bank with the missing lemmas (mostly, proper nouns).

#### 4.3.2. One-to-many Matches

This category includes 932 lemmas of LASLA that yield a positive match with more than one lemma in the Lemma Bank, based on the WR and the POS. For instance, the verb *alleuo* ("to lift up") in LASLA can be paired with two verbs with WR *alleuo* in LiLa (respectively meaning "to lift up" and "to make smooth"[26]).
By adding the constraint of inflectional class, we improved the rate of 1:1 matches by 364. 460 still matched multiple lemmas, while 108 lemmas resulted in an empty match based on the new constraints. This set of no-matches is mostly caused by the problematic mapping of the Greek declension. These cases have been solved by manually validating the link to the lemma with the correct inflection class in LiLa.
The 460 remaining multiple matches are mainly due to two reasons. First, in several cases a lemma in LASLA was linked to two or more lemmas that are connected via the the symmetric property 'lemma variant', defined in the LiLa ontology.[27]. The property is used to connect forms of the same lexical item that fill different cells of the inflectional paradigm and can both be used alternatively as lemmas for that item (Passarotti et al., 2020). For instance, the LASLA lemma *specus* ("cave") matches both with the LiLa masculine/feminine lemma and with its neuter lemma variant[28]. As the use of the 'lemma variant' property makes the two forms practically equivalent, this case is not problematic.
The second source of ambiguous matches is the diachronic range covered by the LiLa's Lemma Bank. The LASLA corpus features Classical Latin texts only, whereas the Lemma Bank is built also over Late and Medieval Latin lexical resources, which might contain lemmas with the same POS and inflectional class of a Classical Latin lemma, but with different meaning. One example is given by the noun *conditor*: LASLA has only the Classical Latin lemma ("founder", from the verb *condo*), whereas LiLa includes also the Late Latin lemma ("the seasoner", from the verb *condio*), thus resulting in a case of homography. These matches were manually disambiguated.
Finally, for 4 lemmas showing a multiple match, we performed a manual disambiguation on the level of their single tokens[29]. In these cases, the LASLA corpus contains a single lemma for two LiLa lemmas that are homographic and cannot be distinguished on the basis neither of linguistic features (like the POS or the inflectional class), nor of formal features (like the plural vs singular form). Given that the distinction between the two LiLa lemmas is exclusively semantic, only the meaning of their single occurrences in the LASLA corpus can be used to link to the correct LiLa lemma.

### 4.4. Results

The publicly shared part of the Latin section of the LASLA corpus is now entirely linked to the LiLa Knowledge Base. In total, 1,738,435 tokens from LASLA are now connected to the LiLa Knowledge Base via the lemmas of the Lemma Bank. Manual linking by one expert annotator was necessary for 3,791 lemmas, for a total of ca. 50 hours of work. Figures 2 and 3 visualize some of the information attached to tokens from the corpus.
The LASLA corpus, its texts and the tokens are modeled using the POWLA ontology (Chiarcos, 2012). Figure 2 shows an example of a document (the philosophical dialogue "Of Friendship" (*De Amicitia*) by Cicero, pink node in the middle of the figure), i.e. one of the 130 works in the corpus. The document is subdivided in a series of structural units, that are grouped in three layers. The sentence and citation layers (light blue node on the top and bottom left) aggregate respectively all the sentences and the structural units (in this case, the numbered paragraphs) that make up the text.

[24] http://lila-erc.eu/data/id/lemma/89129; http://lila-erc.eu/data/id/lemma/89365
[25] http://lila-erc.eu/data/id/lemma/91078; http://lila-erc.eu/data/id/lemma/32914
[26] https://lila-erc.eu/data/id/lemma/88348; https://lila-erc.eu/data/id/lemma/88385
[27] https://lila-erc.eu/lodview/ontologies/lila/lemmaVariant.
[28] http://lila-erc.eu/data/id/lemma/125318; http://lila-erc.eu/data/id/lemma/125319
[29] *clauiger* ("club-bearing"), *insomnium* ("dream"), *myrrheus* ("of myrrh"), *propola* ("forestaller").

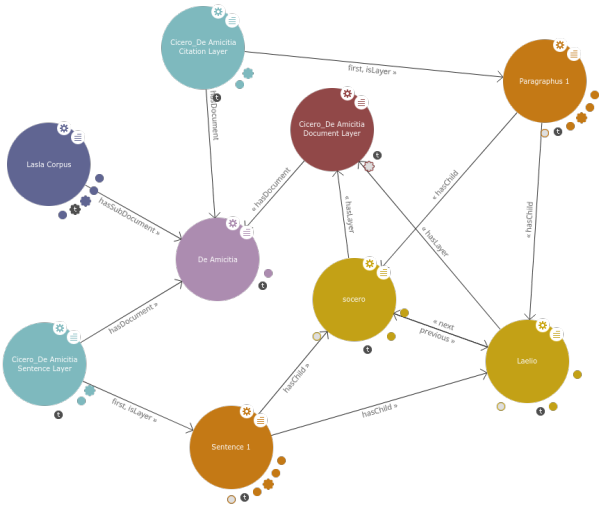


Figure 2: A LASLA token in LiLa: structural relations

The document layer (red node) links directly to the tokens. Two of them from the first paragraph, *socero* ("father in law") and the word that immediately precedes it (*Laelio*, "Laelius") are reproduced in the figure (yellow nodes).

Figure 3 represents some of the lexical information that the network of resources linked to LiLa allows to uncover for the same token *socero* from Cicero's *De Amicitia*. The lemma from the Lemma Bank is represented in the center (purple node), with the three WRs attested for the form (*socer*, *socrus*, *socerus*). The lemma is used as canonical form for entries in a series of lexical resources, two of which are reported in the Figure. In the bottom part, an entry in an etymological dictionary (Mambrini and Passarotti, 2020) accounts for the hypothetical origin of the word from the reconstructed Proto-Italic root **swekuro-* (de Vaan, 2008)[30]. The entry in the Latin-English dictionary L&S documents the two senses of the word, i.e. the main one ("father in law") and the transferred sense ("own child's father in law", properly *consocer*). Figure 3 visualizes the latter. On the top part of the Figure, we also see two lemmas linked to the same derivational family of *socer*: *consocrus* (a variant of *consocer*), composed with the preposition *cum* "with" (lit. "one who is father in law with"), and *prosocer* ("wife's grandfather").

## 5. Querying LASLA in LiLa

A query interface for the Lemma Bank can be accessed at `https://lila-erc.eu/query/`. Lemmas can be searched by string of characters (also using regular expressions), POS, affix, lexical base, inflectional category, and gender (for nouns). Results are provided both as data sheet and in a network-like graphical visualization. The entries in lexical resources and the tokens in corpora linked to each lemma in LiLa are reported as well[31].

A SPARQL endpoint is also available at `https://lila-erc.eu/sparql/`, to query the contents of all the textual and lexical resources currently interlinked in the Knowledge Base. A number of precompiled queries is provided, including a query that counts the number of occurrences of those tokens from the LASLA corpus that are linked to a lemma of the Lemma Bank connected to a lexical entry provided with a negative polarity in the *Latin Affectus* lexicon (Sprugnoli et al., 2020a)[32].

As it is to be expected, this query returns words like: *hostis* "enemy" (2,109 occurrences), *mors* "death" (1,555), *periculum* "danger" (1,299), *grauis* "heavy, grievous" (1,232), or *malum* "evil" (1,220).

If we disaggregate the results by the different documents, we can rank the texts by the relative frequency of negative terms. Table 4 reports the 6 highest results, excluding fragmentary works that are too short to be meaningful. Not surprisingly, 4 out 6 slots are taken by tragedies of Seneca, the only tragic poet represented in the corpus. It is very interesting to note, however, that the other two works in the table, the moral treatises "On Wrath" (*De Ira*) and "On the Firmness of the Wise" (*De Constantia Sapientium*), are also authored by Seneca. The presence of the former text is certainly accounted for by the high occurrence of the word referring to the subject (*ira* "wrath", 242 occurrences, 1.07 x100 words). The latter treatise, on the other hand, is concerned with the ability of the Stoic philosopher to withstand abuse and suffering.

The first text not written by Seneca to figure in the list is only found at rank number 12; the work is "The Conspiracy of Catilina" (*De coniuratione Catilinae*, 5.45 negative words x100) by the historian Sallust, an essay dedicated to an infamous political plot that is certainly lavish of many sinister details about the protagonists and the moral decadence of the Roman society.

| Work | Author | Tokens | Neg x100 |
|---|---|---|---|
| Medea | Seneca | 5,700 | 7.4 |
| Phaedra | Seneca | 7,281 | 7.27 |
| Phoenissae | Seneca | 4,182 | 7.17 |
| De Ira | Seneca | 22,541 | 7.02 |
| Thyestes | Seneca | 6,321 | 6.64 |
| De Constantia | Seneca | 5,323 | 6.63 |

Table 4: Works with highest x100-frequency of negative words in LASLA

[30] As usual in historical linguistics, the star is used to mark unattested forms reconstructed with the help of the comparative method.

[31] The Turtle files of the resources interlinked in LiLa are available at `https://github.com/CIRCSE`.

[32] The pre-made queries can also be downloaded at `https://github.com/CIRCSE/SPARQL-queries`.

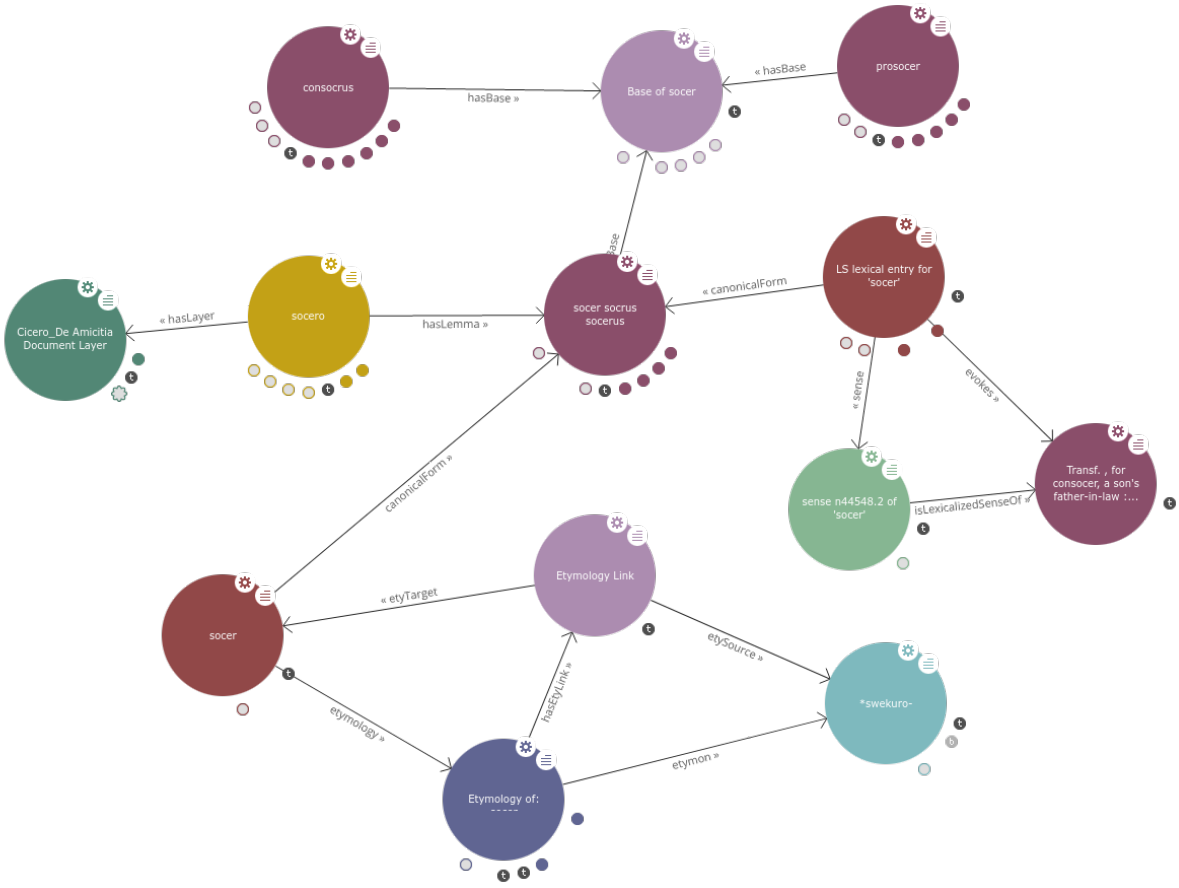


Figure 3: A LASLA token in LiLa: lexical information

## 6. Conclusion

The (soon) freely available portion of the LASLA corpus (ca. 1.7M tokens) is now linked to the LiLa Knowledge Base[33]. This result is a major achievement for both projects. As for LASLA, making its texts interoperable with other (kinds of) linguistic resources extends the degree of granularity of information extraction from the corpus, by focusing on words with specific lexical properties, and supports comparative research, by collecting relevant occurrences of words from corpora of different era and genre. As for LiLa, beyond enlarging the number and diversity of texts interlinked, the inclusion of the LASLA corpus will favor the use and dissemination of the Knowledge Base among Classicists, who are used to consider LASLA as one of the reference corpora of their community. Indeed, one of the objectives of the "LiLa - Linking Latin" project is to make digital linguistic resources and NLP tools finally become part of the everyday work of Classicists. Such objective can be achieved also by leading to a new level of accessibility those resources that are already well known in that community, to show how much more helpful they can become once made interoperable.

Not only is LiLa based on the principles of the Linked Data paradigm, but it reflects as much as possible the common grounds of the Linguistic Linked Open Data community. Such openness of the (meta)data of the resources interlinked through LiLa impacts the community of Classicists in that the entire process followed to collect the empirical evidence supporting their claims is made repeteable, replicable and reproducible (Cohen-Boulakia et al., 2017). Given the highly empirically-based nature of any linguistic, literary, or philological research on ancient languages, such an aspect is a very valuable added value, which is supposed to impact heavily how research in Classics is performed and published.

An open challenge to the community is represented by the management of the flow-back of information from the LiLa Knowledge Base to resources developed outside the Linked Data paradigm: for example, LASLA users would benefit from the integration of LiLa URIs in the current LASLA database and search interfaces.

## 7. Acknowledgements

The "LiLa - Linking Latin" project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

[33] https://lila-erc.eu/lodview/data/corpora/Lasla/id/corpus

## 8. Bibliographical References


Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., and Declerck, T. (2011). Ontology Lexicalization: The *lemon* Perspective. In *Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 33–36.

Cecchini, F. M., Sprugnoli, R., Moretti, G., and Passarotti, M. (2020). UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7, Bologna. CEUR-WS.org.

Chiarcos, C. and Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Chiarcos, C. (2012). POWLA: Modeling Linguistic Corpora in OWL/DL. In Elena Simperl, et al., editors, *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 225–239, Berlin, Heidelberg. Springer.

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Le Bras, Y., Lemoine, F., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75:284–298.

de Vaan, M. (2008). *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Amsterdam.

Lassila, O. and Swick, R. R. (1998). Resource Description Framework (RDF) Model and Syntax Specification.

Litta, E. and Passarotti, M. (2019). (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, et al., editors, *Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, 12. Interrogable online at `http://wfl.marginalia.it/`.

Litta, E., Passarotti, M., and Mambrini, F. (2019). The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Mambrini, F. and Passarotti, M. (2020). Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, pages 20–28, Paris. European Language Resources Association (ELRA).

Mambrini, F., Litta, E., Passarotti, M., and Ruffolo, P. (2021). Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy, January 26-28, 2022*, Milan, Italy, December. CEUR.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, pages 587–597.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 299–319. De Gruyter, Berlin.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Philippart de Foy, C. (2014). Lasla - nouveau manuel de lemmatisation du latin.

Sprugnoli, R., Moretti, G., and Passarotti, M. (2020a). Towards the modeling of polarity in a Latin knowledge base. In Alessandro Adamou, et al., editors, *WHiSe 2020 Workshop on Humanities in the Semantic Web 2020*, pages 59–70, Heraklion, Greece. CEUR.

Sprugnoli, R., Passarotti, M., Corbetta, D., and Peverelli, A. (2020b). Odi et amo. creating, evaluating and extending sentiment lexicons for latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3078–3086.

Verkerk, P., Ouvrard, Y., Fantoli, M., and Longrée, D. (2020). L.A.S.L.A. and Collatinus: a convergence in lexica. *SSL*, 1(LVIII):95–120.


## 9. Language Resource References


Facciolati, J. and Forcellini, E. (1771). *Totius Latinitatis lexicon*. Patavii: typis Seminarii.

Lewis, Ch. and Short, Ch. (1879). *A Latin Dictionary*. Clarendon Press.

Passarotti, M. et al. (2020). *LEMLAT 3.0*. CIRCSE, Università Cattolica del Sacro Cuore, and Zenodo, DOI:10.5281/zenodo.1492134, v. 3.0.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 35–44
Marseille, 24 June 2022


# Use Case: Romanian Language Resources in the LOD Paradigm

**Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Andrei-Marius Avram, Maria Mitrofan**
Romanian Academy Research Institute for Artificial Intelligence
13 Calea 13 Septembrie, Bucharest, Romania
{vergi,elena,vasile,andrei.avram,maria}@racai.ro

**Abstract**

In this paper, we report on (i) the conversion of Romanian language resources to the Linked Open Data specifications and requirements, on (ii) their publication and (iii) interlinking with other language resources (for Romanian or for other languages). The pool of converted resources is made up of the Romanian Wordnet, the morphosyntactic and phonemic lexicon RoLEX, four treebanks, one for the general language (the Romanian Reference Treebank) and others for specialised domains (SiMoNERo for medicine, LegalNERo for the legal domain, PARSEME-Ro for verbal multiword expressions), frequency information on lemmas and tokens and word embeddings as extracted from the reference corpus for contemporary Romanian (CoRoLa) and a bi-modal (text and speech) corpus. We also present the limitations coming from the representation of the resources in Linked Data format. The metadata of LOD resources have been published in the LOD Cloud. The resources are available for download on our website and a SPARQL endpoint is also available for querying them.

**Keywords:** Romanian, OntoLex, LLOD Cloud

## 1. Introduction

According to the recently collected data[1] on the existence and availability of resources and technologies for different languages, Romanian is a language with fragmentary technological support[2]. This means a presence of between 3% and 10% in the catalogue of language resources available in the European Language Grid[3]. The vast majority of these resources are available in various formats adopted according to the requirements or needs of the projects in which they were created: e.g., the Romanian Wordnet was created in XML format (Tufiș et al., 2004a), the corpus annotated with verbal multiword expressions was released in CUPT format (Ramisch et al., 2018), etc.

In the last few years, within the Natural Language Processing group of the Romanian Academy Research Institute for Artificial Intelligence[4], steps have been taken to convert the resources developed herein throughout time (Tufiș, 2022) to the specifications of the Linked Open Data (LOD) paradigm, so as to ensure them the benefits derived from this: higher visibility, accessibility, contextualization (by linking them to other resources) and, eventually, further increase of the technological development of Romanian. The most important decision was to make them open. They have been made freely available and enriched with metadata, which have been added to the Linked Open Data Cloud[5] (LOD Cloud).

The representation of Romanian in the Linguistics LOD Cloud (LLOD Cloud) is not only due to our contribution. Four resources (all created in a multilingual context) had metadata already recorded in the LLOD Cloud when we started our endeavour. They were: EuroVoc[6], Universal Dependencies[7] Treebank Romanian, Multext-East[8] and Romanian WordNet (as part of Open Multilingual WordNet[9]). We present the resources we converted (providing a brief description of their content and of their representation in the LOD paradigm) in Section 2. Their publication methods are enumerated in Section 3. A presentation of the way in which they are interlinked among themselves or with other resources is available in Section 4. Some potential use cases are designed in Section 5, before concluding the paper.

## 2. Romanian Language Resources converted to LOD

During the last year we have added metadata of 8 more resources[10] for Romanian in the LLOD Cloud: the whole Romanian Wordnet (bigger than the one available in Open Multilingual WordNet), the morpho-phonemic lexicon RoLEX, four treebanks (the Romanian Reference Treebank RRT, the medical treebank SiMoNERo, the law treebank LegalNERo, the treebank annotated with verbal multiword expressions PARSEME-Ro), lemmas and tokens frequencies and word embeddings extracted from the corpus of contemporary Romanian (CoRoLa), and a bimodal (written and oral) corpus (RTASC). These have been chosen to be converted to the LOD specification because they are the main resources our group have created throughout time, they are still relevant in the international linguistic context and are, thus, worth being made more visible and accessible. We describe each resource below and present the decisions made about their conversion to the LOD specifications, as well as the limitations of this representation.

While there are more formats to represent Linked Data (N-Triples, RDF Turtle, JSON-LD and RDF/XML among the most common), we chose RDF Turtle[11] for its advantage of human readability, due to the possibility of defining prefixes in the beginning of the file. For syntactic

[1] Within the project European Language Equality (ELE) (https://european-language-equality.eu/) the LT support of the 24 official and 32 additional EU-languages as well as 33 endangered minority languages has been evaluated and a report on the state of the art in language technology and language-centric AI has been released for each language.

[2] https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_29__Language_Report_Romanian_.pdf

[3] https://live.european-language-grid.eu/

[4] www.racai.ro

[5] https://lod-cloud.net/

[6] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

[7] https://universaldependencies.org/

[8] http://nl.ijs.si/ME/Vault/V4/

[9] http://compling.hss.ntu.edu.sg/omw/

[10] https://lod-cloud.net/datasets?search=racai

[11] https://www.w3.org/TR/turtle/

validation of the .ttl files, we used an open source tool[12]. Their semantic validation was done for the original format: e.g., for RoWN an in-house tool was developed (Tufiș et al., 2004a) for detecting semantic incorrectness of the resource. A semantic evaluation of the LD format of the resources is yet to be made, according to acknowledged criteria (Zaveri et al, 2015).

The URIs were created individually for each resource, using a resource-specific format. Of course, when different objects were reused between resources, potentially allowing for interlinking, the same URI was employed. The URIs do not correspond to real-world URLs, and thus individual objects cannot be accessed via the Internet. The resources are intended to be used either as a complete download of the entire file or through the provided SPARQL endpoints.

## 2.1 The Romanian Wordnet

### 2.1.1 Description of the Romanian Wordnet

The lexical ontology for Romanian, i.e. the Romanian Wordnet (RoWN, Tufiș and Barbu Mititelu, 2014), was developed by translating the Princeton WordNet (PWN, Miller, 1995; Fellbaum, 1998) synsets and transferring the relations between equivalent synsets. It contains 56,591 synsets in which 53,092 (noun, verb, adjective or adverb) literals occur. In PWN *semantic relations* are established between synsets, which are lexicalizations of concepts; thus this type of relations has cross-lingual validity (to a certain extent), which offers the grounds for transferring them between equivalent synsets in networks for other languages. RoWN was created and maintained in an XML format, with a DTD specific to the BalkaNet project principles[13].

### 2.1.2 Conversion of RoWN to LOD specifications

We used the following *OntoLex-lemon* classes and properties to represent RoWN, as *lemon*[14], the lexicon model for ontologies developed by the Ontology-Lexica (OntoLex) community group, is the recommended standard for wordnets[15]:

1. *ontolex: LexicalEntry* (mono- or multi-word), described by a lemma (*ontolex:CanonicalForm*), a part of speech (*wn:partOfSpeech*) and a reference to a *LexicalSense* object.

2. *ontolex: LexicalSense* (represents one of the meanings of the lexical entry and contains a reference to a synset in the network, encoded with the *ontolex: reference* property);

3. *ontolex:LexicalConcept*: encodes the synset referenced by the *LexicalSense*, described by a definition, an *ILI* (an id in the the collaborative interlingual index of concept for wordnets[16] (Bond et al., 2016)) and a part-of-speech (all defined by the wordnet specialised vocabulary wn).

The following listing shows an entry in the original XML format, corresponding to the lemma "pom" (En. "tree"). For simplicity, only one semantic relation (out of 26 in which this synset is involved) is presented here.

```
<SYNSET>
   <ID>rown-12651821-n</ID>
    <POS>n</POS>
     <SYNONYM>
          <LITERAL>pom<SENSE>1</SENSE></LITERAL>
    </SYNONYM>
   <DEF>nume generic pentru orice arbore sălbatic sau cultivat,
care produce fructe</DEF>
 <ILR>ENG30-13109733-n<TYPE>hypernym</TYPE></ILR>
   ...
</SYNSET>
```

**Listing 1**. A synset from RoWN in its original XML form

Listing 2 illustrates the LLOD encoding model for the same lemma "pom". The LITERAL XML-attribute becomes the main class LexicalEntry in the *lemon* model, the SENSE attribute of the LITERAL is encoded as an ontolex:LexicalSense, while the synset, which was the basic entry in the XML file, is now just a reference property of the LexicalSense. The reference r*own:12651821-n* is described further in the file by the wn:partOfSpeech *n,* the wn:ili *103362* and a definition in Romanian. The LexicalEntry ID is generated by concatenating the literal, the pos and the synset number to differentiate it from the same literal in other possible synsets. The *lemon* variation and translation module *vartrans* is then used to express synset relations, by encoding the source, the target and the relation category (hypernym in our example).

```
rown:pom-n-12651821 a ontolex:LexicalEntry ;
        ontolex:canonicalForm [
                ontolex:writtenRep "pom"@ro ] ;
        wn:partOfSpeech wn:n ;
        ontolex:Sense rown:pom-n-12651821-1 .
rown:pom-n-12651821-1 a ontolex:LexicalSense ;
        ontolex:reference rown:12651821-n .
...
rown:12651821-n a ontolex:LexicalConcept ;
        wn:partOfSpeech wn:n ;
        wn:ili ili:i103362 ;
        wn:definition [
                rdf:value "nume generic pentru orice arbore
sălbatic sau cultivat, care produce fructe"@ro] .
…

vartrans:source  <https://www.racai.ro/p/llod/resources/rown-
3.0.ttl/12651821-n> ;
vartrans:category wn:hypernym ;
vartrans:target  <https://www.racai.ro/p/llod/resources/rown-
3.0.ttl/13109733-n> .
```

**Listing 2**. The same synset from RoWN in Turtle form

Some information from RoWN cannot be represented in LD format yet: (i) Balkan-specific concepts and (ii) a different treatment of PWN lexical relations. (i) During the BalkaNet project[17] (Tufiș et al., 2004b), some synsets lexicalizing Balkan-specific concepts have been implemented. They were added as hyponyms of synsets already translated from PWN. These new synsets have a specific identifier (BILI). Lacking a correspondent in the interlingual index (so they cannot be assigned an *ili*), these BILI synsets are not included in the LD format of RoWN. (ii) As far as the *lexical relations*[18] (antonymy and derivational relations) are concerned, their semantic

[12] https://github.com/IDLabResearch/TurtleValidator

[13] http://www.dblab.upatras.gr/balkanet/deliverables/d.2.pdf

[14] https://lemon-model.net

[15] http://bpmlod.github.io/report/WordNets/index.html

[16] https://github.com/globalwordnet/cili

[17] http://www.dblab.upatras.gr/balkanet/

[18] PWN distinguishes between semantic relations, holding between senses (so between synsets), and lexical relations, holding between word forms.

component has also been acknowledged in the development of RoWN: for example, the antonymy relation between two word forms can be safely exported to the synsets to which these two words belong as a conceptual opposition. This makes such relations transferable between equivalent synsets in other wordnets.

These two characteristics are relevant with respect to the LD representation of RoWN (as compared to that of other wordnets). Future work will seek to offer solutions for these cases.

## 2.2 RoLEX

### 2.2.1 Description of RoLEX

RoLEX is the most extensively validated phonemic lexicon available for the Romanian language. It contains 330,886 entries and it was initially developed in tabular format[19], with 6 columns: the lexical form, its lemma, a morphosyntactic description in the MSD[20] format, the syllabification with syllable boundaries marked by (.), the stress marked by placing (‘) in front of the stressed vowel, and the phonemic transcription using an extended version of Speech Assessment Methods Phonetic Alphabet (SAMPA)[21] for Romanian (see Listing 3).

It is an organically developed resource, based on the textual component of speech corpora collected in the ReTeRom project[22]. It is built on a list of lexical items extracted from a corpus of Romanian Wikipedia texts that was read by volunteers, news, interviews, talk shows, spontaneous speech, read fairy tales and novels. The morphosyntactic and lemma information associated with each item in the list was taken from a Romanian lexicon[23] that our team maintains, while syllabification, stress and phonemic transcription information was partially extracted from existing resources – RoSyllabiDict (Barbu, 2008) and MaRePhor (Toma et al., 2017) – and partially automatically generated using Stan et al.’s (2011) tool. On top of aggregating all this data, we applied a thorough curation, using techniques of automatic validation and correction, but also manual correction of automatically identified errors. The information about lemma and morphosyntactic descriptions was entirely manually validated and corrected in the original extensive Romanian lexicon TBL, while information about syllabification, stress and phonemic transcription has gone through a process of (1) automatic selection of entries with a high syllabification error probability (entries with abnormally long syllables, with syllables containing two vocals, with syllables containing one vowel followed by one or two semi-vowels, with syllables containing some specific letter groups “ce/ci/ge/gi/che/chi/ghe/ghi”, entries that contain proper nouns, etc.) followed by manual or automatic correction; (2) automatic selection of entries containing homographs that are not homophones: in these cases, stress can be correctly marked by taking into account information like lemma or the morphosyntactic description and its correct marking can consequently impact syllabification and phonemic transcription; (3) implementation of phonemic transcription rules based on correct syllabification and stress marking. All the corrections were made by two linguists; the work was distributed and inter-annotator agreement was not pursued, since correction tasks were rather trivial.

### 2.2.2 Conversion of RoLEX to LOD specifications

*OntoLex-lemon* was the main frame of representation for the RoLEX conversion to LLOD specification. *Ontolex:LexicalEntry* was used to encode the unique lemmas in RoLEX. The inflectional paradigm of the lemma is encoded as a list of *ontolex:lexicalForm* and the set of senses traditionally associated with a LexicalEntry is absent, but compensated by interlinking RoLEX lexical entries with RoWN synsets via *ili*. We exemplify with the six inflected forms of the lemma “pom” (see Listing 3). They were converted in the *LexicalEntry* :lex_pom that has, among other properties, 6 ontolex:LexicalForm associated with it (see Listing 4).

| | | | | | |
|---|---|---|---|---|---|
| *pom* | *=* | *Ncms-n* | *pom* | *pom* | *p o m* |
| *pomul* | *pom* | *Ncmsry* | *po.mul* | *p'omul* | *p o m u l* |
| *pomului* | *pom* | *Ncmsoy* | *po.mu.lui* | *p'omului* | *p o m u l u j* |
| *pomi* | *pom* | *Ncmp-n* | *pomi* | *pomi* | *p o m i_0* |
| *pomii* | *pom* | *Ncmpry* | *po.mii* | *p'omii* | *p o m i j* |
| *pomilor* | *pom* | *Ncmpoy* | *po.mi.lor* | *p'omilor* | *p o m i l o r* |

**Listing 3**. Example of 6 entries associated to lemma “pom” from RoLEX in tabular format

```
:lex_pom_n a ontolex:LexicalEntry;
        rdfs:label "pom_n"@ro;
        ontolex:canonicalForm :form_pom_n;
        wn:partOfSpeech wn:n;
        wn:ili ili:i103362;
        wn:ili ili:i105570;
ontolex:lexicalForm :form_pom_n_noun_ind_masc_sing;
ontolex:lexicalForm :form_pom_n_noun_ind_masc_plur;
ontolex:lexicalForm:form_pom_n_noun_acc_nom_def_masc_plur;
ontolex:lexicalForm:form_pom_n_noun_dat_gen_def_masc_plur;
ontolex:lexicalForm:form_pom_n_noun_acc_nom_def_masc_sing;
ontolex:lexicalForm:form_pom_n_noun_dat_gen_def_masc_sing .
```

**Listing 4**. Example of an ontolex:LexicalEntry from RoLEX

We used *ontolex:canonicalForm* property to specify the canonical form (i.e. lemma) associated with the specific lexical entry: :form_pom_n. The canonical form is later described through the property *ontolex:writtenRep* as "pom"@ro. Although encoding the same information as ontolex:canonicalForm, *rdfs:label* is also used, as OntoLex recommends it for compatibility with RDFS-based representation systems. The *partOfSpeech* property from the *wn*[24] vocabulary encodes the part-of-speech information “n” (noun).

To generate lexicalForm labels, we used the MSD tag uniquely associated with each form and expanded it in a Universal Dependencies (UD) feature list[25]. For example, the label form_pom_n_noun_acc_nom_def_masc_plur

[19]RoLEX in tabular format is freely available for download at https://www.racai.ro/p/reterom/results.html.

[20]https://github.com/clarinsi/mte-msd/blob/master/tables/msd-canon-ro.tbl

[21]https://www.phon.ucl.ac.uk/home/sampa/romanian.htm

[22] https://www.racai.ro/p/reterom/

[23]https://raw.githubusercontent.com/racai-ai/Rodna/master/data/resources/tbl.wordform.ro,

[24] http://globalwordnet.github.io/schemas/wn

[25] https://universaldependencies.org/u/feat/

was created starting from the "Ncmpry" MSD, encoded by the *conll:POS* property as can be seen in Listing 5:

```
:form_pom_n_noun_acc_nom_def_masc_plur a ontolex:Form;
        ontolex:writtenRep "pomii"@ro;
        conll:POS "Ncmpry";
        ontolex:writtenRep "po.mii"@syl;
        ontolex:writtenRep "p'omii"@stress;
        ontolex:phoneticRep "p o m i j"@ro-RO-sampa .
```

**Listing 5**. Description of an ontolex:Form in RoLEX

Finally, the OntoLex properties *writtenRep* and *phoneticRep* are used to represent the lexicalisation and the phonemic transcription associated with the form. In the absence of specific OntoLex properties, we also used writtenRep to encode syllabification and stress.

## 2.3 Treebanks

### 2.3.1 Description of the Romanian Treebanks

**RoRefTrees** (RRT) (Barbu Mititelu, 2018) is the reference treebank for standard Romanian. It has 9,523 sentences containing 218k tokens. It covers a variety of genres (legal, news, fiction, medical, science, academic writing, Wikipedia) and reflects the contemporary language. The RRT treebank is distributed in the UD releases and is freely available[26].

**SiMoNERo** (Barbu Mititelu and Mitrofan, 2020) is the medical treebank for the Romanian language and has three levels of annotation: gold standard morphological annotation, hand validated annotation with medical named entities and syntactic annotation in compliance with UD specifications. SiMoNERo contains texts from the BioRo corpus (Mitrofan and Tufiș, 2018) belonging mainly to three main medical domains: diabetes, cardiology and endocrinology. All the texts are extracted from three types of documents: medical scientific journal articles, scientific medical books and medical blog posts. Currently, the treebank contains 4,681 sentences distributed in 146k tokens. SiMoNERo also has 14,133 medical named entities distributed in the four types: anatomical parts (ANAT), chemicals (CHEM), disorders (DISO) and procedures (PROC). The TTL tool (Ion, 2007) was used for tokenization, lemmatization, part-of-speech tagging and dependency parsing, while the dependency parsing level was added using NLP-Cube (Boroș et. al, 2018). The treebank is also distributed in the UD releases and is freely available for download[27].

**LegalNERo** (Păiș et. al, 2021a; Păiș and Mitrofan, 2021) is the first legal treebank for the Romanian language and it also has three annotation levels: morphological, syntactic and named entities annotation. The LegalNERo corpus contains a total of 370 documents selected from MARCELL-RO corpus (Tufiș et al., 2020) and 265k tokens. It provides gold annotations for five entity classes: organisations (ORG), locations (LOC), persons (PER), time expressions (TIME) and legal resources mentioned in legal documents (LEGAL). The UDPipe tool (Straka et al., 2016) was used for tokenization, lemmatization, part-of-speech tagging and dependency parsing. The treebank is available[28] in multiple formats, including span-based, token-based and RDF.

**PARSEME-Ro** (Barbu Mititelu et al., 2019a) is a journalistic corpus automatically morpho-syntactically annotated using UDPipe. It was further manually enriched with semantic information of the type verbal multiword expressions (VMWEs), within the PARSEME project (Savary et al., 2018). Three main types of VMWEs were annotated: light verb constructions, verbal idioms and reflexive verbs. The corpus is also freely available[29] together with the corpora annotated for other languages (Ramisch et al., 2020).

### 2.3.2 Conversion of the Romanian treebanks to LOD specifications

The Romanian treebanks were converted to the LOD specifications by using the CoNLL-U (the original format of the corpus) to an RDF graph (the target format for LOD) tool that was developed by the Applied Computational Linguistics (ACoLi) laboratory (Chiarcos et al., 2017). This tool converted the CoNLL-U files into the Turtle format.

We used the NIF[30] format to achieve interoperability between the sentences and the words found in the Romanian treebanks. Thus, we employ objects of type *nif:Sentence* to denote the id of the sentence and its text and objects of type *nif:Word* together with *conll* properties to describe the words:

- *conll:ID* - the word index
- *conll:WORD* - the original word, as it is found in the sentence
- *conll:LEMMA* - the lemma of the word form
- *conll:UPOS* - the universal part-of-speech
- *conll:POS* - the extended part-of-speech
- *conll:FEAT* - list of morphological features
- *conll:HEAD* - the head of the word in the dependency tree of the sentence
- *conll:EDGE* - the syntactic relation established with the head
- *conll:MISC* - the semantic information added to some of the treebanks, namely the medical entities in SiMoNERo, the legal ones on LegalNERo, the verbal multiword expressions in PARSEME-Ro.

We also specify the next sentence using the *nif:nextSentence* object and the next word with the *nif:nextWord* object. An example of a sentence and its first word from RRT in the LOD format is given in Listing 6.

```
# sent_id = dev-6
# text = Prin însăși natura lucrurilor era imposibil.
:rrt_dev_s5_0 nif:nextSentence :rrt_dev_s6_0 .
:rrt_dev_s6_0 a nif:Sentence;
   rdfs:comment
      "sent_id = dev-6
       text = Prin însăși natura lucrurilor era imposibil." .
:rrt_dev_s6_1 a nif:Word;
    conll:WORD "Prin";
    conll:EDGE "case";
    conll:FEAT "AdpType=Prep|Case=Acc";
    conll:HEAD :rrt_dev_s6_3;
```

[26]https://github.com/UniversalDependencies/UD_Romanian-RRT

[27]https://github.com/UniversalDependencies/UD_Romanian-SiMoNERo

[28] https://zenodo.org/record/4772095#.YkHJT3pByM8

[29]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3367

[30]https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

```
conll:ID "1";
conll:LEMMA "prin";
conll:POS "Spsa";
conll:UPOS "ADP";
nif:nextWord :rrt_dev_s6_2 .
```

**Listing 6**. Example from the RRT treebank

## 2.4 Corpus-based Frequencies and Embeddings

### 2.4.1 Description of CoRoLa

The representative corpus of contemporary Romanian (CoRoLa) (Tufiș et al., 2019) was created as a priority project of the Romanian Academy and made available to the community in 2017: its 1.2 billion words cover imaginative, journalistic, scientific, administrative, law, memoirs, and blogpost texts, assignable to 4 domains (Society, Science, Nature, Arts & Culture) and tens of subdomains thereof. Each file has associated metadata, with information about the text author, title, year of publication, text type, domain, subdomain, source, etc. The vast majority of the texts included in the corpus are cleared with respect to the Intellectual Property Right, but for some we only have the right to store and process them. Thus, the corpus is not downloadable, though it is indexed and queryable in the KorAP platform[31] (Diewald et al., 2016; Diewald et al., 2019). Besides the written component, CoRoLa also contains an oral component, made up of over 100 hours of speech, which is indexed and can be queried on a different platform[32].

### 2.4.2 Conversion of CoRoLa-based frequencies to LOD specifications

One of the ways in which the corpus can be exploited is by creating corpus-driven data, such as word and lemma frequencies. This information, initially available as simple lists, were converted to linked data specifications.

For conversion purposes, we used the OntoLex-FRAC[33] specification, designed by OntoLex for frequency, attestation and corpus information. An automatic conversion tool, written in Java, was developed, taking as input a CSV file with frequencies and producing a file in RDF Turtle format. For both word and lemma frequencies we used *frac:frequency* objects of type *frac:CorpusFrequency* to hold the actual values as integers. An example is given below. All the vocabularies used for the representation are available in Table 1 and an example is given in Listing 7.

```
:sat a ontolex:LexicalEntry;
      ontolex:canonicalForm "sat"@ro;
      frac:frequency [
        a frac:CorpusFrequency;
        rdf:value "33059"^^xsd:int;
        dct:source <http://corola.racai.ro/> ].
```

**Listing 7.** Word frequency representation

### 2.4.3 Conversion of CoRoLa-based word embeddings to LOD specifications

Multiple word embeddings representations were trained using the CoRoLa corpus (Păiș and Tufiș, 2018b). For conversion to OntoLex-FRAC specifications we considered only the best performing model as described by Păiș and Tufiș (2018a). The model contains vector representations of dimension 300 for 250,942 words. Data is encoded using the *frac:Embedding* class. An example is given in Listing 8.

```
:de a ontolex:LexicalEntry;
  ontolex:canonicalForm "de"@ro;
  frac:embedding [
    a :CoRoLaEmbeddings_300;
    rdf:value "0.058826 0.050749 0.094646 0.059437 0.014913 -
0.14699 0.31223 0.25699 0.020498 0.1497 -0.045657 -0.16574 -
0.14085 0.053746 -0.016113 -0.11879 0.11086 -0.086826 -
0.11564 -0.1137 -0.21041 0.12873 0.074748 -0.1439 -0.11781 -
0.14723 0.080661 0.18918 0.079647 -0.043609 -0.024831
0.058612 0.0028617 0.074098 0.048036 ..... ] .
```

**Listing 8**. Word embedding representation

## 2.5 ROBIN Technical Acquisition Speech Corpus (RTASC)

### 2.5.1 Description of RTASC

RTASC (Păiș et al., 2021b; 2021c) is a bimodal (speech and text) corpus, resulting from the ROBIN[34] project. The dataset was initially created to facilitate the construction of a dialogue component (Ion et al., 2020) for human-robot interaction in the context of a micro-world scenario of purchasing computers. It allowed us to improve the performance of our general speech recognition system (Avram et al., 2020) by approximately 16 WER on this domain. RTASC contains 3,786 audio files, with a total duration of 6h25m, read by multiple Romanian native speakers, associated with 711 text files. Being a read speech corpus, the audio files are aligned with the text variants. The text component was processed automatically in the RELATE platform (Păiș et al., 2020; Păiș, 2020), being tokenized, lemmatized and enhanced with morphosyntactic annotations and dependency parsing.

### 2.5.2 Conversion of RTASC to LOD specifications

The processed text component of the corpus became a treebank. Therefore, the conversion followed a process similar to the one described in Section 2.3.2 for the other treebanks. However, RTASC also contains the speech component which was linked to the texts. Furthermore, additional metadata, such as the recording devices used was included using the Studio Ontology Framework[35] (Fazekas and Sandler, 2011). An example speaker representation is in Listing 9.

```
:speaker_1
   a ma:Person, foaf:Person, studio:Device;
   foaf:gender "m";
   studio:microphone "Realtek HD Audio/Speedlink SL-8703-
BK" .
```

**Listing 9**. Speaker description in the RTASC corpus

[31] https://korap.racai.ro/

[32] http://89.38.230.23/corola_sound_search/index.php

[33]https://github.com/ontolex/frequency-attestation-corpus-information

[34] http://aimas.cs.pub.ro/robin/en/

[35] http://isophonics.org/content/studio-ontology

Audio files were added using the Ontology for Media Resources 1.0[36]. Thus a document in the corpus reflects its bimodal characteristics by making use of both *powla:Document* and *ma:DataTrack* classes. Then the actual audio file is represented using the class *ma:AudioTrack* referencing the wav file using the property *ma:hasFragment*. Additionally, the language, format and sampling rate are specified as well as the speaker who recorded the file. The link with the text representation is given as a *ma:hasSubtitling* property. An example is provided in Listing 10.

```
:d1 a powla:Document, ma:DataTrack ;
  powla:documentID "S0" ;
  powla:hasSuperDocument :c1 ;
  ma:hasLanguage [ rdfs:label "ro" ] .

:S0_4_wav a ma:MediaResource ;
  ma:hasTrack :S0_4_wav_audio ;
  ma:hasSubtitling :d1 .

:S0_4_wav_audio a ma:AudioTrack ;
  ma:hasLanguage [ rdfs:label "ro" ] ;
  ma:hasFormat "audio/wav" ;
  ma:samplingRate "44.1" ;
  ma:hasContributor :speaker_4 ;
  ma:hasFragment <S0_4.wav> .
```

**Listing 10**. Linking audio files in the RTASC corpus

## 2.6. LOD Vocabularies Usage

In our endeavour, we purposefully focused on reusing vocabularies already widely used in the LOD community and we did not create new classes or properties to represent our specificities, but adapted already existing ones. Tables 1 and 2 summarise the vocabularies used in each of the converted resources to represent the encoded data and metadata, respectively.

| | RoWN | RoLEX | RRT | LegalNERo | SiMoNERo | PARSEME-Ro | CoRoLa-based freq | RTASC |
|---|---|---|---|---|---|---|---|---|
| **ontolex-lemon**[37] | x | x | | | | | x | |
| **wn** | x | x | | | | | | |
| **ili**[38] | x | x | | | | | | |
| **frac**[39] | | | | | | | x | |
| **rdfs**[40] | | x | x | x | x | x | | |
| **rdf**[41] | | | | | | | x | |
| **vartrans**[42] | x | | | | | | | |
| **nif**[43] | | | x | x | x | x | | x |
| **powla**[44] | | | | | | | | x |
| **conll**[45] | | x | x | x | x | x | x | |
| **conllu**[46] | | | | | | | x | |
| **ma**[47] | | | | | | | | x |
| **xsd**[48] | | | | | | | x | x |

Table 1. Vocabularies used to encode data in Romanian resources converted to LOD.

| | RoWN | RoLEX | RRT | LegalNERo | SiMoNERo | PARSEME-Ro | CoRoLa-based freq | RTASC |
|---|---|---|---|---|---|---|---|---|
| **dct**[49] | | | | | | | x | |
| **foaf**[50] | | | | | | | | x |
| **studio**[51] | | | | | | | | x |
| **owl**[52] | | | | | | | | x |
| **dcat**[53] | | | | | | | | x |
| **prov**[54] | | | | | | | | x |
| **pav**[55] | | | | | | | | x |
| **dc**[56] | x | x | x | x | x | x | | |
| **cc**[57] | x | x | x | x | x | x | | |
| **schema**[58] | x | x | x | x | x | x | | |

Table 2. Vocabularies used to encode metadata in Romanian resources converted to LOD.

# 3. Publication of Romanian LOD Resources

Two ways of publishing (Verborgh, 2021) resources in the LD format have been adopted: data dump and availability of SPARQL endpoint. A dedicated page[59] was created for these resources, where they can be downloaded from.

A SPARQL Apache Jena Fuseki server has been installed on one of our servers. It can upload RDF-Turtle files, similar to those produced in this project, and then allow them to be queried online by providing a SPARQL endpoint. The server can be accessed at https://relate.racai.ro/datasets/. It first presents a list of

[36] https://www.w3.org/TR/mediaont-10/
[37] Ontology-lexicon interface: http://www.w3.org/ns/lemon/ontolex#
[38] https://github.com/globalwordnet/cili/blob/master/ili.ttl
[39] Frequency, attestation and corpus information: http://www.w3.org/ns/lemon/frac#
[40] RDF Schema: http://www.w3.org/2000/01/rdf-schema#
[41] RDF: http://www.w3.org/1999/02/22-rdf-syntax-ns#
[42] http://www.w3.org/ns/lemon/vartrans
[43] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core\#
[44] http://purl.org/powla/powla.owl\#
[45] http://ufal.mff.cuni.cz/conll2009-st/task-description.html#
[46] https://universaldependencies.org/format.html\#
[47] http://www.w3.org/ns/ma-ont\#
[48] XSD: http://www.w3.org/2001/XMLSchema\#
[49] DCMI Metadata Terms: http://purl.org/dc/terms/
[50] http://xmlns.com/foaf/0.1/
[51] http://isophonics.net/content/studio-ontology
[52] http://www.w3.org/2002/07/owl\#
[53] http://www.w3.org/ns/dcat\#
[54] http://www.w3.org/ns/prov\#
[55] http://pav-ontology.github.io/pav/
[56] http://purl.org/dc/elements/1.1/
[57] http://creativecommons.org/ns
[58] http://schema.org/
[59] https://www.racai.ro/p/llod

available resources and then allows the user to select the desired resource and run a query.

Another way of advertising our resources is registering their metadata in the LOD Cloud. They are now retrieved while browsing the cloud.

## 4. Interlinking of Romanian LOD Resources

The resources we converted to LOD specifications are either linked to other resources or among themselves.

The use of *wn:ili* property in the representation of RoWN ensures its linking to the other wordnets linked to it. The mapping to ILI was done automatically via the intrinsic mapping to PWN3.0. The mapping of ILI to different versions of PWN[60] is publicly offered by the Global Wordnet initiative. Through *wn:ili* property, 59,348 links were created to concepts in any wordnet linked to ILI.

RoLEX is linked to RoWN via ILI, thus ensuring the phonemic description for the words therein. Lemma forms in RoLEX are linked to all occurrences of the respective word in RoWN. However, homographs (i.e., words spelt identically but pronounced differently) are manually semantically disambiguated and then linked to the RoWN correspondent. The linking was automatic, by matching LexicalEntry labels/URIs in RoLEX with corresponding LexicalEntry labels in RoWN and recovering all the synsets (ontolex:LexicalSense) associated with them. As we mentioned, all the LexicalSense descriptions have a *wn:ili* property, and all these properties were extracted and transported to their corresponding LexicalEntry label/URI in RoLEX. As a LexicalEntry object in RoLEX is matched with more Lexical Entry objects in RoWN (e.g. *RoLEX:lex_pom* is matched with both *RoWN:pom-n-12651821* and *RoWN:pom-n-13104059*), each entry in RoLEX has a list of corresponding ILI.

Location entities in the LegalNERO corpus were mapped to the GeoNames resource, when linking was possible. This was automatically performed by using a lookup script created in the Python language that matched GeoNames entries with Location entities at lemma level. For each of the matched Location entities, the corpus was subsequently enriched with the GeoNames code id as an unique identifier for the GeoNames resource. In this process there were 1,411 Location entities that have been matched with a GeoNames code.

## 5. Use cases

The power of LOD is seen when combining multiple resources to produce powerful usage scenarios. This interlinking process exploits the internal structure of these resources and the points they have in common. These can be either identifiers (like those in ILI) or specific word forms found in multiple resources (for example the lemma of a word in the treebanks, expressed by the conll:lemma property can be linked to the *ontolex:canonicalForm* property of the ontolex:LexicalEntry class in RoLEX or RoWN).

Having all the resources available as SPARQL endpoints (see Section 3) allows for formulating complex SPARQL queries exploiting multiple datasets, using the SERVICE keyword (otherwise known as a federated query[61]). A powerful example of a complex federated query exploiting multiple resources is a conceptual search in a speech corpus. In this case, the concept is first looked up in RoWN and related concepts are retrieved. Then, the RoLEX lexicon can be employed to extract the word forms associated with the identified concepts, making use of the ILI identifiers. Finally, the resulting words are looked up in the speech corpus (in our case RTASC, see Section 2.5). For the result, the user obtains a list of audio files containing words related to the concept used in the initial query. This is depicted in Figure 1. The associated SPARQL query implementing the process is given in Listing 11. Finally, example results obtained from running the query through the Fuseki-provided SPARQL endpoints is given in Figure 2.

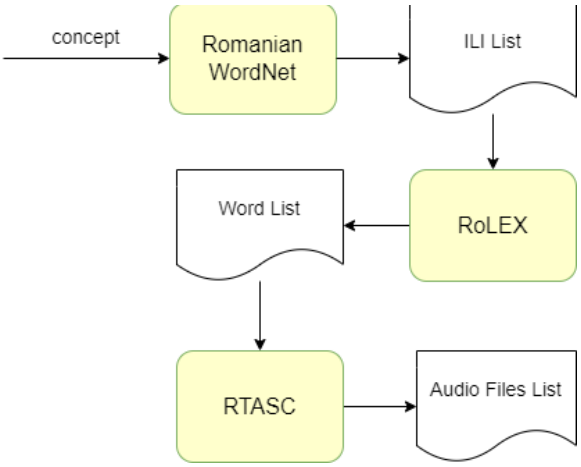


Figure 1. Example concept-based searching in a speech corpus by means of a federated query exploiting multiple Romanian resources

```
SELECT ?ili ?formString ?audio
WHERE {
 { SELECT DISTINCT ?ili1 WHERE {
  ?idConcept ontolex:writtenRep "calculator"@ro .
  ?idLex a ontolex:LexicalEntry .
  ?idLex ontolex:canonicalForm ?idConcept .
  ?idLex ontolex:Sense ?sense .
  ?sense a ontolex:LexicalSense .
  ?sense ontolex:reference ?ref .
  ?ref a ontolex:LexicalConcept .
  ?ref wn:ili ?ili .
  ?x vartrans:source ?ref .
  ?x vartrans:category wn:hypernym .
  ?x vartrans:target ?ref1 .
  ?senser1 ontolex:reference ?ref1 .
  ?ref1 wn:ili ?ili1 .
  ?idr1 ontolex:Sense ?senser1 .
  ?idr1w ontolex:writtenRep ?w .
  ?idr1 ontolex:canonicalForm ?idr1w .
  }}

 SERVICE <https://relate.racai.ro/datasets/rolex/sparql> {
  ?idRolex wn:ili ?ili1 .
  ?idRolex ontolex:lexicalForm ?idRolexForm .
  ?idRolexForm ontolex:writtenRep ?formWritten .
  FILTER (lang(?formWritten)="ro") .
  BIND (str(?formWritten) as ?formString) .
 }
  SERVICE <https://relate.racai.ro/datasets/rtasc/sparql> {
  ?id_tok conllu:FORM ?formString .
  ?id_tok powla:hasLayer ?id_layer .
```

[60] https://github.com/globalwordnet/cili

[61] https://www.w3.org/TR/sparql11-federated-query/

```
  ?id_layer powla:hasDocument ?id_doc .

  ?idr a ma:MediaResource .
  ?idr ma:hasSubtitling ?id_doc .
  ?idr ma:hasTrack ?ida .

  ?ida a ma:AudioTrack .
  ?ida ma:hasFragment ?audio .
}
}
```

**Listing 11**. Federated query providing concept-based retrieval of speech files, combining the Romanian Wordnet, the RoLEX lexicon and the RTASC speech corpus

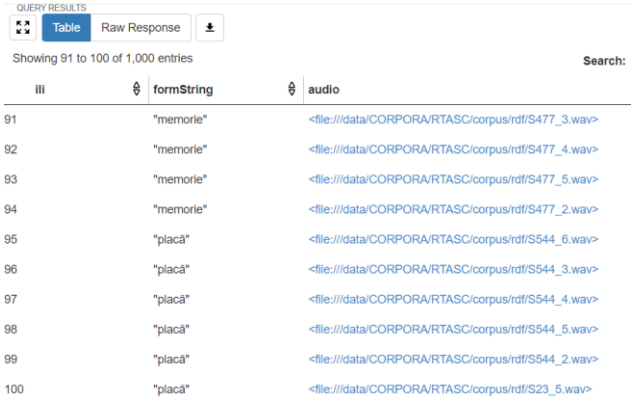


Figure 2. Example results when searching for concepts related to the word "calculator" ("computer")

Apart from this example, other usage scenarios of different degrees of complexity can be envisaged, exploiting other features of the resources. Thus, one can start with a phonetic search in the RoLEX lexicon (possibly as a result of obtaining phonemes from an ASR system) and then go to RoWN to retrieve concepts. In case of multiple words having similar phonetic representations (as indicated by RoLEX results) one can make use of the frequencies list extracted from the CoRoLa corpus to identify the most likely word form.

It is possible to make use of the interlinking process to query multilingual resources. Thus, one can start with a phonetic query in RoLEX, identify relevant RoWN concepts and then employ the ILI identifier to find similar concepts in another language (or multiple languages). Finally, the result can be used to perform concept-based retrieval in the foreign language(s).

The presence of references to the GeoNames ontology in the LegalNERo corpus opens up the possibility of performing queries filtered by geospatial criteria. These can be performed directly or combined with other resources (RoLEX, RoWN) to obtain more complex results.

## 6. Conclusions

We have taken important steps towards making a pool of Romanian language resources available to the community in LD format. They have been released in an open manner and are accessible in standard formats, reusing existing vocabularies, and can be queried using a SPARQL endpoint. They are now ready to be exploited to the potential offered by Linked Data. Our conversion of resources to LD specifications has led to what seems to be a duplication of resources in the LLOD Cloud. RoWN was already available from MultiWordNet and RRT was already available as the Romanian treebank in UD. However, the version of RoWN that we have converted is the whole resource that has been created, as opposed to the sample that is available in MultiWordNet. The RRT we converted is a newer version (including some recent corrections) of that already available in the LLOD Cloud.

Some resources require further enhancement by adding extra information: e.g., derivational relations existing between Romanian word senses to be added to the RoWN (Barbu Mititelu, 2012), links between the verbs in RRT and their corresponding senses in the valence lexicon of Romanian verbs (Barbu et al., 2022), links between the verbal multiword expressions in PARSEME-Ro corpus and their corresponding senses in RoWN (Barbu Mititelu et al., 2019b), etc.

The treebanks that are released via UD pose the problem of keeping track of their biannual versions. A solution to this will be seeked.

With regard to the sustainability of the resources, developed tools allow execution on primary data, such as the RoWN development format. Therefore, when new versions of the resources become available, they will be exported to the LOD format. Nevertheless, this is a manual process, requiring a developer to also execute the corresponding LOD export tool.

## 10. Bibliographical References


Avram, A. M., Păiș, V., and Tufis, D. (2020). Towards a Romanian end-to-end automatic speech recognition based on Deepspeech2. In *Proc. Rom. Acad*. Ser. A, Vol. 21, pp. 395-402.

Barbu, A.-M. (2008). Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp. 1937-1941.

Barbu, A.-M., Barbu Mititelu, V., Mititelu, C. (2022). Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs. In *Proceedings of LREC 2022* (in press).

Barbu Mititelu, V. (2012). Adding Morpho-Semantic Relations To The Romanian Wordnet. In *Proceedings of LREC2012*, Istanbul, Turkey, pp. 2596-2601.

Barbu Mititelu, V. (2018). Modern Syntactic Analysis of Romanian. In Ofelia Ichim, Luminița Botoşineanu, Daniela Butnaru, Marius-Radu Clim, Ofelia Ichim, Veronica Olariu (eds.), *Clasic şi modern în cercetarea filologică românească actuală*, Iași, Publishing House of "Alexandru Ioan Cuza" University, pp. 67-78.

Barbu Mititelu, V., Cristescu, M. and Onofrei, M. (2019a). The Romanian Corpus Annotated with Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, Florence, Italy, pp 13–21.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Mitrofan, M., Dimitrova, T. and Todorova, M. (2019b). Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*. Association for Computational Linguistics, Florence, Italy, pp. 2-12.

Barbu Mititelu, V. and Mitrofan, M. (2020). The Romanian Medical Treebank – SiMoNERo. In

*Proceedings of the 15th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, Editura Universității A. I. Cuza, Iași, pp. 7–16.

Boroş, T., Dumitrescu, S.D. and Burtica, R. (2018). NLP-Cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 171-179.

Chiarcos C. and Fäth C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., Hellmann S. (eds) *Language, Data, and Knowledge. LDK*. pp 74-88.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Banski, P. and Witt, A. (2016). KorAP architecture – Diving in the Deep Sea of Corpus Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S.Piperidis (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 3586–3591.

Diewald, N., Barbu Mititelu, V., Kupietz, M. (2019). The KorAP User Interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique*, 64(3): 265–277.

Fazekas, G. and Sandler, M. B. (2011). The studio ontology framework. In 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 471–476.

Fellbaum, Ch. (1998, ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Bond, F., Vossen, P., McCrae, J. P., & Fellbaum, C. (2016). Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pp. 50-57.

Ion, R. (2007). *Word sense disambiguation methods applied to English and Romanian*. PhD diss., PhD thesis (in Romanian). Romanian Academy, Bucharest.

Ion, R., Badea, V.G., Cioroiu, G., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Tufiș, D. (2020). A Dialog Manager for Micro-Worlds. In *Studies in Informatics and Control*. 29(4):411-420

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39-41.

Mitrofan, M. and Tufiș, D. (2018). BioRo: The biomedical corpus for the Romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Păiș, V. and Tufiș, D. (2018a). Computing distributed representations of words using the CoRoLa corpus. In Proceedings of the Romanian Academy, series A, 403-410.

Păiș, Vasile and Tufiș, Dan. (2018b). More Romanian word embeddings from the RETEROM project. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language - CONSILR, 91-100.

Păiș, V. (2020). Multiple annotation pipelines inside the RELATE platform. In *Proceedings of the 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*. pp. 65--75.

Păiș, V., Tufiș, D. and Ion, R. (2020). A Processing Platform Relating Data and Tools for Romanian Language. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pp. 81-88.

Păiș, V. and Mitrofan, M. (2021). Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus. In *Workshop on Deep Learning and Neural Approaches for Linguistic Data*. Skopje, North Macedonia, pp. 16--17.

Păiș, V., Mitrofan, M., Gasan, C.L., Coneschi, V. and Ianov, A. (2021a). Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 9-18.

Păiș, V., Ion, R., Avram, A.M., Irimia, E., Barbu-Mititelu, V., and Mitrofan, M. (2021b). Human-Machine Interaction Speech Corpus from the ROBIN project. In *Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 91-96.

Păiș, V., Ion, R., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Avram, A. (2021c). ROBIN Technical Acquisition Speech Corpus. Zenodo, https://doi.org/10.5281/zenodo.4626539.

Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A. and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In the *Proceedings of Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, workshop at COLING 2018, Santa Fe, USA, pp. 222-240.

Ramisch, C. et al., 2020, *Annotated corpora and tools of the PARSEME Shared Task on Semi-Supervised Identification of Verbal Multiword Expressions (edition 1.2)*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3367.

Toma, Ş.A, Stan, A., Pura, M.L and Bârsan, T. (2017). MaRePhoR—An open access machine-readable phonetic dictionary for Romanian. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1-6.

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S., Eryiğit, G., Giouli, V., Van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Parra Escartín, C., van der Plas, L., Qasemi Zadeh, B., Ramisch, C. and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary and V. Vincze (Eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop Phraseology and Multiword Expressions*. Berlin: Language Science Press, pp. 87–147.

Stan, A., Yamagishi, J., King, S. and Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication* 53, no. 3 (2011): 442-450.

Straka, M., Hajič, J. and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Tufiș, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004a). The Romanian Wordnet. *Romanian Journal*

*of Information Science and Technology*, vol. 7, nr. 1-2, pp. 107-124

Tufiș, D., Cristea, D., Stamou, S. (2004b). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology*, 7(2-3): 9-34.

Tufiș, D. and Barbu Mititelu, V. (2014). The Lexical Ontology for Romanian. In N. Gala, R. Rapp, N. Bel-Enguix (Eds.), *Language Production, Cognition, and the Lexicon*, series Text, Speech and Language Technology, vol. 48. Springer, pp. 491-504.

Tufiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian. *Revue Roumaine de Linguistique*, 64(3): 227–240.

Tufiș, D., Mitrofan, M., Păiș, V., Ion, R. and Coman, A. (2020). Collection and Annotation of the Romanian Legal Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association*, Marseille, France, pp. 2766-2770.

Tufiș, D. (2022). Romanian Language Technology – a view from an academic perspective. *International Journal of Computers Communication and Control*, vol. 17, no. 1, 2022.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, vol. 7, pp. 63-93.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 45–51
Marseille, 24 June 2022


# Fuzzy Lemon: Making Lexical Semantic Relations More Juicy

**Fernando Bobillo, Julia Bosque-Gil, Jorge Gracia, Marta Lanau-Coronas**
Aragon Institute of Engineering Research (I3A)
University of Zaragoza, Spain
{fbobillo, jbosque, jgracia, mlanau}@unizar.es

**Abstract**

The OntoLex-Lemon model provides a vocabulary to enrich ontologies with linguistic information that can be exploited by Natural Language Processing applications. The increasing uptake of Lemon illustrates the growing interest in combining linguistic information and Semantic Web technologies. In this paper, we present *Fuzzy Lemon*, an extension of Lemon that allows to assign an uncertainty degree to lexical semantic relations. Our approach is based on an OWL ontology that defines a hierarchy of data properties encoding different types of uncertainty. We also illustrate the usefulness of Fuzzy Lemon by showing that it can be used to represent the confidence degrees of automatically discovered translations between pairs of bilingual dictionaries from the Apertium family.

**Keywords:** linked data, ontologies, fuzzy logic, uncertainty

## 1. Introduction

Managing linguistic information is important in many real-world applications, in particular in those taking advantage of Natural Language Processing (NLP) techniques. For this reason, there is an increase in the interest in combining linguistic information and Semantic Web technologies (Cimiano et al., 2020). Such Semantic Web technologies include ontologies, or formal and shared specifications of the vocabulary of a domain of interest (Staab and Studer, 2004), usually expressed in OWL (Cuenca-Grau et al., 2008); and linked data, a set of best practices for publishing and connecting data on the Web (Bizer et al., 2009), usually expressed in RDF (Schreiber and Raimond, 2014).

A very good example is the *Ontolex-Lemon* model, which intends to provide a vocabulary to enrich ontologies with information about how ontology elements can be realized in natural languages (Cimiano et al., 2016; McCrae et al., 2017). Lemon includes support to represent lexical semantic relations (between pairs of lexical entries, pairs of lexical senses, or pairs of lexical concepts) by means of its *Vartrans* module, as illustrated in Figure 1, borrowed from (Cimiano et al., 2016).

One of the limitations of Lemon is its inability to represent and manage uncertainty in the linguistic information. To manage uncertainty, the literature includes many extensions of Semantic Web technologies, such as Description Logics (Bobillo et al., 2015; Lukasiewicz and Straccia, 2008), ontologies (Zhang et al., 2016), SPARQL (Pan et al., 2008), or RDF (Straccia, 2009). The objective of this paper is to propose *Fuzzy Lemon*, an extension of Lemon to assign an uncertainty degree to lexical semantic relations. We understand the term "uncertainty" in a wide sense, and it is intended to embrace a variety of aspects of imperfect knowledge, including incompleteness, inconclusiveness, vagueness, ambiguity, and others (Laskey et al., 2008).

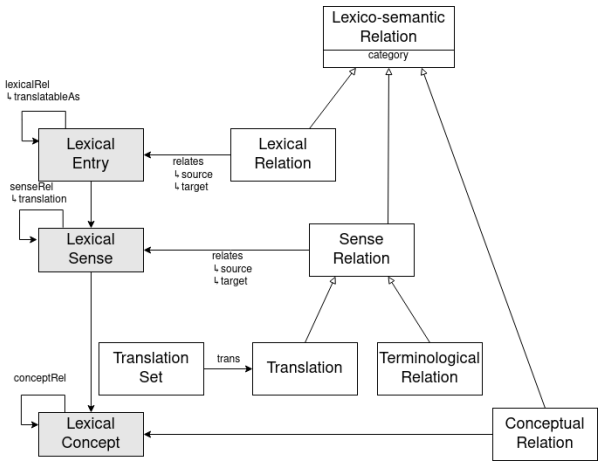


Figure 1: Scheme of the Lemon Vartrans module

The remainder of this paper is organized as follows. Section 2 discusses the need to support different types of uncertainty and some formalisms to do so. Section 3 presents Fuzzy Lemon ontology, enabling modeling uncertainty knowledge in lexical semantic data. Then, Section 4 discusses a use case: the representation of the confidence in automatically discovered translations between pairs of Apertium dictionaries. Finally, Section 5 ends up with some conclusions and ideas for future work.

## 2. Uncertainty in Lexical Semantic Relations

Current approaches to represent semantic relationships between pairs of lexical elements can be extended in several ways.

- Firstly, we might be interested in relations that partially hold, i.e., that hold to some *degree of*

*truth*. For example, there can be a translation between two terms in different languages that is imprecise or partially true, e.g., a Spanish "siesta" is slightly different than a "nap". A more involved example, borrowed from (León-Araúz et al., 2012) is that a Spanish "dique" is similar to an English "breakwater" with degree 0.9. This also makes it possible to model the degree of semantic overlap between two meanings of two terms (in the same or different languages), or between two definitions of the same sense in different dictionaries with different granularity. For example, M. Gonzàlez et al. show different meanings of the senses of "fog" in an English monolingual dictionary and in an English-Spanish bilingual dictionary (Gonzàlez et al., 2021).

- Secondly, we might be interested in relations that we are not sure about, so we would attach a *confidence degree* to them. For example, there could be a term in a source language which can be perfectly translated as another term in a target language, but we are not sure if the translation is correct, i.e., if it is the right one. For example, the Spanish term "primo" has two senses and can be translated into English either as "prime" (number) or as "cousin". This could be the case if we use an automatic software (e.g., Google Translate) to compute the translation of a term.

Both types of degrees require different formalisms to deal with them (Dubois and Prade, 2001).

- On the one hand, fuzzy logic can manage statements with an associated degree of truth (Zadeh, 1965; Klir and Yuan, 1995), expressing that the statement is partially true or, in other words, the extent to which the event described by such statement holds in the world. For example, "the bottle is full with fuzzy degree 0.5" means that the amount of liquid in the bottle is half of its total capacity. There can be completely full bottles (fuzzy degree 1) and completely empty bottles (fuzzy degree 0), but there are also bottles which are full up to some degree.

- On the other hand, possibilistic (Dubois and Prade, 1988) or probabilistic (Nilsson, 1986) logics can manage confidence degrees, which quantify our certainty about an event. In this case, there are several *worlds* or possible scenarios, but we are not sure which is the right one. The statement "the bottle is full with confidence degree 0.5" means that we are not sure about the status of the bottle, in some worlds it could be full, and in others it could be not full (but the amount of liquid does not need to be half of the total capacity).

  Probability logic is a well known formalism that tries to quantify how likely an event is. Possibilistic logic differs by the use of a pair of dual measures (possibility and necessity) rather than just one. A possibility degree quantifies how possible an event is (by taking the supremum value over all worlds), while the necessity degree quantifies how necessarily an event happens, by computing one minus the possibility of the negated event. For instance, "tomorrow it will rain with a possibility degree 1" means that there is a world where it will rain, but there could be other scenarios where it will not. Note also that if it is absolutely impossible that tomorrow will rain (the possibility degree is 0), it is necessarily true (the necessity degree is 1) that it will *not* rain.

Note that both approaches are actually orthogonal, and we might want to represent that we are partially confident on a statement being partially true.

## 3. Fuzzy Lemon Ontology

This section describes the elements of Fuzzy Lemon Ontology and how to use them to extend the Lemon model. We will also discuss how to populate and use the ontology, possible extensions, links to existing ontologies, and some reasoning strategies.

**Elements of the Ontology.** Fuzzy Lemon has been written in OWL 2 (Cuenca-Grau et al., 2008) and is publicly available[1]. It includes data properties linking a lexical semantic relation with a numerical or textual data type value representing the degree of the relation. The main data property is semanticRelationDegree, and it has as domain the class vartrans:LexicoSemanticRelation. Note in particular that it is possible to consider other lexico-semantic relations different than translations, for instance equivalence or hyponymy relations between pairs of Wordnet synsets (or similar resources), as long as they are represented using lemon as already proposed by (McCrae et al., 2014).
Next, we built a hierarchy of subproperties of semanticRelationDegree to support different uncertainty types (see Figure 2). In particular, we propose to consider fuzzyDegree and confidenceDegree. The latter one has two subproperties probabilisticDegree and possibilisticDegree. The latter property has two subproperties possibilityDegree and necessityDegree. Properties fuzzyDegree and confidenceDegree have as range the decimal numbers in the interval $[0, 1]$.
Properties fuzzyDegree, probabilisticDegree, possibilityDegree, and necessityDegree are functional. However, semanticRelationDegree, confidenceDegree, and possibilisticDegree are not. Therefore, it is possible to combine a single degree of truth with one or more confidence degrees, but we cannot combine several confidence degrees of the same type.

[1] http://sid.cps.unizar.es/ontology/fuzzyLemon.owl

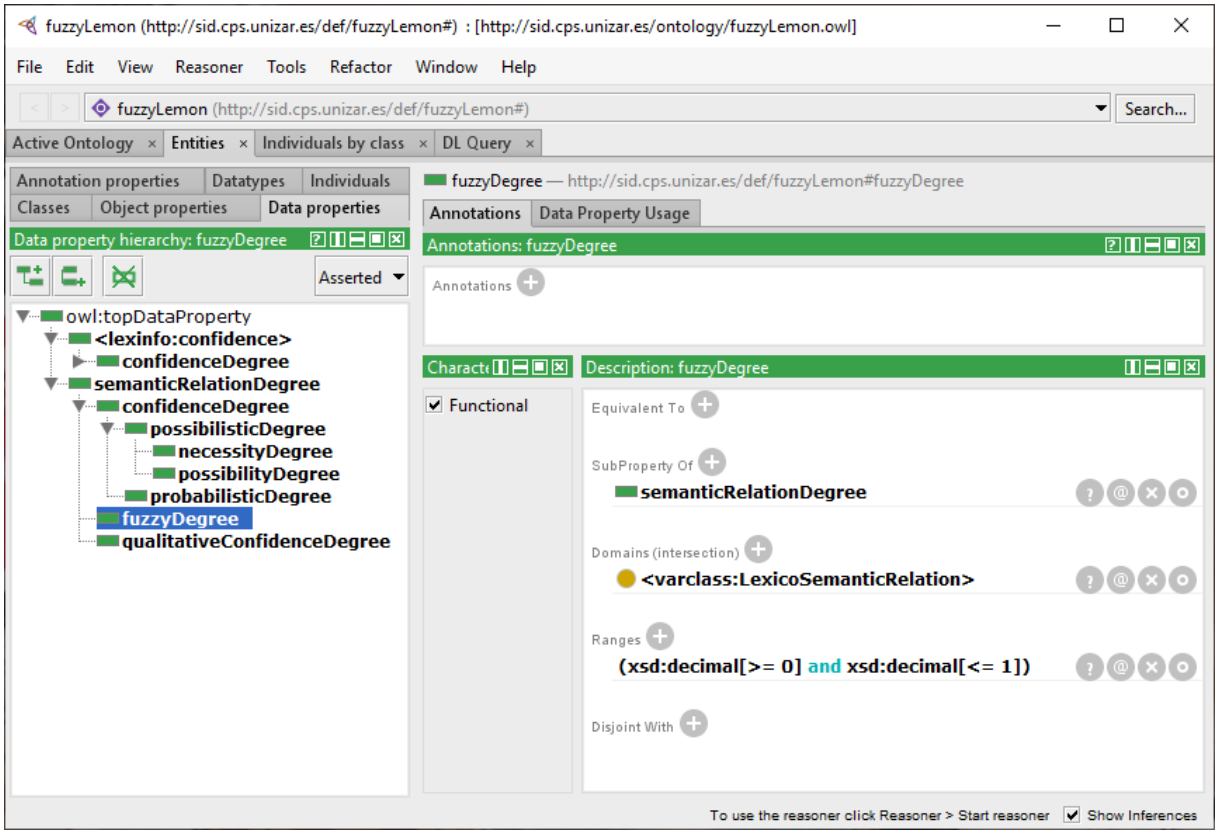


Figure 2: Subproperty hierarchy of the Fuzzy Lemon ontology

The default value of these semantic degrees is 1, making our extension backwards compatible. Therefore, if the value is 1, there is no need to represent it explicitly. Other authors have proposed using non-numerical (i.e., categorical) values to categorize the type of links between two lexical semantic elements. For example, possible values are perfect, partial, unknown, narrowerThan or widerThan (Gonzàlez et al., 2021). In order to support such non-numerical values, we have added a subproperty of semanticRelationDegree, called qualitativeConfidenceDegree, having as range an xsd:string value. Note that numerical values provide more information, e.g., two relations that hold with degrees 0.1 and 0.9 are both partial, but the second one is truer than the second one.

It is worth to mention that there is a previous ontology of uncertainty with more approaches to manage uncertainty, such as rough sets, belief functions, or random sets (Laskey et al., 2008). We have not reused it because uncertainty types are expressed as classes, but properties are more appropriate in our scenario. In particular, this allows to express domain and range restrictions, as well as non-numerical confidence degrees.

**Using the Ontology.** After having defined all these properties, we propose to extend the syntax of Lemon so that we can attach to a lexical semantic relationship (between senses, entries, or concepts) a degree via a subproperty of semanticRelationDegree. For example, we could add a degree to a translation (i.e., to an instance of the vartrans: Translation class). Example 1 shows how to add a fuzzy degree to a translation involving two entities ex:siesta and ex:nap.

**Populating the Ontology.** A common problem when managing uncertainty is how to obtain the concrete values of the degrees. A first option is to ask a human expert, or a group of them, to assign the values. In some cases, the proportion of lexical semantic relations with an attached degree seems to be very small, so this could be a feasible solution. This could be the case, for example, if a human expert is encoding the translations, and only translations which do not fully hold are annotated. Another option is to use some automatic or semiautomatic machine learning procedure to obtain the degrees from examples. The problem here is the need to obtain large amounts of data to learn from. In Section 4 we will discuss in detail one of the many possible ways to do it: learning the confidence degrees of translations between pairs of terms in different languages based on the cycles density (Villegas et al., 2016; Lanau-Coronas and Gracia, 2020).

**Extending the Ontology.** We have restricted to three logics (fuzzy, possibilistic, and probabilistic) because there has been some previous work to extend ontology axioms with them (Lukasiewicz and Straccia, 2008). Clearly, our ontology could be extended with more sub-

properties of semanticRelationDegree.
In the case of fuzzy degrees, it would also be possible to further generalize our approach by replacing the interval $[0, 1]$ with a more general structure, such as another interval or a lattice. Despite these possibilities, we argue for a simple but flexible approach, which could be further extended in the future if there is need to.

**Linking the Ontology.** Lexinfo ontology (Cimiano et al., 2011)[2] has a data property lexinfo:confidence and a sub-property lexinfo:translationConfidence. Our properties confidenceDegree (with a numerical range) and confidenceDegree (with a textual range) are stated to be subproperties of lexinfo:confidence, which is more general as it does not restrict the domain or the range.
SKOS vocabulary (Miles et al., 2005)[3] includes some object properties that are relevant to our work, such as skos:exactMatch or skos:closeMatch. skos:exactMatch could be used to represent relations without uncertainty, as in classical Lemon, whereas skos:closeMatch could be used to represent relations affected by uncertainty. However, our approach gives more information, as it makes it possible to specify the uncertainty type (e.g., probabilistic or fuzzy) and quantify the uncertainty (e.g., with a numerical degree), uses data properties rather than object properties, and make it possible to represent both exact (e.g., if the degree is 1) and close matches.

**Reasoning with the Ontology.** Another important problem is whether it is possible to infer new degrees from existing ones. In some cases, it is possible to exploit transitivity of the relationships. For example, if "siesta" can be translated as "nap" with degree $\alpha$, and "nap" can be translated as "sonnellino" (in Italian) with degree $\beta$, can "siesta" be translated as "sonnellino" with degree $\gamma$?
In fuzzy logic, given a transitive relation, one can infer that $\gamma \geq \alpha \otimes \beta$, where $\otimes$ is a t-norm function that generalizes the classical conjunction to the fuzzy case (Klement et al., 2000). Examples of t-norm functions are the minimum and the product. Note in particular that the product is subidempotent, which means that $\alpha \otimes \alpha < \alpha, \forall \alpha \in (0, 1)$. Note also that a similar approach is not possible with possibility degrees, as possibilistic logic is not truth-compositional (Dubois and Prade, 2001).
In our scenario, we claim that the retrieved candidate relations should be revised by a human expert before incorporating them into our knowledge base. For example, given two relations that partially hold because there is an overlapping between lexical senses, we might not be able to infer a third relation because overlapping is not transitive, as Figure 3 shows (blue and red squares overlap, red and green squares, but blue and green do not).

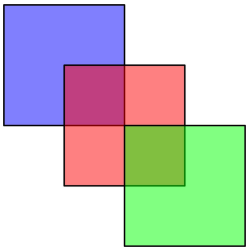

Figure 3: Example of non-transitivity of overlapping

## 4. Use case: Automatic translations

Apertium (Forcada et al., 2011) is a free open-source machine translation platform, initially created by Universitat d'Alacant and released under the terms of the GNU General Public License. In its core, Apertium relies on a set of bilingual dictionaries, developed by a community of contributors, which covers more than 50 languages pairs.
Apertium RDF (Gracia et al., 2018) is the result of publishing the Apertium bilingual dictionaries as linked data on the Web. The result groups the data of the (originally disparate) Apertium bilingual dictionaries in the same graph, interconnected through the common lexical entries of the monolingual lexicons that they share. In its current version, it contains 44 languages and 53 language pairs, with a total number of 1,540,996 translations between 1,750,917 lexical entries (Gracia et al., 2020).
Apertium RDF has been used in a number of campaigns of the Translation Inference Across Dictionaries (TIAD) initiative (Gracia et al., 2019; Kernerman et al., 2020)[4]. In this task, the participating systems were asked to generate new translations automatically among three languages, English, French, Portuguese, based on known translations contained in the Apertium RDF graph. As these languages (EN, FR, PT) are not directly connected in this graph, no translations can be obtained directly among them there. Based on the available RDF data, the participants applied their methodologies to derive translations, mediated by any other language in the graph, between the pairs EN/FR, FR/PT and PT/EN.
Motivated by the outcomes of this campaign, we are proposing in this work a way to semantically represent the inferred translations between pairs of senses resulting of such translation inference algorithms, which usually come with a confidence degree per translation pair.[5] In that way, the new translations can be "materi-

[2] https://www.lexinfo.net
[3] https://www.w3.org/2004/02/skos
[4] http://tiad2021.unizar.es
[5] Both the inferred data sets as well as their linked data representation are available at https://github.com/sid-unizar/fuzzy-lemon-translations.

Example 1. Representation of the fact that siesta can be translated as nap with a fuzzyDegree 0.5.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix ex: <http://example.org/> .
@prefix fuzzyLemon: <http://sid.cps.unizar.es/def/fuzzyLemon#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix trcat: <http://purl.org/net/translation-categories#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

ex:siesta a ontolex:LexicalEntry ;
          dct:language <http://lexvo.org/id/iso639-1/es> ,
                       <http://id.loc.gov/vocabulary/iso639-2/spa> ;
          ontolex:sense ex:siesta_sense .
ex:siesta_sense ontolex:reference <http://dbpedia.org/ontology/Nap> .
ex:nap a ontolex:LexicalEntry ;
       dct:language <http://lexvo.org/id/iso639-1/en> ,
       <http://id.loc.gov/vocabulary/iso639-2/eng> ;
       ontolex:sense ex:nap_sense .
ex:nap_sense ontolex:reference <http://es.dbpedia.org/resource/Siesta> .
ex:trans a vartrans:Translation ;
         vartrans:category trcat:directEquivalent ;
         fuzzyLemon:fuzzyDegree "0.5"^^xsd:decimal ;
         vartrans:source ex:siesta_sense ;
         vartrans:target ex:nap .
```

alized" in RDF and re-introduced in the overall Apertium RDF graph, with suitable provenance information and confidence degree, thus coexisting with human made translations from the original Apertium and enabling new ways of processing and enriching the Apertium data.

In particular, we focus on an algorithm that exploits the existence of cycles in the Apertium RDF graph structure in order to infer new translations, and computes their confidence degree based on the cycles density (Villegas et al., 2016; Lanau-Coronas and Gracia, 2020).

For the newly inferred translations we had thus three representation needs. We wanted to capture (1) the confidence score assigned for the translation, (2) the path followed in the Apertium RDF graph, that is, the entries making up the cycle leading to such a translation, and (3) their provenance as outcome from a specific translation inference software. For (1), we turned to the use of the property fuzzyLemon:confidenceDegree.

Figure 4 represents the inferred translation of *calendar* (English) into French (*calendrier*) as linked data. The inferred vartrans:Translation linking two senses, in English and French respectively, is given a confidence score of 0.83. For (2), note that we have kept the cycle information inside a comment linked to the translation: starting from English (*calendar*), and traversing through Spanish, French, and Esperanto, the cycle is closed with English again.

For aspect (3), the Prov Ontology (Belhajjame et al., 2012) has been used. In our case, the inferred translation belongs to a translation set (vartrans:TranslationSet) which is attributed to a prov:SoftwareAgent representing our inference system via the property prov:wasAttributedTo. This system is related to a prov:Activity representing the inference activity itself, which produces as output the translation set. This reflects the interplay between Agent-Activity-Entity accounted for in Prov-O.

The following pairs of dictionaries were inferred following the approach based on cycle density, and concern English, French, Esperanto, Italian and Sardinian (in parentheses, the number of inferred translations per dictionary): EN-FR (7772), EO-IT (7607), FR-IT (6497), SC-FR (4607), SC-EO (3473).

## 5. Conclusions

In this paper, we have presented *Fuzzy Lemon*, an extension of Lemon model that makes it possible to assign an uncertainty degree to lexical semantic relations. This can be achieved by means of an OWL ontology that defines a hierarchy of data properties supporting the management of different uncertainty types. The model has also been designed in such a way that future extensions to support more uncertainty types.

Because uncertainty is inherent to many real-world domains, Fuzzy Lemon can be useful in many Natural Language Processing or, more generally, Artificial Intelligence applications. To illustrate the usefulness of

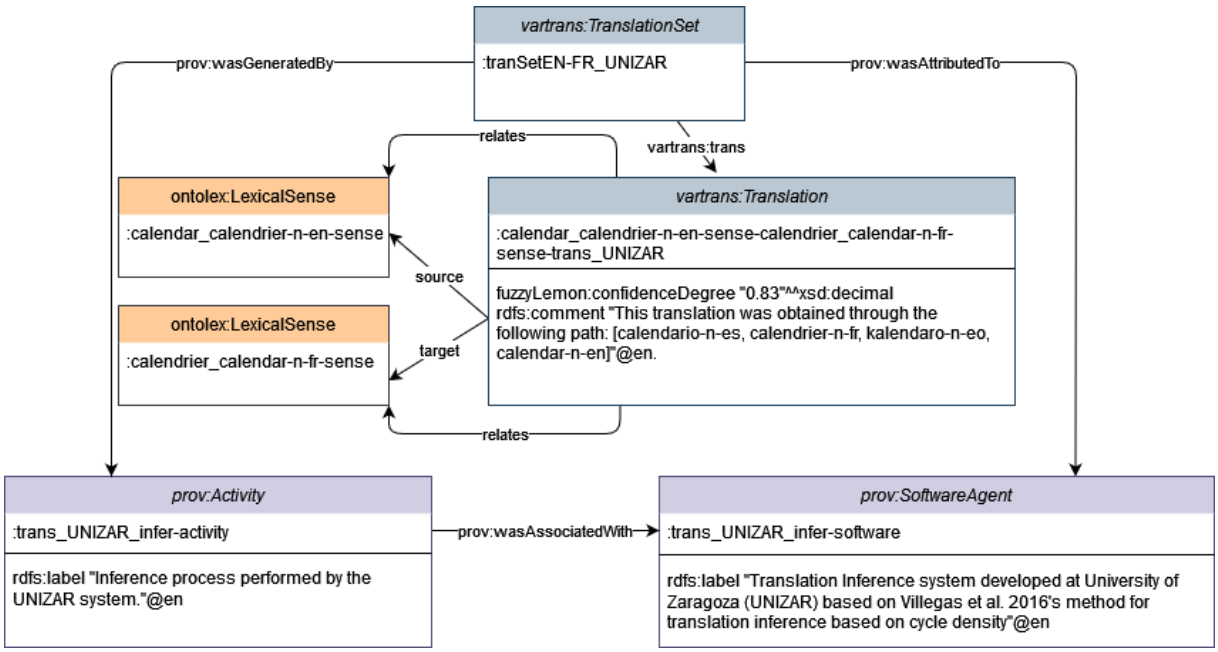


Figure 4: The inferred EN-FR translation *calendar-calendrier* represented as linked data with Fuzzy Lemon and Prov-O. The elements shaded in orange were already available in the Apertium graph, gray elements are inferred, and purple entities are common to all the inferred pairs and refer to the implemented system.

Fuzzy Lemon, we have shown how it can be used to represent the confidence degrees of automatically discovered translations between pairs of Apertium dictionaries.

As a next step, we plan to involve the broader W3C Ontolex community in order to gather feedback for our modelling proposal, to identify other possible use cases, and maybe to incorporate our proposed extension into the family of "official" Lemon modules in the future.

It would also be interesting adding to the model details about the creator of the uncertainty information, particularly when it comes from a machine learning software. A possible idea is to reuse the ideas behind the Internationalization Tag Set (ITS) tool annotation.[6].

## 6. Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program through the project Prêt-à-LLOD (grant agreement No 825182), by the I+D+i project PID2020-113903RB-I00, funded by MCIN/AEI/10.13039/501100011033, by DGA/FEDER, and by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

## 7. Bibliographical References

Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2012). PROV-O: The PROV ontology. Technical report, W3C Recommendation. `http://www.w3.org/TR/prov-o`.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Bobillo, F., Cerami, M., Esteva, F., García-Cerdaña, À., Peñaloza, R., and Straccia, U. (2015). Fuzzy description logics. In Petr Cintula, et al., editors, *Handbook of Mathematical Fuzzy Logic Volume III*, volume 58 of *Studies in Logic, Mathematical Logic and Foundations*, chapter XVI, pages 1105–1181. College Publications.

Cimiano, P., Buitelaar, P., McCrae, J. P., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon model for ontologies: Community report. Final community group report, W3C. `http://www.w3.org/2016/05/ontolex`.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

Cuenca-Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. (2008). OWL 2: The next step for OWL. *Journal of Web Semantics*, 6(4):309–322.

Dubois, D. and Prade, H. (1988). *Possibility Theory - An Approach to Computerized Processing of Uncer-*

[6] `https://www.w3.org/TR/its20`

*tainty*. Springer.

Dubois, D. and Prade, H. (2001). Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):35–66.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gonzàlez, M., Buxton, C., and Saurí, R. (2021). XD-AT: A cross-dictionary annotation tool. In *Proceedings of the XIX International Congress of the European Association for Lexicography (EURALEX 2020)*, volume 1, pages 503–514.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2018). The Apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., and Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. In Jorge Gracia, et al., editors, *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries, at 2nd Language, Data and Knowledge Conference (LDK 2019)*, volume 2493, pages 1–12, Sintra (Portugal). CEUR-WS.

Gracia, J., Fäth, C., Hartung, M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., and Orlikowski, M. (2020). Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain. In Bo Fu et al., editors, *Proceedings of 19th International Semantic Web Conference (ISWC 2020)*, pages 499–514. Springer.

Kernerman, I., Krek, S., Mccrae, J. P., Gracia, J., Ahmadi, S., and Kabashi, B. (2020). Introduction to the Globalex 2020 Workshop on Linked Lexicography. In Ilan Kernerman, et al., editors, *Proceedings of Globalex'20 Workshop on Linked Lexicography at LREC 2020*. ELRA.

Klement, E. P., Mesiar, R., and Pap, E. (2000). *Triangular Norms*, volume 8 of *Trends in Logic*. Kluwer.

Klir, G. J. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall, Inc.

Lanau-Coronas, M. and Gracia, J. (2020). Graph exploration and cross-lingual word embeddings for translation inference across dictionaries. In Ilan Kernerman, et al., editors, *Proceedings of Globalex'20 Workshop on Linked Lexicography at LREC 2020*, pages 106–110. ELRA.

Laskey, K. J., Laskey, K. B., Costa, P. C. G., Kokar, M. M., Martin, T., and Lukasiewicz, T. (2008). Uncertainty reasoning for the World Wide Web (URW3) incubator group report. Technical report, W3C Incubator Group Final Reports. `http://www.w3.org/2005/Incubator/urw3/XGR-urw3`.

León-Araúz, P., Gómez-Romero, J., and Bobillo, F. (2012). A fuzzy ontology extension of WordNet and EuroWordnet for specialized knowledge. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*, pages 139–154.

Lukasiewicz, T. and Straccia, U. (2008). Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics*, 6(4):291–308.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking Wordnet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (2014)*.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, in Leiden, Netherlands*, pages 587–597. Lexical Computing CZ s.r.o., sep.

Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005). SKOS Core: Simple knowledge organisation for the Web. In *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications (DCMI 2005)*. Dublin Core Metadata Initiative.

Nilsson, N. J. (1986). Probabilistic logic. *Artificial Intelligence*, 28(1):71–87.

Pan, J. Z., Stamou, G., Stoilos, G., Thomas, E., and Taylor, S. (2008). Scalable querying service over fuzzy ontologies. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 575–584.

Schreiber, G. and Raimond, Y. (2014). RDF 1.1 primer. W3c working group note, W3C. `http://www.w3.org/TR/rdf11-primer`.

Staab, S. and Studer, R. (2004). *Handbook on Ontologies*. International Handbooks on Information Systems. Springer.

Straccia, U. (2009). A minimal deductive system for general fuzzy RDF. In *Proceedings of the 3rd International Conference on Web Reasoning and Rule Systems (RR 2009)*, volume 5837 of *Lecture Notes in Computer Science*, pages 166–181. Springer-Verlag.

Villegas, M., Melero, M., Gracia, J., and Bel, N. (2016). Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In Nicoletta Calzolari Conference Chair, et al., editors, *Proceedings of 10th Language Resources and Evaluation Conference (LREC'16) Portorož (Slovenia)*, pages 868–876. European Language Resources Association (ELRA).

Zadeh, L. A. (1965). Fuzzy sets. *Information Control*, 8:338–353.

Zhang, F., Cheng, J., and Ma, Z. (2016). A survey on fuzzy ontologies for the semantic web. *Knowledge Engineering Review*, 31(3):278–321.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 52–60
Marseille, 24 June 2022


# A Cheap and Dirty Cross-Lingual Linking Service in the Cloud

**Christian Chiarcos**[1,2], **Gilles Sérasset**[3]
[1]Applied Computational Linguistics, Goethe University Frankfurt, Germany
[2]Institute for Digital Humanities, University of Cologne, Germany
[3]GETALP, LIG, Univ. Grenoble Alpes, CNRS, Grenoble, France
chiarcos@cs.uni-frankfurt.de, gilles.serasset@imag.fr

**Abstract**

In this paper, we describe the application of Linguistic Linked Open Data (LLOD) technology for dynamic cross-lingual querying on demand. Whereas most related research is focusing on providing a static linking, i.e., cross-lingual inference, and then storing the resulting links, we demonstrate the application of the federation capabilities of SPARQL to perform lexical linking on the fly. In the end, we provide a baseline functionality that uses the connection of two web services – a SPARQL end point for multilingual lexical data and another SPARQL end point for querying an English language knowledge graph – in order to perform querying an English language knowledge graph using foreign language labels. We argue that, for low-resource languages where substantial native knowledge graphs are lacking, this functionality can be used to lower the language barrier by allowing to formulate cross-linguistically applicable queries mediated by a multilingual dictionary.

**Keywords:** Linguistic Linked Open Data, DBnary, DBpedia, cross-lingual querying

## 1. Introduction

Since its conception about a decade ago (Chiarcos et al., 2011), Linguistic Linked Open Data (LLOD) technology has begun to establish itself in the areas of language technology, linguistics and lexicography, most notably demonstrated by the development of a Linguistic Linked Open Data cloud,[1] and its increasing degree of maturity is demonstrated in a number of collected volumes, e.g., (Pareja-Lora et al., 2020), as well as a designated monography (Cimiano et al., 2020) that summarizes the state of the art in the field. As already observed by (Chiarcos et al., 2013), Linguistic Linked Open Data (LLOD) technology has a number of benefits in its application to language resources and language technology: The use of web standards such as RDF and SPARQL, as well HTTP-resolvable URIs for identifying and referring to content elements allows to establish links between resources published on the web of data, and this *linkability* entails advantages with respect to representation and modelling (graphs can represent any linguistic data structure), structural and conceptual interoperability (generic data structures, shared vocabularies, uniform access protocol), federation (querying over distributed data), dynamicity (access remote resources at query time) and the availability of a mature technical ecosystem for language technology.

In this context, especially the field of lexical resources has flourished, mostly due to the establishment and wide-spread adaptation of the OntoLex vocabulary[2] that expanded from its initial field of application from ontology lexicalization and the addition of linguistic information to general-purpose knowledge graphs to become a general community standard for machine-readable dictionaries on the web of data. With an increasing number of lexical data sets available as Linked Data or in RDF over the web, interest in *lexical* linking has been on the rise in the past years, e.g., in the Question-Answering over Linked Data challenges (QALD, since 2011)[3] or in the more series of Shared Tasks on Translation Inference Across Dictionaries (TIAD, since 2017).[4]

However, as far as lexical inference is concerned, all results we are aware of, be it on translation inference or the enrichment with multilingual labels, are concerned with precompiling links which are afterwards stored and distributed as novel or along with existing data sets. Curiously, the benefit of dynamicity, although being emphasized throughout the entire history of LLOD (Chiarcos et al., 2013; Cimiano et al., 2020), does not seem to have been explored for lexical data or cross-lingual linking, so far. We assume that this is mostly due to the fact that it is taken for granted, however, concrete applications of dynamic linking don't seem to ever have been brought forward.

With this paper, we aim to address this apparent gap. We describe the conjoint application of two web services, taking advantage of the federation capabilities of SPARQL to perform cross-lingual linking on the fly and thereby to enable querying over an English language knowledge graph using foreign language labels. Although this method is constrained by runtime considerations and thus largely restricted to string matching,[5]

[1] http://linguistic-lod.org
[2] https://www.w3.org/2016/05/ontolex/
[3] http://qald.aksw.org
[4] https://tiad2017.wordpress.com
[5] More advanced methods for lexical linking evaluate the wider lexicographical context, e.g., the number of pivot words that connect translation candidates (Lanau-Coronas and Gracia, 2020), but these are time-consuming analyses

we show that it provides a baseline functionality to enable the querying of knowledge graphs using foreign language query terms (words, or labels). This functionality, despite the noise it may introduce, is a valuable, and practically relevant option for low-resource languages for which no substantial knowledge graphs or ontologies exist, but whose speakers can then, for example, consult the English DBpedia in their own language.

We demonstrate this for two datasets and their associated web services (SPARQL end points): DBnary (Gilles Sérasset, 2012), described by Sérasset (2015), a machine-readable edition of Wiktionary data in RDF, and DBpedia (Auer et al., 2007), a machine-readable edition of Wikipedia data in RDF. However, instead of DBnary, any dictionary server could be used (e.g., the Apertium dictionaries hosted at UPM, (Gracia et al., 2018)), and instead of DBpedia, any knowledge graph (say, YAGO, (Suchanek et al., 2007)).

Overall, this paper is structured as follows: Sect. 2 discusses fundamentals of LLOD and RDF technology, Sect. 3 describes the use case. Then, Sect. 4 shows how we query DBpedia and DBnary. Finally Sect. 5 shows how we wrap up everything in one single federated query.

## 2. Linguistic Linked Open Data

As researcher specialised in Linguistic Linked Open Data, we are regularly faced with questions from other NLP or CL researchers on the advantages and drawback of using Semantic Web technologies to model, store or serve linguistic data. These questions are indeed justified as the Semantic Web approach seams to incur a steep learning curve and also may incur a higher workload on the resource publisher than on the resource consumer.

### 2.1. A Tree is a Graph, but a Graph is not Necessarily a Tree

Resource Description Format is the ground basis of the representation of Linked Open Datasets. RDF is not a language, but rather an abstract format that can be expressed using several syntax (one of which being indeed based on XML). RDF data is interpreted as a directed graph where (almost) each node has a name (an URI) and each arc is labelled using a relation name (also an URI).

As the data format is interpreted as a graph it's representation power is strictly higher than the representation power of a tree. As most existing Linguistic resources heavily use XML and seldom use XPaths/Xpointer to go beyond the basic XML tree structure, this leads to potentially more natural modelings for linguistic data.

One may argue that the OntoLex core vocabulary (McCrae et al., 2017) may mostly be viewed as a tree structure where root is the Lexicon, with LexicalEntry as children and Forms and LexicalSenses as grand-children. But you are able to provide additional information from other OntoLex vocabularies (e.g., OntoLex-VarTrans for Variation and Translation). For example, you can model a set of LexicalEntries/Forms/LexicalSenses using both the OntoLex model and give it additional structure with OntoLex-Lexicog (Bosque-Gil et al., 2017). OntoLex core will allow you to describe your LexicalEntries and LexicalSenses (and their relation to other ontologies or knowledge graphs) in a flat structure, while Lexicog model will allow you to precisely model the hierarchy of LexicalEntries/LexicalSenses as it was described in your original lexicon. As most of the nodes in both models are shared, the resulting structure may indeed be interpreted as 2 different trees covering the same node set.

### 2.2. No more Document Boundaries

By using URIs to name the nodes (and arcs) of the graph, each atomic part of your dataset is known outside of any file or document that may describe it. In essence, nodes in any RDF graph are globally defined and may be reused anywhere in the world. This is not the case for nodes of any XML files which may be shared between several documents only if the resource provider has given it a global name.

Moreover, the open world assumption clearly states that any document describing an entity can be complemented by any other source of information. Indeed, it is a common use case to describe the very same entity in different files or datasets. Among such use case are:

- by providing a set of relations between nodes from different datasets, one may link datasets together without being the producer of any of the linked datasets.
- in the DBnary dataset, the core of a dictionary (i.e. the lexical entries, word senses and canonical forms) is described in a dump file, while other dump files will complement lexical entries with all inflected forms or with translation links.

### 2.3. Standardizing Leads to Interoperability

The LLOD community is deeply involved in re-using common models for the publication of their linguistic data. As RDF allows for the extension of any vocabulary, it is easy for researcher to adopt a standard model even if some of its concepts are too coarse grained to be faithfully used for the description of a specific language.

One can easily refine the standard concepts by subclassing and provide a very detailed description of its lexicon. Then, consumers of the resource may be able to use the very fine-grain description or fallback to a coarser grained description for their specific use case.

that operate over large sets of complete dictionaries which are polynomial in time (over the size of the entire vocabulary). This is not an option here, as we require real-time performance, i.e., effectively linear lookup time.

### 2.4. Achieving Web Scale

When achieving the 5 stars of linked open data, each node in the LLOD graph has its own description available on the web (through it URI that is required to be resolvable through HTTP). Hence, consumers may be able to use LLOD data without necessarily having to import any dataset in their own database. Indeed any process that lacks knowledge on a specific entity may fetch it from the web.

Moreover, one can query different datasets through public SPARQL endpoints.

This means that the data available for your application goes far beyond what is available locally or in your own databases. In this paper, we show that it is possible to prototype a multilingual service without having any local database installed on premises.

## 3. Use Case: Cross-Lingual Querying of DBpedia

Our use case is the cross lingual querying of an ontology available in English. For this use case, we chose to query the English edition of DBpedia. Cross-linguality is achieved by querying DBnary, a multilingual dictionary available as Lexical Linked Open Data.

### 3.1. Knowledge graph: DBpedia

Since more than a decade, DBpedia is firmly established one of the most widely used general-purpose knowledge graphs in the web of data. At its core, it is automatically constructed from information provided in Wikipedia infoboxes. DBpedia started as a joint effort of researchers from Free University of Berlin and Leipzig University, Germany, in collaboration with OpenLink Software, and is now maintained by the University of Mannheim and Leipzig University. The first publicly available dataset has been made available in 2007 and published under the same license as the underlying Wikipedia information (CC-BY-SA), allowing others to reuse the dataset. DBpedia provides structured information extracted from Wikipedia pages and made available in a a uniform dataset which can be queried. As of June 2021, it contains over a trillion entities.

DBpedia has a broad scope of entities covering different areas of human knowledge. This allows external datasets to link to its concepts and has subsequently estalished DBpedia as a central hub in the web of data: The DBpedia dataset is interlinked with various other Open Data datasets on the Web, e.g., OpenCyc, UMBEL, GeoNames, MusicBrainz, CIA World Fact Book, DBLP, Project Gutenberg, Eurostat, UniProt, Bio2RDF, and US Census data.

DBpedia data can be queried via a public SPARQL endpoint under `https://dbpedia.org/sparql/`, which provides access to the underlying OpenLink Virtuoso data base.

### 3.2. Machine-readable dictionary: DBnary

The DBnary dataset (Gilles Sérasset, 2012) has grown steadily since its first description (Sérasset, 2012; Sérasset, 2015) and, at the time of writing, contains more than 275M relations describing 6.3M lexical entries in 22 languages. Its structure was originally based on lemon format but is now using the ontolex model.

The DBnary dataset now contains lexical data extracted from 22 wiktionary[6] language editions[7]. Up to now, DBnary used to only provide the wiktionary edition *endolexicon*, i.e. the subset of the wiktionary data that describe the language of the edition. That means that French language data is exclusively extracted from French language edition while English data was extracted from the English language edition. This choice was made so that data will achieve linguistic felicity as it is provided by the language's wiktionary community. Very recently DBnary is also providing the Wiktionary edition *exolexica*, i.e. all the lexical entries that do not belong to the edition's language. This means that many more languages may be described, but usually with a coarser grained description.

Translations are represented using an adhoc vocabulary based on the `dbnary:Translation` class which encodes a single translation from one of the extracted lexicon (endolex) to a target language. The translation entity is linked to a source lexical entry, but the target of the translation is encoded as a string, along with an entity representing the target language. Figure1 shows an example of such a translation.

```
fra:__tr_aze_1_animal__nom__1
    rdf:type dbnary:Translation ;
    dbnary:isTranslationOf
        fra:animal__nom__1 ;
    dbnary:targetLanguage lexvo:aze ;
    dbnary:writtenForm "heyvan"@az .
```

Figure 1: An example French to Azeri translation.

At the time of writing, the DBnary dataset contains 8.6M[8] such translations accounting for 22 source languages and 4396 different target languages. These number are constantly evolving as the DBnary dataset is extracted from Wiktionary everytime a dump is made available (i.e. twice a month). They should be compared with the 2.8M translations that where available in 2015. This changes in the datasets and in the whole LLOD cloud fully justify the use of dynamic approaches to lexical inferences.

[6] `http://wiktionary.org/`

[7] A language edition of wiktionary correspond to a site managed by its own community (e.g. `http://en.wiktionary.org` for English or `http://fr.wiktionary.org` for French). A language edition contains lexical entries in all possible language, with a description in the edition language.

[8] Exactly $8,619,352$ in the 20220401 semi-monthly extract from 1st April 2022, growing around 0.5% every month.

The DBnary dataset chose not to use ontolex `vartrans` (Bosque-Gil et al., 2015) by default as it is designed to link existing lexical entries through translation relations. In the case of DBnary, we do not have lexical entries in all the target languages and we chose not to adopt the LexVo (de Melo, 2015) attitude consisting in crafting a URI for every term in a language, as we are not guarantied that the value of a translation would qualify as a legitimate lexical entry in the target language (indeed, some translations are sometimes inflected forms or explanations rather than fully legitimate terms).
Note that translations from/to the 22 extracted languages are additionally represented using `vartrans` when the translation string can be linked to a lexical entry for the correct Part Of Speech, provided that there are no homonymy in the target language.
DBnary data can be queried via a public SPARQL endpoint under `http://kaiko.getalp.org/sparql`. Like DBpedia, this operates over an OpenLink Virtuoso data base.

## 4. Querying one end point at a time

### 4.1. SPARQL

In its current version 1.1, the SPARQL Protocol and RDF Query Language (SPARQL) [9] provides a standard for querying and manipulating RDF graph data over the Web or in an RDF store. SPARQL 1.1 defines a query language, result formats, update language, protocol and web service specifications. Features that set it apart from general query languages for graph data in general include query federation (accessing and integrating data from multiple remote end points at query time), entailment regimes (the possibility to infer implicit statements from an ontology associated with the data) as well as its orientation towards processing RDF data, i.e., a generic directed labelled multigraph characterized by using URIs (rather than internal IDs or strings) to denote nodes and edges, which can be serialized in or read from numerous formats (including, but not limited to, XML (Beckett and McBride, 2004), (X)HTML (Adida et al., 2008), JSON (Sporny et al., 2014), CSV (Ermilov et al., 2013), RDBMS (Dimou et al., 2014), as well as native RDF sources represented by Turtle (Beckett et al., 2014), HDT (Fernández et al., 2013), RDF-Thrift (Käbisch et al., 2015) or web services). If an RDF data set uses URIs that resolve via the HTTP prototocol to other RDF data, this constitutes Linked Data (Bizer et al., 2011), and linked data technology can been used to develop and to refer to widely used standards and community standards such as for knowledge graphs and ontologies – e.g., SKOS (Miles and Bechhofer, 2009), RDFS (McBride, 2004) and OWL (Antoniou and Harmelen, 2004) – as well as lexical data and other linguistic information – e.g., using SKOS-XL (Miles and Bechhofer, 2009), LexInfo (Cimiano et al., 2011), and OntoLex-Lemon (McCrae et al., 2017). With shared vocabularies described by resolvable URIs, Linked Data provides explicit, machine-readable semantics for its data structures that can be consulted at query time, and beyond shared vocabularies, the federation mechanism allows to also consult, retrieve and integrate information from different data providers.
Here, we focus on aspects of querying, the following section illlustrates federated search. In general, a SPARQL select query consists of a number of key words, including `PREFIX` (namespace declarations), `SELECT` (query operator), `FROM` (data source), and `WHERE` (graph pattern). The `WHERE` block contains the actual query, expressed using Turtle-style statements extended with variables, XPath-style functions for filtering and binding and operators such as conjunction (`.`), grouping (`{...}`), disjunction (`UNION`), negation (`MINUS`), optional statements (`OPTIONAL`), as well as the possibility to include embedded `SELECT` statements and to address named data sources (`GRAPH`) and remote end points (`SERVICE`). SPARQL contains a number of extensions over this basic model, including the possibility to not only query for individual statements, but also to formulate complex patterns over sequences of statements by means of SPARQL property paths.

### 4.2. Relations: DBpedia

For illustrating a general-purpose queries against a knowledge graph, imagine a simple question-answering setup in which we want to consult DBpedia to return (a human-readable label for) the type of entity a user enters. Say, for the query 'What is a horse?' (or, more briefly, 'horse'), we expect it to return 'animal'. The query itself, shown in figure 2, needs to be constructed on the basis of the RDF vocabularies used in DBpedia, but it contains a variable part with the actual search term, here, `"horse"@en`, i.e., *horse* in English:[10]
This query can be executed against the DBpedia end point.[11] It has a number of pecularities, so, the search term must be upper cased, also, we eliminate technical (W3C) terms that just describe the data model, and finally, we want to restrict the results (`?category`) to English labels. If these things are being respected, this query returns *animal* and *personal function*[12].
Another possible strategy is, instead of returning an English language label, to parse the local name of the URI

[9] `https://www.w3.org/TR/sparql11-overview/`

[10] Note that from future queries, we omit prefix declarations for reasons of space. All of the namespaces used can be retrieved using `http://prefix.cc/`. For example, `http://prefix.cc/dbp` will return `http://dbpedia.org/property/` for the namespace `dbp:`.

[11] `https://dbpedia.org/sparql`

[12] The latter value comes as a surprise, but it possibly comes from an logical inference based on the fact that 4 persons had "Horse" as their "function" in DBpedia

```
SELECT distinct ?category
WHERE {
  ?a rdfs:label "Horse"@en.
  ?a rdf:type ?type.
  FILTER(!strstarts(str(?type),
            'http://www.w3.org'))
  ?type rdfs:label ?category.
  FILTER(lang(?category)="en")
} LIMIT 10
```

Figure 2: Querying all categories an article labelled "Horse"@en belongs to.

as illustrated in Figure 3.

```
SELECT distinct *
WHERE {
  ?a rdfs:label "Horse"@en.
  ?a rdf:type ?type.
  FILTER(!strstarts(str(?type),
           'http://www.w3.org'))
  BIND(replace(str(?type),".*[/#]","")
         as ?localname)
  BIND(lcase(replace(?localname,
               "([a-z])([A-Z])","$1 $2"))
          as ?category)
}
```

Figure 3: Getting the name of the category by lower casing its URI localname.

As we cannot guarantee that URIs are not human-readable or in any particular language, the latter is usually discouraged, in this case, however, it retrieves additional categories that were not associated to an English label from an external vocabulary: *biological living object*, *eukaryotic cell* and *mammal*.

## 4.3. Translations: DBnary

There are several ways to query translations from the DBnary dataset.

### 4.3.1. Querying Translations Through Ontolex `vartrans`

Standard multilingual OntoLex modeled lexical datasets may be queried using `vartrans` ontolex extension. In `vartrans`, translations are represented by linking 2 lexical entries through `vartrans:translatableAs` relation. Figure 4 shows the corresponding query which looks for 2 lexical entries, related with this relation (in either direction), one of which is the term "Stadt"@de we want to translate.

This query will return *city*, *town*, *center*, *centre*, *stead* and *independent city* (where the two latter come from an inverse translation relation).

Such a query may be used on other datasets like Apertium or ACoLi and it indeed works on DBnary, but this will restrict our use case to the 22 language editions

```
SELECT DISTINCT * WHERE {
  ?source
    ontolex:canonicalForm/
      ontolex:writtenRep "Stadt"@de;
    (vartrans:translatableAs|
        ^vartrans:translatableAs)/
      ontolex:canonicalForm/
      ontolex:writtenRep ?translation.
    FILTER(lang(?translation)="en")
}
```

Figure 4: Querying translation from and to German term "Stadt" using `vartrans` modeling.

of DBnary (as these relations require a lexical entry at both end of the relation). This accounts for 3.5M available translation relations and 22 source/target languages while DBnary contains 8.6M translations to 4396 languages.

### 4.3.2. Querying Translations using DBnary's `Translation` class

As most translations in DBnary are encoded as `dbnary:Translation` instances. We can look for English translations of German entries or, in reverse, for German translations of English entries. Figure 5 retrieves translations for German *Pferd*.

```
SELECT distinct ?translation
WHERE {{
  ?t dbnary:isTranslationOf/
       ontolex:canonicalForm/
       ontolex:writtenRep "Pferd"@de;
     dbnary:writtenForm ?translation.
  } UNION {
  ?t dbnary:writtenForm "Pferd"@de;
     dbnary:isTranslationOf/
       ontolex:canonicalForm/
       ontolex:writtenRep ?translation.
  }
  FILTER(lang(?translation)="en")
}
```

Figure 5: Looking for translations from and to German *Pferd* using DBnary translations modeling.

This particular query returns *horse*, *equidae*, *knight*, *vaulting horse*, *vault*, *equine* and *horsy* and is equivalent to the preceding one for German to English, but it is now possible to query from languages that are not part of the 22 DBnary language edition. For instance, querying for translations of Romanian *cal* leads to *knight* and *horse*.

### 4.3.3. Querying translation using a pivot strategy

Previous strategies look for a translation instance from or to English. However, the DBnary dataset contains many more indirect translation links. The idea here is to find a translation relation from our query language to English by pivoting through one lexical entry of any

of the DBnary languages. Figure 6 shows such a query again for German term *Pferd*.

```
SELECT distinct ?translation
WHERE {
  ?entry ^dbnary:isTranslationOf/
          dbnary:writtenForm "Pferd"@de.
  ?t dbnary:isTranslationOf ?entry;
     dbnary:targetLanguage lexvo:eng;
     dbnary:writtenForm ?translation.
}
```

Figure 6: Translations of German *Pferd* pivoting on any lexical entry.

Executed against the DBnary SPARQL end point, this retrieves a total of 49 translations, while querying for Romanian *cal* leads to 46 translations.

## 5. Cheap and dirty cross-lingual querying by federated search

The idea of cheap and dirty cross-lingual querying is to use the functionalities illustrated above to translate labels from the user language to English, to extrapolate the expected DBpedia labels (by doing upper case conversion), to retrieve category labels from DBpedia and then to use DBnary to translate these back into the user language.

Note that we use German as an example only – and more precise results would be expected if we just query German DBpedia labels –, but this approach can be applied to *any* language for which lexical data in OntoLex is provided, and in particular, low resource languages. To a considerable extent, these are covered by DBnary, and as it is regularly updated from Wiktionary which is a crowd-sourced resource, its coverage is continuously increasing, but other portals provide OntoLex-compliant lexical data, as well, e.g., bilingual dictionaries from the Apertium project (Gracia et al., 2018), the GlobalWordNet family of resources (McCrae et al., 2021) or the ACoLi Dictionary Graph (Chiarcos et al., 2020).

In this case, both end points run on independent installations of the same database management system, however, it is important to note that no provider-specific technology is being used, but that we only rely on standardized, portable SPARQL 1.1 functionalities. In particular, this includes the keyword `SERVICE` which allows to consult an external SPARQL end point at query runtime.

Using the `SERVICE` keyword, it is possible to consult an external SPARQL end point (or another webservice) when running a local SPARQL query.[13] For example, as shown in figure 7, we can call DBpedia from DBnary, i.e., we first translate German to English, adjust the result to match the upper case convention of DBpedia labels, then query DBpedia and then translate the results back to German.

```
SELECT distinct ?result
       (count(distinct *) as ?confidence)
WHERE {
  ?entry ^dbnary:isTranslationOf/
          dbnary:writtenForm "Pferd"@de.
  ?t dbnary:isTranslationOf ?entry;
     dbnary:targetLanguage lexvo:eng;
     dbnary:writtenForm ?translation.

  BIND(concat(
         ucase(substr(?translation,0,1)),
         substr(?translation,2))
       as ?dbp_label)

  SERVICE <https://dbpedia.org/sparql> {
    ?a rdfs:label ?dbp_label.
    ?a rdf:type ?type.
    FILTER(!strstarts(str(?type),
            'http://www.w3.org'))
    ?type rdfs:label ?category.
    FILTER(lang(?category)="en")
  }

  ?t2 dbnary:isTranslationOf ?entry2;
      dbnary:targetLanguage lexvo:eng;
      dbnary:writtenForm ?category.
  ?entry2 dct:language lexvo:deu;
      ontolex:canonicalForm/
        ontolex:writtenRep ?result;
      a lexinfo:Noun
} ORDER BY desc(?confidence) asc(?result)
  LIMIT 10
```

Figure 7: Federated query calling DBpedia from DBnary after translations has been queried.

We return nominal concepts only, and with this particular query, we also calculate confidence, i.e., the number of paths (English translations, DBpedia concepts) that will lead to a particular translation. This is a very useful feature as the multitude of paths can lead to unexpected associations, and these can be detected as possible but unlikely.

Results of the query above are shown in Tab. 1. The top-level match is the expected result, *Tier* 'animal', *Couleur*, *Farbe*, *Farbton* 'color' and *Beere* 'berry' refer to types of horses designated by characteristics of their color, *Leut* 'people', *Mensch* 'human' and *Person* 'person' originate in the DBpedia concept *dpo:PersonalFunction* – this seems to reflect the sense of a 'workhorse' which can be metaphorically extended to people. It is less clear where *Chaot* 'slob' and *Gelähmter* 'paralyzed' originate from.

As this requires substantial aggregation, this is not really cheap, yet, but we can speed it up by restricting

[13] For security reasons and load balancing, this functionality may be disabled. It is, however, part of the SPARQL specification and should be supported by all SPARQL 1.1 compliant RDF stores.

| result | confidence |
|---|---|
| "Tier"@de | 271 |
| "Couleur"@de | 28 |
| "Chaot"@de | 22 |
| "Gelähmter"@de | 20 |
| "Leut"@de | 20 |
| "Mensch"@de | 20 |
| "Person"@de | 20 |
| "Farbe"@de | 14 |
| "Farbton"@de | 14 |
| "Beere"@de | 4 |

Table 1: Quick and dirty cross-lingual search: German category labels for German *Pferd* 'horse' retrieved from DBnary and the English DBpedia

the number of responses and suppressing aggregation as shown in figure 8

```
SELECT ?result WHERE {
  {SELECT DISTINCT ?translation
   WHERE {
     ?e ^dbnary:isTranslationOf/
         dbnary:writtenForm "Pferd"@de.
     ?t dbnary:isTranslationOf ?e;
        dbnary:targetLanguage lexvo:eng;
        dbnary:writtenForm ?translation.
   } LIMIT 3
  }

  BIND(concat(
         ucase(substr(?translation,0,1)),
         substr(?translation,2))
       as ?dbp_label)

 SERVICE <https://dbpedia.org/sparql> {
  SELECT DISTINCT
     ?dbp_label ?type ?category
  WHERE {
    ?a rdfs:label ?dbp_label.
    ?a rdf:type ?type.
    FILTER(!strstarts(str(?type),
            'http://www.w3.org'))
    ?type rdfs:label ?category.
    FILTER(lang(?category)="en")
  } LIMIT 2
 }

?t2 dbnary:isTranslationOf ?entry2;
    dbnary:targetLanguage lexvo:eng;
    dbnary:writtenForm ?category.
?entry2 dct:language lexvo:deu;
    ontolex:canonicalForm/
      ontolex:writtenRep ?result;
    a lexinfo:Noun
} LIMIT 1
```

Figure 8: Restricting the number of translation and suppressing aggregation.

It is to be noted that SPARQL query results are considered to be unsorted. Accordingly, the confidence we explicitly measured in the last query translates to a *probability* that the result of this query is correct. With *Tier* returned via 271 different paths in the last query from a total of 433 paths, and considering *Tier* to be the only correct response, the precision *for this particular query* would be at 62.5%.

## 6. Outlook

We have shown how two SPARQL webservices can be used in conjunction to perform cross-lingual search using search terms in one language, a knowledge graph in another, and producing results in the search language. We would like to note that this is not a novel functionality, but provided by standard SPARQL technology, albeit one which does not seem to have been documented in LLOD literature before. As such, our contribution is not so much innovative as it fills a gap in the current scientific documentation of LLOD practices and possibilities that can serve as a template for the development of future applications.

The main purpose of this submission is to show that SPARQL allows to stack web services and RDF resources in a meaningful, and specifically, to enable cheap and dirty cross-lingual querying. It's main advantages is that (1) it allows for crafting cheap mockups of cross-lingual services and (2) it fully uses dynamicity and the quality of the service will evolve with the quality/coverage of the resources available in the cloud.

It is clear that in this context, methods that compile static links are superior in quality,[14] but they require designated development time, substantial preprocessing and hosting of generated links, whereas the benefit of this method is that it is immediately applicable to *any* language for which a bilingual OntoLex dictionary can be found that either provides English translations or a link to another bilingual dictionary that does. At the time of writing, hundreds of such dictionaries are available online, e.g., from the OntoLex edition of PanLex (Kamholz et al., 2014) available from the ACoLi Dictionary Graph (Chiarcos et al., 2020).[15]

[14] Due to the trade-off between speed and quality, the comparative performance between static dictionary induction and dynamic methods is relatively hard to evaluate: Static methods are optimized towards managing and minimizing noise in translations, and they achieve this by aggregating confidence scores over the lexical content of pivot translations in larger lexical knowledge graphs. As shown in Tab. 1, aggregation is possible in our approach as well, but it comes at the price of processing speed, and for responses calculated on-the-fly, there are limitations as to how much context can be inspected. Our approach will produce optimal results (in terms of speed) if the number of pivot languages (or pivot translations) is limited. For static methods, where speed at query time is eliminated as a factor (i.e., reduced to a lookup), best results (in terms of quality) will be achieved if multiple pivot languages (or pivot translations per word) are available.

[15] https://github.com/acoli-repo/

## Acknowledgements

The research described in this paper has been initiated in the context of the Cost Action *Nexus Linguarum. European Network for Web-Centered Linguistic Data Science* (CA18209). The work of the first author has been conducted in the context of the H2020 Research and Innovation Action *Prêt-à-LLOD. Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors* (2019-2022, grant agreement 825182).

## 7. Bibliographical References

Adida, B., Birbeck, M., McCarron, S., and Pemberton, S. (2008). RDFa in XHTML: Syntax and processing. Technical report, W3C Recommendation.

Antoniou, G. and Harmelen, F. v. (2004). Web Ontology Language: OWL. In *Handbook on ontologies*, pages 67–92. Springer.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer.

Beckett, D. and McBride, B. (2004). RDF/XML syntax specification (revised). Technical report, W3C recommendation.

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., and Carothers, G. (2014). RDF 1.1 Turtle. *World Wide Web Consortium*, pages 18–31.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global.

Bosque-Gil, J., Gracia, J., Aguado-de Cea, G., and Montiel-Ponsoda, E. (2015). Applying the OntoLex Model to a multilingual terminological resource. In Fabien Gandon, et al., editors, *The Semantic Web: ESWC 2015 Satellite Events*, pages 283–294, Cham. Springer International Publishing.

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. In *LDK Workshops*, pages 74–84.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL Traitement Automatique des Langues*, 52(3):245–275.

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., Fäth, C., and Ionov, M. (2020). The ACoLi dictionary graph. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3281–3290.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer Nature.

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

Dimou, A., Vander Sande, M., Slepicka, J., Szekely, P., Mannens, E., Knoblock, C., and Van de Walle, R. (2014). Mapping hierarchical sources into RDF using the RML mapping language. In *2014 IEEE International Conference on Semantic Computing*, pages 151–158. IEEE.

Ermilov, I., Auer, S., and Stadler, C. (2013). CSV2RDF: User-driven CSV to RDF mass conversion framework. In *Proceedings of the ISEM*, volume 13, pages 04–06.

Fernández, J. D., Martínez-Prieto, M. A., Gutiérrez, C., Polleres, A., and Arias, M. (2013). Binary RDF representation for publication and exchange (HDT). *Journal of Web Semantics*, 19:22–41.

Gracia, J., Villegas, M., Gomez-Perez, A., and Bel, N. (2018). The Apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Käbisch, S., Peintner, D., and Anicic, D. (2015). Standardized and efficient RDF encoding for constrained embedded networks. In *European Semantic Web Conference*, pages 437–452. Springer.

Kamholz, D., Pool, J., and Colowick, S. (2014). PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150.

Lanau-Coronas, M. and Gracia, J. (2020). Graph exploration and cross-lingual word embeddings for translation inference across dictionaries. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 106–110.

McBride, B. (2004). The Resource Description Framework (RDF) and its vocabulary description language RDFS. In *Handbook on Ontologies*, pages 51–65. Springer.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

McCrae, J. P., Goodman, M. W., Bond, F., Rademaker, A., Rudnicka, E., and Da Costa, L. M. (2021). The GlobalWordNet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

Miles, A. and Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation*.

Pareja-Lora, A., Lust, B., Blume, M., and Chiarcos, C. (2020). *Development of Linguistic Linked Open*

acoli-dicts

*Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. The MIT Press.

Sérasset, G. (2012). DBnary: Wiktionary as a LMF based Multilingual RDF network. In *Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey, May.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6(4):355–361.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2014). JSON-LD 1.0. Technical report, W3C recommendation.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

## 8. Language Resource References

Gilles Sérasset. (2012). *DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF*. Univ. Grenoble Alpes, ISLRN 023-163-901-149-4.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 61–68
Marseille, 24 June 2022


# Spicy Salmon: Converting between 50+ Annotation Formats with Fintan, Pepper, Salt and Powla

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

**Abstract**

Heterogeneity of formats, models and annotations has always been a primary hindrance for exploiting the ever increasing amount of existing linguistic resources for real world applications in and beyond NLP. Fintan - the Flexible INtegrated Transformation and Annotation eNgineering platform introduced in 2020 is designed to rapidly convert, combine and manipulate language resources both in and outside the Semantic Web by transforming it into segmented RDF representations which can be processed in parallel on a multithreaded environment and integrating it with ontologies and taxonomies. Fintan has recently been extended with a set of additional modules increasing the amount of supported non-RDF formats and the interoperability with existing non-JAVA conversion tools, and parts of this work are demonstrated in this paper. In particular, we focus on a novel recipe for resource transformation in which Fintan works in tandem with the Pepper toolset to allow computational linguists to transform their data between over 50 linguistic corpus formats with a graphical workflow manager.

**Keywords:** Interoperability, Transformation, Annotation, Corpora, Ontologies, Linguistic Linked Open Data

## 1. Entree

With the continued rise of corpus technologies in language sciences and lexicography that we have seen in the last decades, the number and diversity of linguistic annotations has been growing at an exponential rate – and this trend still continues. At the same time, the increasing maturity of language technology and machine learning and their spread to novel domains calls for ever increasing amounts of homogeneous training and evaluation data, so that a core challenge of applied NLP is, in fact, not so much to come up with innovative algorithms, but to secure the availability and consistency of the data needed for applying state-of-the-art technology.

Indeed, the heterogeneity of formats, models and annotations has always been a primary hindrance for exploiting the ever increasing amount of existing linguistic resources for real world applications in and beyond NLP. The Linguistic Linked Open Data (LLOD) community has a decade-spanning history of creating community standards for homogeneous data publication and interlinking. Relying on Semantic Web technology, Fintan - the Flexible INtegrated Transformation and Annotation eNgineering platform (Fäth et al., 2020) has been designed as a tool to rapidly convert, combine and manipulate heterogeneous language resources in a generic, sustainable and scalable way. Its transformation and export capabilities are tailored towards, but not limited to commonly used LLOD vocabularies such as CoNLL-RDF (Chiarcos and Fäth, 2017) or Ontolex Lemon (Cimiano et al., 2016) and thus alleviate generation of LLOD datasets and their integration into NLP workflows.

Fintan has recently been extended with a set of additional modules increasing the amount of supported non-RDF formats and the interoperability with existing non-JAVA conversion tools. In this paper, parts of this work are demonstrated. In particular, we focus on a novel recipe for resource transformation in which Fintan works in tandem with the Pepper (Zipser and Romary, 2010) toolset to allow computational linguists to transform their data between over 50 linguistic corpus formats with a graphical workflow manager.

## 2. The Fish: Fintan

Although Fintan as a tool doesn't have anything fishy about it, the acronym is actually coined as a metaphor to its generic, variable design. Fintan mac Bóchra in Irish folklore was a shape-shifting sage who survived a great flood in the shape of a salmon (Macalister, 1941), which is also reflected in the logo of the Fintan platform (cf. Fig. 1).

### 2.1. Software design

Fintan is designed to adapt to the flood of data and formats computational linguists are confronted with and also enable users to integrate their data with a wealth of resources from the Semantic Web. This is also reflected by the internal architecture which heavily relies on Semantic Web Standards such as SPARQL (Buil Aranda et al., 2013) and encourages users to take advantage of graph-based transformation capabilities while adhering to the following principles:

- Fintan is *generic* in that it builds on the transformation of linguistic annotations, lexical data structures, etc. into labeled directed multi-graphs and back. In particular, this can represent every type of linguistic annotation (Bird and Liberman, 2001).

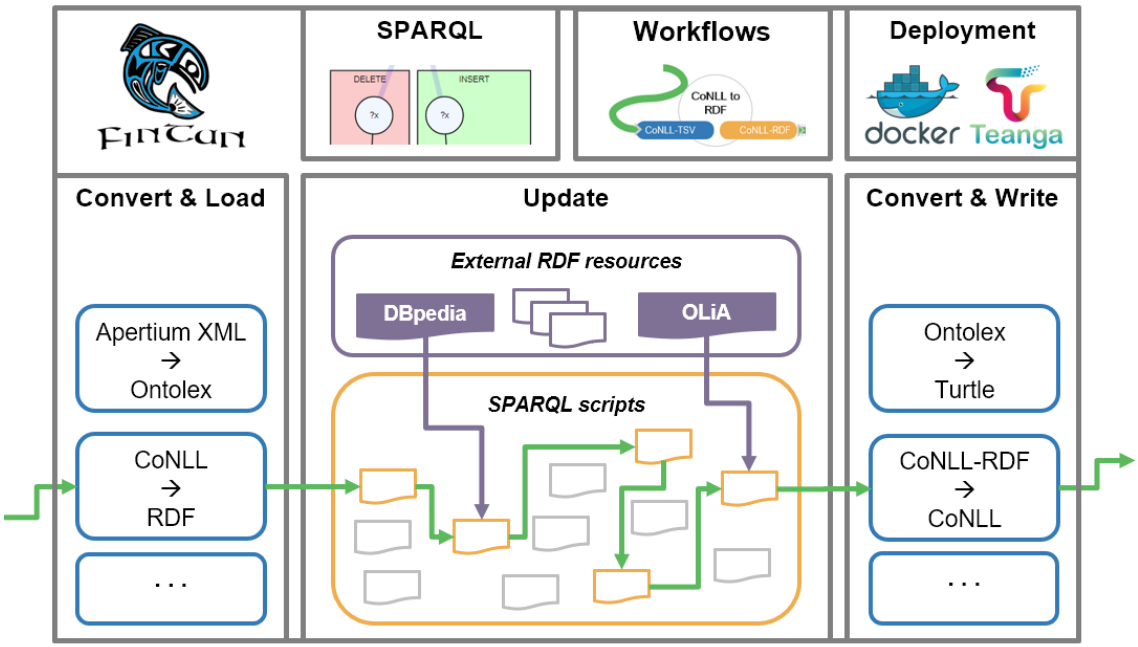


Figure 1: The Fintan plattform

- Fintan is *sustainable* in that it builds on web standards (RDF) and standardized, declarative transformations (SPARQL) for representing and manipulating these graphs. In particular, the SPARQL scripts can be run fully independently from the current Fintan code base, but against any SPARQL end point or with any programming language for which a SPARQL library is available.

- Fintan is designed to be *scalable*: It provides parallelized stream processing. Fintan hereby takes advantage of the inherently localized structure of most linguistic data which can often be divided into self-contained segments like entries in dictionaries or sentences, tokens etc. in corpora. By splitting data in such a way, we can process multiple segments at the same time while also minimizing memory consumption during query execution resulting in increased scalability, stability and faster execution.

Figure 1 provides an overview of the general software architecture and its *modular* structure. Within Fintan, data is transformed and streamed between components each providing specific processing capabilities:

- `Loader` components prepare data for segmented processing.

- `Updaters` apply SPARQL scripts on data segments, optionally also relying on external LLOD resources such as OLiA (Chiarcos and Sukhreva, 2015) for transforming linguistic annotations or DBpedia (Mendes et al., 2012) for entity linking.

- `Writer` components producing RDF serializations such as Turtle or exporting tabular formats like CoNLL.

Additional transformer components may also take other script languages to process various types of data, e.g. XSL for converting XML data as has been applied in a complex workflow enabling the Apertium bilingual dictionaries to be used for cross-lingual transfer learning in the pharmaceutical domain (Gracia et al., 2020). The CoNLL-RDF (Chiarcos and Fäth, 2017) library, Fintan's spiritual predecessor, is now also a native part of the toolchain allowing Fintan to directly execute any existing CoNLL-RDF pipeline.
By treating transformation scripts and pipeline configurations as data fed into standardized transformer components, we also open possibilities to increase *reusability*. Some scripts (like annotation transformation) may be applicable in multiple workflows for several types of resources including lexical and corpus data alike, albeit this is highly dependent on how users structure their pipeline configurations.
To alleviate this design process, Fintan also features a stand-alone graphical workflow manager which renders existing components as processing nodes which can be connected by edges reflecting data streams (cf. Fig. 4). Streams are hereby distinguished between unsegmented text streams and streams of pre-loaded RDF segments.

## 2.2. OpenAPI support

Pipelines created with the workflow manager can directly be inserted and run in the Fintan JAVA backend on a shell environment. However, they can also be

exported as dockerized[1] web services to be integrated into decentralized complex workflows. To achieve this we built a Python server which can expose the Fintan backend as an OpenAPI[2] compliant web service and provides functionality to upload scripts and data and to run pipeline configurations. While this server can be run stand alone, the Workflow manager can also create a makefile to directly build an integrated Docker container containing all relevant data and a specific pipeline configuration. The make script includes all relevant code and resources for automated deployment.
However, OpenAPI support is not just limited to deployment of workflows. Instead, we recently integrated API functions which allow external web services to be run as part of Fintan pipelines. This API is mostly based on the Swagger Code Generator[3] and has originally been created specifically for the use case described in this paper, but has been slightly redesigned to host generic services by exposing most configuration options (i.e. request methods etc.) as parameters in the Fintan JSON configuration. The generic wrapper component is however structurally limited to single requests per transformation as more complex operations would require specific treatment (possible wait operations, poll for success etc.) which are highly service specific. Such peculiarities must be addressed by implementing service specific wrapper components using the Fintan API[4].
This not only allows to address existing web services but also provides a means to wrap other toolsets which are structurally incompatible or not natively available in JAVA directly within Fintan.

## 3. The Spices: Salt and Pepper

Language resource interoperability is, indeed, a problem that has received a lot of attention over the years, and as far as linguistic annotations are concerned, the state of the art in this regard is represented by the Linguistic Annotation Framework (Ide and Suderman, 2014, LAF). LAF defines an abstract data model (a generic labelled directed acyclic multi-graph over streams of primary data) and an XML syntax (GrAF), and it is designed to be able to represent any linguistic annotation.
In terms of processing tools, however, very few technology seems to be around that actually implements LAF/GrAF directly.[5] Instead, real-world tools use a

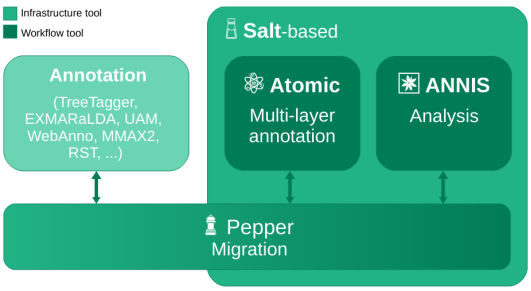


Figure 2: Interdependencies in the Salt/ANNIS universe of tools, taken from Druskat et al. (2016).

number of derivative formats (usually with less generic, but otherwise equivalent features), a notable example being the JSON-LD based LAPPS Interchange Format (Verhagen et al., 2015, LIF) developed by the creators of LAF. One such application is the converter suite Salt'n Pepper, to some extent overlapping in spirit to Fintan, but firmly integrated in its own, independent technological ecosystem

### 3.1. ANNIS and Salt ecosystem

ANNIS is a corpus management system specifically designed for dealing with multi-layer corpora (Dipper et al., 2004). With this goal in mind, its developers adopted early drafts of the LAF standard (Ide and Romary, 2004) as the underlying data model and developed a large set of converters from and to various established corpus formats.
ANNIS provides search and visualization capabilities for of multi-layer corpora, it supports annotations of different types (spans, trees with and without labelled edges, dependencies, arbitrary pointing relations), internally represented as a directed acyclic multi-graph. ANNIS comes with convenient visualizations and with different backend implementations as in-memory database (ANNIS1), relational database (ANNIS2, ANNIS3) and an experimental graph backend (Krause et al., 2016, GraphANNIS). The primary input format to ANNIS is PAULA XML (Sect. 4.1), and the underlying PAULA Object Model represents the basis for a small universe of interconnected tools, specifications and resources (Chiarcos et al., 2008; Druskat et al., 2016), see Fig. 2 for an architectural overview. In addition to ANNIS, these include

- PAULA (XML format and abstract data model)
- AQL (ANNIS Query Language)
- Salt (Java API and theory-neutral meta model)
- Pepper (converter suite)
- Atomix/Hexatomic (annotation tool)
- Laudatio (corpus repository)

[1] https://www.docker.com/
[2] https://www.openapis.org/
[3] https://swagger.io/tools/swagger-codegen/
[4] A detailed description on how to create and integrate custom components into Fintan is available here: https://github.com/acoli-repo/fintan-doc/blob/master/3c-build-custom-components.md
[5] There are a number of third-party converters, but the only tool natively building on LAF/GrAF seems to be the ANC-Tool (Suderman and Ide, 2006).

ANNIS operates a powerful, tagset-independent and theory-neutral meta model, and with its reference implementation in the Salt API, it allows for storing, manipulating, and representing nearly all types of linguistic data.

### 3.2. Pepper plattform

The Pepper (Zipser and Romary, 2010) plattform and its internal theory-neutral Salt meta model compose a framework which enables means of direct conversion between at least 20 formats for annotated corpora including EXMARaLDA, Tiger XML, MMAX2, RST, TCF, TreeTagger format, TEI (subset), PAULA and more[6]. Pepper's general architecture is partly reminiscent of Fintan in that it is based on Java and Maven and divides its processing steps into *Importers* (corresponding to Fintan's Loaders), *Manipulators* (corresponding to Fintan's Transformers and Updaters) and *Exporters* (corresponding to Fintan's Writers). However, the implementation and design principles differ in many regards:

- Pepper uses the Salt model as an internal abstraction layer for the processed data. This introduces compatibility between modules but also narrows the aim towards corpora and may result in a loss of unsupported pieces of information stored in the original data. Fintan on the other hand is completely format-independent.
- Pepper focuses on fully designed converters as modules. Fintan instead emphasises on atomic reusable operations, e.g. by allowing users to directly load and manipulate transformation scripts for components.
- Since Pepper is focused on corpora, it cannot easily replicate Fintan's ability to side-load ontologies or external RDF repositories for Annotation Engineering and resource enrichment tasks.

As a corpus transformation tool, Pepper, nevertheless, is a valuable addition to Fintan's portfolio and extends its coverage of corpus formats. Because of the structural similarities, we were first considering a direct integration as a native Java library, however there were some drawbacks to consider:

- Pepper exclusively uses file I/O and is not natively streamable without major refactoring or caching.
- Pepper uses the OSGi framework while Fintan operates on native Java and Apache Jena, thus introducing additional complexity and risks when trying to directly map Pepper modules as Fintan components.
- Pepper needs a lot of additional module data (bloating a possible direct integration into Fintan's backend)

For these reasons, we decided to treat Pepper as a stand-alone converter module and wrapped it into a dockerized OpenAPI service which is specifically designed to generate POWLA-RDF (cf. Sect. 4.2) data from any corpus format supported by a Pepper Importer. This service can be accessed within Fintan workflows using the OpenAPI transformer component.

## 4. The Scullions: PAULA and POWLA

PAULA is an abstract data model for the ANNIS query language AQL, underlying the Salt API and the PAULA XML format, but also POWLA, an OWL2/DL data model for linguistic annotation on the web (Chiarcos, 2012), and in terms of its expressivity, it is equivalent (but slightly older than) the ISO-standardized Linguistic Annotation Framework (LAF).

### 4.1. PAULA

PAULA's underlying data model, much like RDF, is represented by labeled directed acyclic (multi)graphs (DAGs) and thus contains various types of *nodes*, *labels* and *edges*:

- *Nodes* are distinguished between *terminals* (tokens or spans of characters in the source data), *markables* (flat, positional annotations referring to spans of terminals) and *structs* (functioning as structural parents to other nodes in a tree).
- *Edges* can thus be *dominance relations* (parent to child in structs) or simple *pointing relation* (directed, but without hierarchical implications).
- *Labels* can be attached to nodes and edges alike representing linguistic annotations.

### 4.2. POWLA

The POWLA vocabulary shown in Figure 3 is an OWL2/DL implementation of the PAULA Object Model and preserves similar data structures for linguistic annotations, as an example, syntactic tree structures are rendered in POWLA by means of:

- `powla:Node` for tokens and phrasal nodes,
- `powla:hasParent` for hierarchical relations between nodes,
- `powla:next` for sequential relations between nodes, and
- `powla:Relation` (with `powla:hasSource` and `powla:hasTarget`) for labelled edges.

PAULA thus has a high level of compatibility with existing Fintan workflows, as its OWL2/DL implementation POWLA is used in conjunction with CoNLL-RDF

[6] A full list of "known" modules is provided by the developers: `https://corpus-tools.org/pepper/knownModules.html`

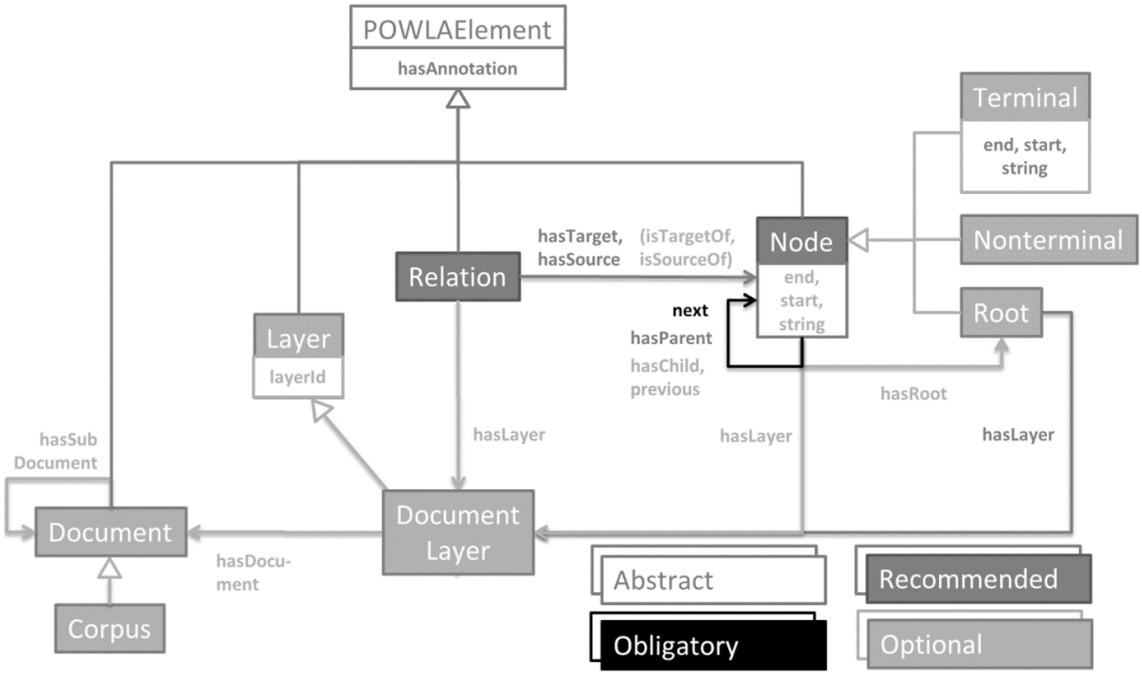


Figure 3: The POWLA vocabulary

(Chiarcos et al., 2021) to model linguistic data structures that exceed beyond labels of or pointers between individual words (Chiarcos and Glaser, 2020). PAULA and POWLA thus form a natural technological bridge between Pepper and Fintan, and here, they are being used to connect both technologies.

## 5. Main Course: Converting 50 Formats

Now that we have all ingredients and helpers ready, we can start concocting our workflow.

### 5.1. Preparations and Workflow

Since POWLA is a supported format in the CoNLL-RDF toolset (Chiarcos and Glaser, 2020) it can be converted to CoNLL-RDF by a set of SPARQL updates:

- Token level annotations in the form of `conll:COL` for typical columns such as `WORD`, `POS` etc. are derived from PAULA labels on markables and terminals.
- Dependencies are derived from dominance relations.
- Since POWLA does not necessarily annotate sentence boundaries, CoNLL-RDF's `nif:Sentence` nodes also need to be derived from the hierarchical data structure in order to produce a fully delimited CoNLL corpus. In case this fails or produces inconsistent output (e.g. if the corpus is annotated with a lot of cross-sentence relations), we can optionally split sentences purely based on punctuation in a post processing step

The sample configuration in Figure 4 shows how to convert data from the original PAULA XML format via POWLA into segmented CoNLL-RDF for further processing within Fintan. As a first step, we use the Pepper API service to transform PAULA to POWLA, then we use the RDF Splitter to induce the CoNLL-RDF data structure and split it by sentences into segmented graphs, which can be simultaneously processed with the RDF Updater for further customization. The splitting process uses Fintan's `ITERATE_CONSTRUCT` method by first selecting all sentence nodes in order with an iterator query:

```
SELECT ?s
WHERE {
  ?s a nif:Sentence
  BIND(xsd:integer(
    REPLACE(STR(?s),'[^0-9]','')
  ) AS ?snr)
} order by asc(?snr)
```

Subsequently, for each sentence `?s`, we execute a construct statement in which the wildcard `<?s>` is replaced by the specific sentence identifier:

```
CONSTRUCT {
  <?s> ?sp ?so .
  ?w ?wp ?wo .
} WHERE {
  <?s> ?sp ?so .
  ?w conll:HEAD+ <?s> .
  ?w ?wp ?wo .
}
```

The RDF Writer and CoNLL-RDF Formatter can then

output structured CoNLL or the CoNLL-RDF canonical format.

### 5.2. Adjusting the Recipe

The workflow depicted in Figure 4 is prepared to introduce additional processing steps. Since the resulting CoNLL-RDF data is already split into sentences, it is possible to directly execute updates on the segments transforming the existing annotations to commonly used schemes such as Universal Dependencies e.g. by using OLiA. OLiA at this point supports over 50 annotation schemes in its stable branch and features partial support for various additional models or reference catalogues including ISOcat and GOLD. In a similar manner, dictionaries could be side-loaded to infer foreign language lemmatization. Depending on the input data, even a complete recombination and restructuring of corpora is possible as we demonstrated by engineering a gold corpus for Role and Reference Grammar (Chiarcos and Fäth, 2019).

For native CoNLL output the CoNLL-RDF Formatter also alleviates structural customization, such as column reordering to directly feed data into subsequent NLP tools. With the CoNLL-RDF Ontology and CoNLL Transform (Chiarcos et al., 2021) we even introduced a means to automatically derive transformation pipelines from one CoNLL dialect to another which we aim to use as a blueprint for other formats as well.

In addition to producing CoNLL-RDF and CoNLL, also, other conventional corpus formats can be produced from POWLA. This includes bracketing formats as commonly used in treebanks such as the Penn Treebank (Marcus et al., 1993). XML-augmented TSV formats are SketchEngine (Kilgarriff et al., 2014) and the Corpus Workbench (Evert and Hardie, 2011) as also supported by the Fintan/CoNLL-RDF tool chain, but at the moment primarily as input formats.

## 6. One for the road

We have not just been cooking this up. Taking the combined capabilities of Pepper, Fintan and their configuration options into consideration, we are capable to cross-transform and recombine over 50 formats with a multitude of annotation schemes also taking advantage of parallelized stream processing. At the moment, this includes any CSV format (via Fintan's Tarql wrapper), 24 TSV formats (different CoNLL formats, Universal Morphology format, Sketch Engine/Corpus Workbench formats, OMW TSV format via CoNLL-RDF), 28 common corpus formats (via Pepper), all XML formats (with format-specific XSLT scripts for individual formats, e.g. for the Apertium dictionaries), the TBX format and numerous serializations of RDF data (RDF/XML, Turtle, JSON-LD, etc.). In addition to supporting different types of input data, Fintan also supports side-loading SKOS taxonomies, OWL ontologies, RDF and RDFS knowledge graphs as well as any custom XML or C/TSV resource when preprocessed into RDF graphs. As output formats we primarily support RDF serializations and customizable TSV formats. Since Fintan is an open platform, the number of supported formats can always be extended by building custom transformer components.

This combined support for taxonomies and both dictionary and corpus data inside a complex workflow manager which not only allows recombining existing converter components but alleviates full customization of transformation scripts is a somewhat unique approach to the data transformation challenge. Instead of dishing a buffet we provide ingredients and recipes in a prepare-your-own-pipeline package. Surely this is not a one-click solution and workflow accuracy is tied to the transformation components used. Specifically the preprocessing of data into POWLA using the Pepper framework heavily depends on individual Pepper modules and how lossless they render data in the internal SALT model. However, the resulting data can always be optimized within Fintan by additional resources or transformation steps in order to meet the requirements for specific use cases. Such preconfigured workflows can then be exported as stand-alone dockerized web services and made available in a sustainable way on Docker Hub[7] or as part of the European Language Grid[8] (ELG), which could also establish an interface to create additional processing nodes in large-scale infrastructure efforts such as the Switchboard or WebLicht platforms from CLARIN (de Jong et al., 2020).

We would like to emphasize that the integration of Pepper into Fintan not only increases the number of input and output formats supported by the platform. More importantly, it means that existing Fintan and CoNLL-RDF workflows can now be complemented with support for doing manual annotation (from HexAtomic, via Pepper), linguist-friendly means of querying and visualization (from ANNIS, via Pepper). And from the perspective of the Pepper/ANNIS universe, the addition of Fintan means that more advanced means of automated annotation and annotation engineering now become available that the native Java implementation of Salt'n'Pepper did not provide. Finally, via Fintan, Salt'n'Pepper can be connected with general-purpose NLP workflow management systems such as Teanga[9], linking it to the more algorithmic side of corpus linguistics with a possibility to integrate external webservices and Docker containers in annotation pipelines.

## 7. Acknowledgements

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

[7] https://hub.docker.com/

[8] https://www.european-language-grid.eu

[9] https://github.com/Pret-a-LLOD/teanga

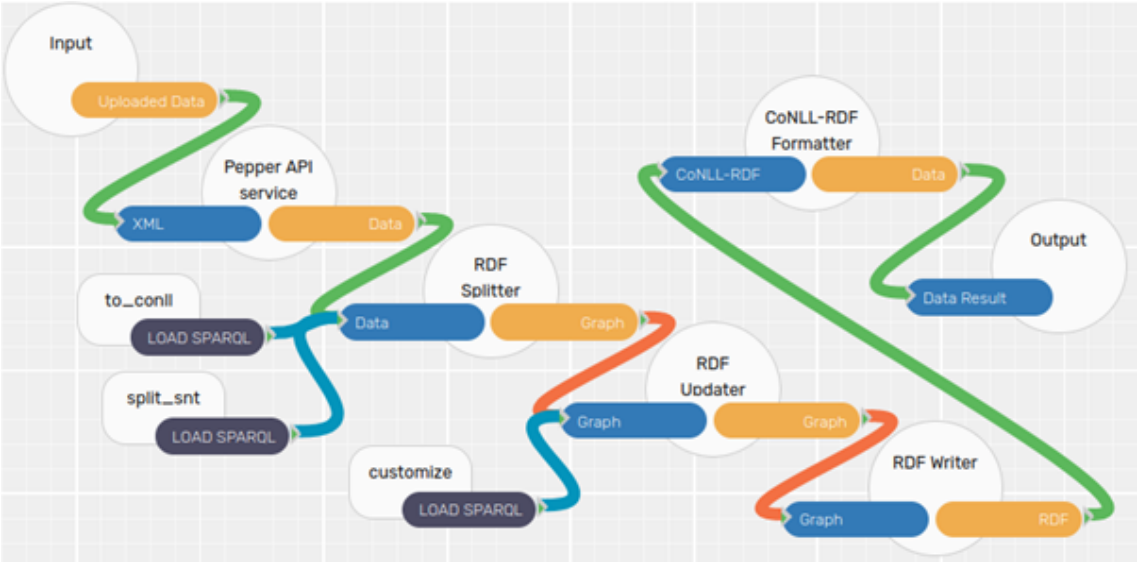


Figure 4: Fintan workflow converting PAULA to CoNLL-RDF

## 8. Bibliographical References


Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech communication*, 33(1-2):23–60.

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 Fäth, C. (2019). Graph-based annotation engineering: towards a gold corpus for role and reference grammar. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chiarcos, C. and Glaser, L. (2020). A tree extension for CoNLL-RDF. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7161–7169.

Chiarcos, C. and Sukhreva, M. (2015). OLiA – ontologies of linguistic annotation. *Semantic Web Journal*, 518:379–386.

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., Ionov, M., Glaser, L., and Fäth, C. (2021). An ontology for CoNLL-RDF: Formal data structures for TSV formats in language technology. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Chiarcos, C. (2012). POWLA: Modeling linguistic corpora in OWL/DL. In *9th Extended Semantic Web Conference (ESWC-2012)*, pages 225–239, Heraklion, Crete, May.

Cimiano, P., McCrae, J., and Buitelaar, P. (2016). Lexicon Model for Ontologies. Technical report, W3C Community Report, 10 May 2016.

de Jong, F., Maegaard, B., Fišer, D., van Uytvanck, D., and Witt, A. (2020). Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3406–3413, Marseille, France, May. European Language Resources Association.

Dipper, S., otze, M. G., Stede, M., and Wegst, T. (2004). ANNIS: A linguistic database for exploring information structure. In *Interdisciplinary Studies on Information Structure*, ISIS Working papers of the SFB 632 (1), pages 245–279. Universitätsverlag Potsdam.

Druskat, S., Gast, V., Krause, T., and Zipser, F. (2016). corpus-tools. org: An interoperable generic software tool set for multi-layer linguistic corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4492–4499.

Evert, S. and Hardie, A. (2011). Twenty-1st century Corpus Workbench: Updating a query architecture for the new millennium. In *Proc. of the Corpus Linguistics 2011 conference*, Birmingham. University of Birmingham.

Fäth, C., Chiarcos, C., Ebbrecht, B., and Ionov, M. (2020). Fintan - Flexible, Integrated Transformation and Annotation eNgineering. In *Seventh conference*

*on International Language Resources and Evaluation, LREC 2020*.

Gracia, J., Fäth, C., Hartung, M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., and Orlikowski, M. (2020). Leveraging linguistic linked data for cross-lingual model transfer in the pharmaceutical domain. In *International Semantic Web Conference*, pages 499–514. Springer.

Ide, N. and Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *Proc. the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 135–139, Lisbon, Portugal.

Ide, N. and Suderman, K. (2014). The linguistic annotation framework: a standard for annotation interchange and merging. *Language Resources and Evaluation*, 48(3):395–418.

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

Krause, T., Leser, U., and Lüdeling, A. (2016). graphannis: A fast query engine for deeply annotated linguistic corpora. *J. Lang. Technol. Comput. Linguistics*, 31(1):1–25.

R. A. S. Macalister, editor. (1941). *Lebor Gabála Érenn: Book of the Taking of Ireland*, volume 2 and 3. Irish Texts Society, Dublin, Ireland.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330.

Mendes, P., Jakob, M., and Bizer, C. (2012). DBpedia for NLP: A multilingual cross-domain knowledge base. In *8th international Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, May.

Suderman, K. and Ide, N. (2006). Layering and merging linguistic annotations. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*.

Verhagen, M., Suderman, K., Wang, D., Ide, N., Shi, C., Wright, J., and Pustejovsky, J. (2015). The lapps interchange format. In *Proc. of the Int. Workshop on Worldwide Language Service Infrastructure*, pages 33–47. Springer.

Zipser, F. and Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, Malta.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 69–77
Marseille, 24 June 2022


# A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data

**Anas Fahad Khan[1], Christian Chiarcos[2], Thierry Declerck[3], Maria Pia di Buono[4], Milan Dojchinovski[5,6], Jorge Gracia[7], Giedre Valunaite Oleskeviciene[8], Daniela Gifu[9,10]**

[1] Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Italy, fahad.khan@ilc.cnr.it
[2]Applied Computational Linguistics, Goethe University, Frankfurt, Germany, chiarcos@cs.uni-frankfurt.de
[3]DFKI GmbH, Saarland Informatics Campus, Saarbrücken, Germany, declerck@dfki.de
[4]UNIOR NLP Research Group, University of Naples "L'Orientale", Italy, mpdibuono@unior.it
[5]Faculty of Information Technology, Czech Technical University in Prague, milan.dojcinovski@fit.cvut.cz
[6]DBpedia Association/InfAI, Leipzig University, Germany, dojchinovski@informatik.uni-leipzig.de
[7]Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain, jgracia@unizar.es
[8]Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania, gvalunaite@mruni.eu
[9]Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania, daniela.gifu@info.uaic.ro
[10]Institute of Computer Science, Romanian Academy - Iasi Branch, Romania, daniela.gifu@iit.academiaromana-is.ro

**Abstract**

This article discusses a survey carried out within the NexusLinguarum COST Action which aimed to give an overview of existing guidelines (GLs) and best practices (BPs) in linguistic linked data. In particular it focused on four core tasks in the production/publication of linked data: generation, interlinking, publication, and validation. We discuss the importance of GLs and BPs for LLD before describing the survey and its results in full. Finally we offer a number of directions for future work in order to address the findings of the survey.

**Keywords:** Linguistic Linked Data, Guidelines, Best Practices

## 1. Introduction

This article has its origin in a survey on the use of specific Linked Data (LD) vocabularies for different categories of language resources. The survey was carried out by the COST Action "CA18209 - European network for Web-centred linguistic data science"[1]. At the time of writing this article, the Linguistic Linked Open Data (LLOD) Cloud,[2] consisting of datasets belonging to the linguistics domain, makes up one of the largest subsets of the Linked Open Data (LOD) cloud, with 227 datasets out of a total 1301 in the whole LOD cloud[3]. However, after over a decade of LLD, and despite the advantages and opportunities of LLD, there is still room for improvement, not least in terms of languages covered[4] and types of linguistic dataset represented in the LLOD cloud. The provision of clearly formulated guidelines and best practices written in different languages (and featuring use cases dealing with a range of languages) and types of resources could help to close these gaps and make Linguistic Linked Data (LLD) more accessible. In addition, such a documentation can also help in the exploitation of the datasets in the LLOD cloud, i.e., to help it realise its full potential. We therefore reflect in this paper on the current state of such guidelines and best practises, the main topics they cover, their targeted audience, etc. as well as their limitations and the aspects that future materials of this type should fulfil. Indeed, there exist caesurae and weaknesses in the available documentation which prevent the full exploitation of LD principles for linguistic data. The remainder of the paper is organised as follows. In Section 2 we identify some desirable aspects that guidelines and best practises on LLD should fulfil. Then, Section 3 describes our survey of currently available materials and, finally, Section 4 features a discus-

[1]The short name of the COST Action being "NexusLinguarum", https://nexuslinguarum.eu/

[2]The LLOD cloud is accessible at http://linguistic-lod.org/llod-cloud

[3]https://lod-cloud.net

[4]According to a recent policy brief (Bosque-Gil et al., 2021) on under-resourced languages the availability of language resources demonstrates tremendous differences across languages. Some languages like English have an abundance of the resources available for LLOD technologies, while some other languages show scarcity of resources. This lack of language resources is damaging for at least two reasons: first, the application of advanced data processing technologies is limited as they require extensive data and next, the automated development and enrichment of language resources becomes really scarce. In addition, linguistic resources vary in their depth of the information available on numerous linguistic features, thus the use and re-use of the data poses a twofold challenge of processing in width and in depth of the available linguistic resources. The resources are unevenly developed in different languages and some languages may not have the material developed available for LLOD technologies.

sion of what is missing, with a number of suggestions addressed to the LLD community to produce useful guidelines and best practices.

## 2. The Role of Guidelines and Best Practices in LLD

### 2.1. Some Definitions

One aim of the survey was to give an overview of existing guidelines (GLs) and best practices (BPs) with respect to four core tasks in the production/publication of linked data. These are: **generation**, **interlinking**, **publication**, and **validation**. The survey also helped in determining what is missing or needs to be updated in those areas, leading to the intention to work on these gaps, also in collaboration with other initiatives. Before we continue, however, we should clarify what we mean here by 'guidelines" and "best practices" In the first instance, we can adopt the definition given by the Cambridge English Dictionary[5], stating that a *guideline* is:

> *information intended to advise people on how something should be done or what something should be.*

For *best practice*, we can adopt the Merriam-Webster definition[6]:

> *a procedure that has been shown by research and experience to produce optimal results and that is established or proposed as a standard suitable for widespread adoption.*

Understood in this way, there are relatively few resources which can label themselves either as guidelines or best practices, or anything that could be construed as a synonym of these, in the context of Linguistic Linked Data (LLD). But there is a reasonably large number of other types of material and resources which fulfil, in part, the role of a set of guidelines and best practices as we have defined them above. These include, for instance, one or more sections in the technical report for a standard or individual chapters in an introductory textbook. Our survey therefore took into consideration all of these types of material and resources. We describe our methodology, data gathering process and results in Section 3.

### 2.2. Desiderata

Understanding the advantages of LLD and the many opportunities it offers as a means of publishing linguistic data as FAIR data[7] requires some level of technical appreciation of the Semantic Web, of RDF and other formalisms as well as a number of other technologies. Nonetheless, in order to increase the uptake of LLD amongst non-specialists, it is essential that materials are made available which are accessible to non-specialists and which give clear instructions and ways of doing common tasks (the role of GLs/BPs). A related issue here is the need for LLD specific technologies which target non-specialists (as opposed to more generic Semantic Web oriented applications and technologies such as protégé[8])[9]. The use of more accessible tools will in turn make the production of more accessible guidelines more viable; something we discuss in Section 4.

In other cases, the provision of clear and easy-to-understand guidelines have been essential in helping to introduce standards and technologies to target audiences. This, for instance, is the case with the Text Encoding Initiative Guidelines[10], which in addition to describing the Text Encoding Initiative approach to annotation themselves (and the elements of which it consists) also incorporates a valuable introduction to XML itself targeted towards humanists. In this context, therefore, there is no clear line between what counts as didactic materials and guidelines and best practises; this is why we have included two self-contained online courses in our list of miscellaneous materials in Section 3.3.

As in any other domain, the use of GLs/BPs in LLD helps to fill the gap between a technical description of a standard and its use in practice; and indeed both kinds of documentation help to ensure the interoperability, and therefore FAIRness of resources[11]. However, it takes on a special significance for LLD given that Linked Data is one of the core technologies which is helping to make FAIR a reality. We end this section with a list of desiderata for LLD GLs/BPs based on the experience of the authors as both consumers and compilers of such documents:

- Multilinguality: they should not just be in English, but should make LLD accessible to speakers of other languages;
- They should be easy to find and access, preferably with an open licence and not behind a paywall; this very fits in the spirit of LL**O**D;
- They should give clear instructions for how to carry out different tasks and be as self-contained as possible (and save users from having to wade through text that is not relevant for their information need). In particular, they should be organised according to the task they are developed for;
- They should be pitched at different levels of expertise but especially for beginners (given we need to

[5] https://dictionary.cambridge.org/dictionary/english/guideline
[6] https://www.merriam-webster.com/dictionary/best%20practice
[7] https://www.go-fair.org/
[8] https://protege.stanford.edu/
[9] There are few generally accessible tools that offer specific provision for LLD use cases, one of those that does exist is VocBench, see (Stellato et al., 2020)
[10] https://tei-c.org/Guidelines/
[11] https://www.go-fair.org/fair-principles/

increase uptake of the technology);

- They should cover (at least) the types of resources listed in the LLOD cloud, and the four tasks of generation, interlinking, publication, and validation;
- They should be aware of existing tools which can be integrated in the workflow
- Be regularly updated to ensure they keep up to date with the latest technology/models/tools.

This list of desiderata will help us evaluate the already existing materials which we have found in our survey and which we look at in the following section, as well as to suggest what to prioritise when it comes to producing new materials.

## 3. A Survey of Already Available Materials

In order to come up with a candidate list of resources for our survey, we solicited input from the members of the NexusLinguarum COST Action, a group that consists of researchers and linguistic linked data experts with extensive experience in numerous relevant projects and initiatives. In addition, this work also benefited from the extensive process of data collection which was carried out as part of the survey paper on LLD models (Khan et al., 2022 in press) produced as part of Task 1.1 of the NexusLinguarum Action; this included the compilation of a survey of LLD-relevant projects and other relevant initiatives (i.e. W3C community groups). Each of the resources contained in the survey have been described/categorised using a number of salient metadata fields. The fields were chosen with an eye to the potential (re-)usability of these resources. Accordingly, we have specified the level of expertise which is assumed by each resource according to the following categorisation. Note that the "Beginner level" of expertise assumes some basic knowledge of linked data and the Semantic Web, e.g., the concept of a triple, the fact that linked data is structured as a series of subject-object-triples what a SPARQL endpoint is. We do not deal with basic materials for learning about linked data and the Semantic Web here, since our focus is on linguistic linked data and not linked data in general. However, the beginner level of expertise should not assume any specialist knowledge of different areas of (Computational) Linguistics or NLP. For instance, materials which required an intermediate level of familiarity of corpus linguistics but only a basic level of familiarity with (linguistic) linked data would be classed as "Intermediate". An "Intermediate level" of expertise in this context assumes either an intermediate level of familiarity with LLD and/or with some area of Computational Linguistics. The "Advanced level" of expertise is defined similarly.

| Target Audience | Description |
|---|---|
| Beginner | Assumes little or no LLD or technical knowledge |
| Intermediate | Assumes some LLD or technical knowledge |
| Expert | Assumes advanced LLD or technical knowledge |

Table 1: Levels of Expertise.

Additionally, in our survey we have listed a number of keywords for each resource, including the tasks it is useful for and the kind of resource it covers. In the latter case, we have taken the classification used to categorise the resources in the LLOD cloud, namely (abbreviations in parentheses are used in the survey tables below): Corpora **(Corp)**; Lexicons and Dictionaries **(LD)**; Terminologies, Thesauri and Knowledge Bases **(TTKB)**; Linguistic Resource Metadata **(LRM)**; Linguistic Data Categories **(LDC)**; and Typological Databases **(TD)**. In addition, whenever a resource assists in carrying out one or more of the four tasks which we are focusing on in this deliverable, i.e., generation **(Gen)**, interlinking **(InL)**, publication **(Pub)** or validation **(Val)**, we also add it as a keyword. Note that **Gen** here also includes the sub-tasks of data modelling and conversion of datasets into LLD. In the following subsections, we look at the different kinds of materials described in the survey[12].

### 3.1. Guidelines and Best Practices

In this section, we consider GL/BP's that either advertise themselves as such or that very clearly have this purpose, that is, the provision of guidelines and best practices for LLD, as a primary aim (as distinct e.g., from technical reports for standards or textbooks which, while fulfilling the role played by GLs and BPs, also have other, distinct aims). It became clear during the information gathering phase of this survey that there was a dearth of materials or resources fitting this description. Here we can, however, mention two different sets of materials, the first of which was produced as a result of work carried out by the now dormant 'Best Practices for Multilingual Linked Open Data' (BPMLOD) W3C community group, and the second of which was an output of the LIDER project [13].

Table 2 describes the eight guidelines made available as part of the BPLMOD set of guidelines. These comprise guidelines for generating multilingual[14] and bilingual[15]

[12]The full survey will be made available as a NexusLinguarum deliverable in April 2023.

[13]https://lider-project.eu

[14]http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/

[15]https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/

dictionaries, wordnets[16], TBX terminologies[17], developing NIF services[18] and LLOD aware servicess[19] and creating corpora with NIF[20]. Finally, there are guidelines for LLD exploitation[21]. It is notable that all the BPMLOD guidelines are from 2015, seven years from the time of writing and prior to the new version of *lemon*, OntoLex-Lemon, which was published in 2016[23]. This is problematic because there are numerous classes and properties which exist in OntoLex-Lemon and not in *lemon* and vice versa. It is also prior to the publication of the OntoLex-Lemon lexicographic module in 2019[24] (something which clearly affects the first two dictionary related guidelines). As well as being out of date, they do not cover all tasks and all types of resources. This is problematic, given the lack of alternative and more recent materials.

Table 3 summarises the eight reference cards which were made available by the LIDER project. These include guides to publishing linked data[25], language resource licensing[26], inclusion in the LLOD cloud[27], data IDs[28], language resource discovery with Linghub[29], NIF corpora[30], the representation of crosslingual links[31] and language resource documentation in datahub[32].

Such cards are structured as 'how to' instructions to address different types of target audiences, e.g., data publisher, data creator, and different scopes, e.g., publishing LD on the Web. Furthermore, they clearly state the steps and the knowledge needed, e.g., RDF knowledge, together with the resources/tools useful for reaching the goal.

These reference cards were intended to offer sets of guidelines for carrying out a number of tasks, ranging from publication, adding metadata, and including resources on the LLOD cloud, which were accessible for beginners. Again all of these cards date from a specific year, and once again this year is 2015 (an exemplary year for LLD guidelines and best practices!). Unfortunately, we were unable to find any licensing information for these reference cards, so it is unclear how and when they can be re-used. Note also that neither the BPMLOD guidelines nor the reference cards deal directly with the validation of linked data, nor do they offer any special assistance in the case of working with typological databases. Additionally, the reference cards run to two pages each and are limited in the amount of information they offer with respect to the task of enriching a linguistic dataset with metadata or dataset crosslinking.

Finally, the *lemon* cookbook[33], which was an output of the Monnet project[34] which provided an introduction to the *lemon* model, describing each of its submodules and generally fulfilling the role of a set of guidelines. For Ontolex-Lemon, the official W3C community report of the final specifications of the model fulfils the role played by the *lemon* cookbook for Ontolex-Lemon as we discuss in Section 3.2.

## 3.2. Standards

Another group of documents relevant to this discussion are technical reports and specifications for LLD-related standards. These include 'official' formal standards: those that are issued and maintained by designated institutions[35] and subject to a formal, institution-specific process of proposal, review, revision, confirmation and withdrawal. These can be subject to a number of constraints on formats and means of presentation that usually make them less accessible than some other kinds of materials we've looked at above and which take a more didactic stance. In addition to formal standards, a number of specifications exist, which are treated as *de facto* standards specifications by the community without being published as official standards by some standardisation body. In what follows we largely focus

[16] http://bpmlod.github.io/report/WordNets/index.html (Unofficial Draft)

[17] https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/

[18] https://www.w3.org/2015/09/bpmlod-reports/nif-based-nlp-webservices/

[19] http://bpmlod.github.io/report/LLOD-aware-services/index.html

[20] http://bpmlod.github.io/report/nif-corpus/index.html (Unofficial Draft)

[21] https://www.w3.org/2015/09/bpmlod-reports/lld-exploitation/

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

[24] https://www.w3.org/2019/09/lexicog/

[25] http://bpmlod.github.io/report/LLOD-aware-services/index.html

[26] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-publish-linguistic-linked-data-Reference-Card.pdf

[27] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Inclusion-in-the-LLOD-Cloud-Reference-Card.pdf

[28] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/DataID-Reference-Card.pdf

[29] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Discovering-Language-Resources-with-Linghub.pdf

[30] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/NIF-Corpus-reference-card.pdf

[31] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-represent-crosslingual-links-Reference-Card.pdf

[32] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Documenting-a-language-resource-in-Datahub.pdf

[33] https://lemon-model.net/lemon-cookbook/index.html

[34] https://cordis.europa.eu/project/id/248458

[35] These include standardisation bodies such as, for example, W3C, OASIS and ISO.

| Title | License | Target | Keywords | Last Updated |
|---|---|---|---|---|
| Guidelines for Linguistic Linked Data Generation: Multilingual Dictionaries (BabelNet) | W3C Community FSA[22] | Expert | babelnet, lemon, wordnet, generation, **LD**, **Gen** | 2015 |
| Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries | W3C Community FSA | Expert | bilingual dictionary, lemon, translation, multilingual lexical resources, **LD**, **Gen** | 2015 |
| Guidelines for Linguistic Linked Data Generation: Multilingual Terminologies (TBX) | W3C Community FSA | Expert | Multilingual terminologies, TBX, resource conversion, **TTK**, **Gen** | 2015 |
| Guidelines for Developing NIF-based NLP Services | W3C Community FSA | Expert | NIF, NLP services | 2015 |
| Guidelines for LLD exploitation | W3C Community FSA | Intermediate | LLD services, use cases | 2015 |
| Guidelines for Linguistic Linked Data Generation: WordNets | W3C Community FSA | Expert | wordnet, lemon, **LD**, **TTK** | 2015 |
| Guidelines for Linked Data corpus creation using NIF | W3C Community FSA | Expert | NIF, **Corp** | 2015 |
| Guidelines for LLOD aware services | W3C Community FSA | Expert | LLD services, use cases | 2015 |

Table 2: The BPMLOD Guidelines

on such community standards since there do not exist many LLD-specific (as opposed to linked data specific) formal standards. In fact, we look at individual technical reports and specifications for such *de facto* standards to see the extent to which such documentation can fulfil the role of GLs and BPs[36]. The primary purpose of such documentation is undoubtedly to give an exhaustive and unambiguous description of a standard. In many cases, however, they are also intended to assist users in applying the standard, often by providing examples of its use in typical use cases – and in this they play the same role as GLs/BPs.

**Standards for lexical-semantic resources** We will start by looking at the specifications for OntoLex-Lemon, a W3C community standard for lexical resources which we mentioned above and which was originally inspired by the UML-based proprietary ISO standard Lexical Markup Framework (LMF) on its first iteration. The OntoLex-Lemon specifications were published in 2016 in the W3C namespace as a community report by the W3C Ontology-Lexicon group[37]; note however that as reported in document itself OntoLex-Lemon is not a W3C recommendation (and neither is it on the W3C recommendations track). Besides detailed descriptions of the single classes and properties in the model, the specifications also give (simple and fairly accessible) examples of the use of the latter, both in the form of diagrams and snippets of code. In general the text of the specifications is fairly expansive and goes beyond the more technical presentation of, e.g., ISO standards making the document accessible to the Beginner level of user. These specifications can therefore be said to fulfil the role of a set of beginner's guidelines or a primer to OntoLex-Lemon. On the other hand, there are many use cases (especially for generation but also other tasks) which they don't (and given their status as guidelines shouldn't) capture. Moreover, the guidelines are so far only available in English with the examples mostly in English, in addition to a handful of others in Latin, French, Spanish and German. As well as the necessity of translations of the OntoLex-Lemon specifications in other languages

[36] We leave out documentation for vocabularies like SKOS, SKOS-XL, DCAT, DCMI which aren't LLD specific even though they are often used in LLD datasets, neither do we deal with LLD (de facto) standards which do not currently have accessible reports/specifications.

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

| Title | License | Target | Keywords | Last Updated |
|---|---|---|---|---|
| How to publish Linguistic Linked Data | N/A | Beginner | Linking Data, Resolvable URIs, **Gen**, **Pub**, **InL** | 2015 |
| Language Resource Licensing - ODRL Reference Card | N/A | Beginner | RDF Conversion, Data Modeling, Linking Data, Resolvable URIs, **LRM**, **Gen**, **Pub** | 2015 |
| Inclusion in the LLOD Cloud | N/A | Beginner | LLOD Cloud, Datahub, Linked Dataset, **Pub** | 2015 |
| Data ID | N/A | Beginner | Dataset description, DataID, Resource Metadata, Pub | 2015 |
| Discovering Language Resources with Linghub | N/A | Beginner | LingHub, Resource Discovery, Language Resources | 2015 |
| NIF corpus | N/A | Beginner | NIF, RDF, Corpus Conversion, **Corp** | 2015 |
| How to represent crosslingual links | N/A | Beginner | Cross-lingual Linked Data links, Cross-lingual mapping, **Pub**, **InL** | 2015 |
| Documenting a language resource in Datahub | N/A | Beginner | Metadata documentation, DataHub, DCAT, data description, **LRM**, **Gen**, **Pub** | 2015 |

Table 3: Lider Project Reference Cards

(with examples in other languages too), it is also clear that we need more Intermediate and Expert level materials dealing with more advanced modelling topics for OntoLex-Lemon. However, it is not unreasonable to assume that the popularity of OntoLex-Lemon is due in no small part to the accessibility of the specifications, both in terms of the fact that they are openly available and unlike ISO standards like LMF aren't closed or behind a paywall, and their readability.

Two years after the publication of these specifications, the W3C Ontology Lexicon group published the specifications for an extension to the OntoLex-Lemon model, dealing this time with lexicographic resources, namely, the OntoLex-Lemon Lexicography Module (lexicog) [38]. In line with the specifications of the original model, these specifications were furnished with illustrative examples for individual classes and properties. The limitations of these guidelines are the same as those of the original model; as will likely be the case for another two follow-up OntoLex-Lemon modules in an advanced phase of preparation (the first dealing with the representation of morphology, the second with frequency, attestation and corpus data), with others also being planned, including an extension for terminologies (this would make a good start in developing guidelines for the TTK category).

**Standards for linguistic annotation** There is currently no settled consensus as to which is the most suitable linguistic annotation mechanism for LLD. This is important since linguistically annotated data plays a vital role in current NLP/AI technologies.[39] NLP Interchange Format NIF and the Web Annotation standard, a W3C recommendation that developed out of the Open Annotation community. NIF is a community standard developed in a series of research projects at the AKSW Leipzig, Germany, and still maintained by that group. In addition to that, it enjoys a semi-official status as a component of the Internationalization TagSet (ITS 2.0) which is a formal W3C standard that describes the application of NIF. Web Annotation is a W3C recommendation that evolved out of the Open Annotation vocabulary, a community standard originally published

[38] https://www.w3.org/2019/09/lexicog/

[39] Here the importance of collaborating with small and medium enterprises (SMEs) in the development of new standards should be emphasised. This would have the effect of helping them to establish new business relationships and enter new markets early. Vice versa, the experiences of SMEs in working with Semantic Web technologies would likely prove crucial to strategic discussions about the Web's future.

as a community report of the W3C Community Group Open Annotation.

Both Web annotation and NIF build on the use of URIs (resp., IRIs) for addressing corpora, and this coincides with the use of URIs (IRIs) in TEI and XML standoff formats. A typical UR/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For numerous media types and different file formats, different fragment identifiers have been defined, often as best practices (BPs; also referred to as Requests for Comments, RFCs) of the Internet Engineering Task Force (IETF). Other, format-specific standards include the W3C standards SVG (Scalable Vector Graphics),[40] XPointer (for addressing XML documents)[41], or Media Fragments[42]. None of these are specific to linguistic annotation, but they can be used in conjunction with Web Annotation or NIF. The level of presentation in these standards and community standards is relatively technical, its content is normative and oriented towards engineers that are responsible for implementing the corresponding reference functions. None of these standards is particularly user-friendly. In addition to standards and community standards, a URI schema for Web Annotation selectors is provided as a working note that accompanies the W3C recommendation. Again, this document has the same level of technicality. It is therefore clear that this is one area where there is a real necessity for documentation that provides clear GL's and BP's.

### 3.3. Miscellaneous

Finally, we round off this current section by looking at other types of materials or resources which have served, or which might serve, to play the role of GLs and BPs for LLD, alongside a range of other didactic or expository tasks. One category of materials which can often play this role is textbooks and monographs and here in particular we can cite the introductory text *Linguistic Linked Data: Representation, Generation and Applications*, (Cimiano et al., 2020). This book is intended to be primarily introductory, but also contains intermediate and advanced materials. Although designed to be self-contained, it recommends, in each chapter, a number of additional readings to complete the given overview and to get deeper into some details. The book is structured in four main blocks: preliminaries (a basic introduction to linked data and linguistic linked data), modelling (lexical data, annotated texts, linguistic annotations, metadata), generation and exploitation (generation of LLD resources, linking, workflows), and use cases (multilingual wordnets, digital humanities, discovery of language resources). Although not conceived as a set of guidelines in itself, it shares many commonalities with our previous definition of guidelines, and is a valuable source of reference for those interested in LLD in general or in any of its particular aspects. Overall, there are at least a couple of major drawbacks to using such books as sources for GLs/BPs. For a start, and given current publishing practices (and notwithstanding a growing movement towards publication of open edition) their digital editions tend to be paywalled, with the kind of copyright licenses that mean that the information in them can't be shared – at least not legally. More generally, information contained in them and which pertains to GLs/BPs tends not to be in a self-contained format. There are many similar issues with articles (paywalls, copyright, less focus on providing self-contained sets of GLs/BPs). Another category of material or resource that is salient to the current discussion are didactic or course materials. In order to respond to the information needs of users looking for GLs/BPs these should be self-contained (and not depend on other materials) as well as, preferably, made freely available. Although one can often find slides (both from courses and from conference/workshop presentations) which will in many cases answer specific questions, it's difficult to find materials which can more generally take on the function of GLs/BPs. Here, however, we can mention two courses published on the DARIAH-CAMPUS platform (the latter being as the name suggests an initiative of the DARIAH infrastructure) and which were produced as an output of the ELEXIS European Project and which fulfil in large part the role of GLs/BPs. The first is the course *Modeling Dictionaries in OntoLex-Lemon*[43]; the second is the *Lexicography in the Age of Open Data* [44]. These are much closer to the materials we looked at in Section 3.1, especially the BP-like content of the Lider reference cards.

### 3.4. Observations

Returning to the list of desiderata listed in Section 2 and in light of the last few sections, what observations can we make with respect to what exists? The most obvious one is simply that there aren't enough materials available fulfilling the role of GLs/BPs for linguistic linked data, and moreover a lot of what exists hasn't been updated for years and doesn't reflect the latest developments in the field. And this is true of all levels of expertise. In the case of OntoLex-Lemon and its extension(s), these are well served by their specifications; moreover, OntoLex-Lemon is regarded as *the* de facto standard for lexicons and dictionaries. This makes it much easier to produce further materials, at least in contrast to cases where there is no such settled standard (or when there are too many incompatible standards). This would argue in favour of initiatives for consolidating competing standards or rendering them interopera-

[40] https://www.w3.org/TR/SVG11/linking.html

[41] https://www.w3.org/TR/xptr-framework/

[42] https://www.w3.org/TR/media-frags/

[43] https://elexis.humanistika.org/resource/posts/modeling-dictionaries-in-ontolex-lemon

[44] https://elexis.humanistika.org/resource/posts/lexicography-in-the-age-of-open-data

ble might[45]. In the case of books and articles, these can be helpful in providing sets of GLs and BPs, but such materials are usually not published as open source publications, or digital editions are behind paywalls, and might not be the organised in a way that's convenient for those searching for specific GLs/BPs. All of which suggests a real need for new GLs and BPs.

Finally, the question of the languages in which GLs/BPs are written in (as well as the kind of examples which they feature) is a crucial one, especially for the uptake of LLD standards and technologies. The lack of information available in languages other than English reflects a similar disparity in language resources. As suggested in the introduction, the provision of GLs/BPs in other languages and/or with the inclusion of a wider range of linguistic examples from typologically diverse languages could help to improve this situation. Overall, the need to provide easy-to-read guidelines and goal-oriented instructions, addressing different levels of expertise and use cases, calls for a re-organisation and integration of existing documentation.

## 4. Conclusion: What is to be done?

After laying out the current situation with respect to GLs and BPs for LLD, we suggest a number of future work directions. We propose to promote and/or (wherever possible) implement these work directions within the framework of the Nexus Linguarum COST action in collaboration with other initiatives and projects as discussed below.

**Update existing GLs and BPs; Solicit feedback for new GLs/BPs** Perhaps the lowest hanging fruit here: Given the continuing existence of the W3C BPMLOD group (even if currently inactive), one obvious proposal would be for Nexus Linguarum participants to work with that group on updating already existing GLs. In addition, suggestions for new GLs and BPs could be solicited both from that group and other relevant W3C groups such as the W3C Ontology Lexicon group and Nexus Linguarum mailing lists, and indeed any other relevant community list. This brings us onto our next proposal.

**Use case/example driven GLs and BPs; Bridging GLs and BPs and tools** As we have seen, there is a real need to adapt and extend GLs and especially BPs with more use case driven examples. One idea would be to reinstate something like lemon patterns, or to make use of a repository of ontology design patterns (this idea is further discussed in (Khan et al., 2022 in press)). In addition, where possible, GLs and BPs should focus on actual implementation of the particular task using a concrete tool or software.

**A Central Hub for GLs and BPs.** Another proposal would be to establish a central hub for LLD. This would significantly help with the discovery of relevant materials. Currently, there is a lack of a reference point for search and discovery of BPs and GLs.

**Open, editable and collaborative GLs and BPs.** In order to keep materials up-to-date, it is necessary to enable users to directly contribute to the materials and provide updates when necessary. This can be achieved by providing the materials through a wiki system or using markdown documents. Both this and the previous proposal could be undertaken in collaboration with infrastructures like CLARIN or DARIAH (as part of the Social Sciences and Humanities Open Cloud(SSHOC) cluster[46]).[47]). It should not be neglected that some replication even with stable and well maintained infrastructures might be considered. In fact, one of the past initiatives of DARIAH was to enhance communications between five European Research Infrastructures (ERICs) in the Social Sciences & Humanities (SSH): CLARIN, DARIAH, European Social Survey (ESS), Survey of Health, Ageing and Retirement in Europe (SHARE), Consortium of European Social Science Data Archives (CESSDA). In addition, Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS) supports the work of CLARIN and DARIAH. **Interactive BPs and GLs.** While most available materials are static (e.g. PDF documents or static HTML pages) making use of video clips and quizzes would significantly help with knowledge transfer and increase user engagement. In particular, organising a massive open online course (MOOC) could help to deliver learning content online in an interactive way. Different levels could be offered catering to users with different levels of expertise and/or different backgrounds In fact, OER (Open Education Resources) would be more appropriate to cover a wide range of online learning formats, like the ones already mentioned and many more.

## 5. Acknowledgements

This article is based upon work from COST Action NexusLinguarum – "European network for Web-centered linguistic data science" (CA18209), supported by COST (European Cooperation in Science and Technology) `www.cost.eu`. The article is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015), by the I+D+i project PID2020-113903RB-I00, funded by MCIN/AEI/10.13039/501100011033, by DGA/FEDER, and by the *Agencia Estatal de Inves-*

[45] Indeed, an initiative is underway for such a consolidation for RDF vocabularies for linguistic annotation within Nexus Linguarum

[46] https://sshopencloud.eu/

[47] Indeed in addition to the infrastructures mentioned above there are several other European initiatives supporting the development of a unique platform to access language technologies and tools for all European languages, e.g., European Language Grid https://www.european-language-grid.eu/ (ELG) which could also collaborate in the development of shared and user-oriented documentation.

*tigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

## 6. Bibliographical References

Bosque-Gil, J., Mititelu, V. B., Oliveira, H. G., Ionov, M., Gracia, J., Rychkova, L., Oleskeviciene, G. V., Chiarcos, C., Declerck, T., and Dojchinovsk, M. (2021). Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud. https://nexuslinguarum.eu/results/policy-briefs.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., et al. (2022 (in press)). When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web Journal*.

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., et al. (2020). Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5):855–881.

*Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 78–86
Marseille, 24 June 2022


# Computational Morphology with OntoLex-Morph

**Christian Chiarcos[1,2], Katerina Gkirtzou[3], Anas Fahad Khan[4], Penny Labropoulou[3], Marco Passarotti[5], Matteo Pellegrini[5]**
[1]Applied Computational Linguistics, Goethe University Frankfurt, Frankfurt am Main, Germany
[2]Institute for Digital Humanities, University of Cologne, Cologne, Germany
[3] Institute of Language and Speech Processing, Athena Research Center, Athens, Greece
[4]Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Italy
[5]Università Cattolica del Sacro Cuore, Milan, Italy
[1]chiarcos@cs.uni-frankfurt.de, [3]{katerina.gkirtzou,penny}@athenarc.gr,
[4]fahad.khan@ilc.cnr.it, [5]{marco.passarotti,matteo.pellegrini}@unicatt.it

**Abstract**

This paper describes the current status of the emerging OntoLex module for linguistic morphology. It serves as an update to the previous version of the vocabulary (Klimek et al. 2019). Whereas this earlier model was exclusively focusing on descriptive morphology and focused on applications in lexicography, we now present a novel part and a novel application of the vocabulary to applications in language technology, i.e., the rule-based generation of lexicons, introducing a dynamic component into OntoLex.

**Keywords:** OntoLex, computational morphology, inflection, derivation, compounding, finite state transducers

## 1. Background and Introduction

This paper describes the current status of the emerging module for linguistic morphology of the OntoLex vocabulary (Cimiano et al., 2016). It serves as an update to Klimek et al. (2019) and introduces a novel part of the vocabulary designed for rule-based generation of lexicons, introducing a dynamic component into OntoLex. The generation component is intended to allow for the dynamic generation of morphological variants of a single lexical entry; that is, it is intended to permit *intensional* as well as *extensional* morphological descriptions. In the latter kind of description all inflected forms of an entry (in the case of an inflected language) are explicitly listed; in the former, morphological information is given in a manner that allows individual forms to be generated dynamically.

Preliminary results in the development of this module have been published by Klimek et al. (2019), but, at the time, with a strict focus on extensional (descriptive) morphology and use cases from lexicography. Since then, we intensified research on intensional morphology and morphological generation and we now present the revised, consolidated model that has emerged. We consider the current draft to be near-final and would like to use this publication to elicit feedback from a broader audience before finalising it and publishing as a W3C Community Report akin to OntoLex-Lemon (Cimiano et al., 2016) and *lexicog*, the OntoLex module for lexicography (Bosque-Gil and Gracia, 2019).

The OntoLex-Lemon (core) model is illustrated in Fig. 1. It was foreseen in OntoLex that more detailed morphological information would be provided at a later point in time. In particular, the OntoLex core model includes the object property ontolex:morphologicalPattern, which, however, remained

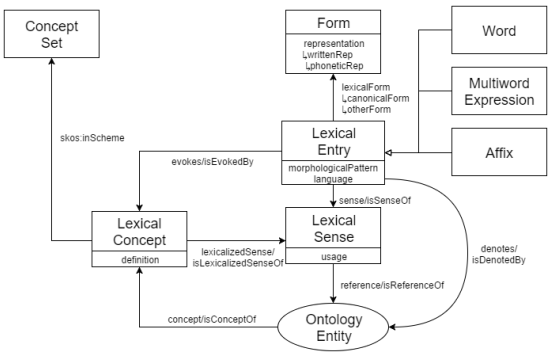


Figure 1: OntoLex-Lemon core model

underspecified until a future module for morphology would have been created. OntoLex-Morph is the current prototype for this module.

In this paper, we first describe the OntoLex-Morph vocabulary (Sect. 2) and then elaborate on three use cases (Sect. 3) for inflection, word formation and compounding in three inflecting languages. When developing approaches for the computational application of OntoLex-Morph, we initially focused on inflecting languages with relatively rich morphology (in comparison to English). Section 4 summarises the main achievements, and presents the open issues currently under investigation. Finally, Sect. 5 gives an outlook towards publishing OntoLex-Morph as a W3C vocabulary, i.e., as Community Report of the W3C Community Group Ontology-Lexica, and thus, as a formal addendum to the OntoLex vocabulary.

## 2. OntoLex-Morph

The current version of OntoLex-Morph module is shown in Fig. 2.

Class **morph:Morph** is a subclass of `ontolex:LexicalEntry` that represents

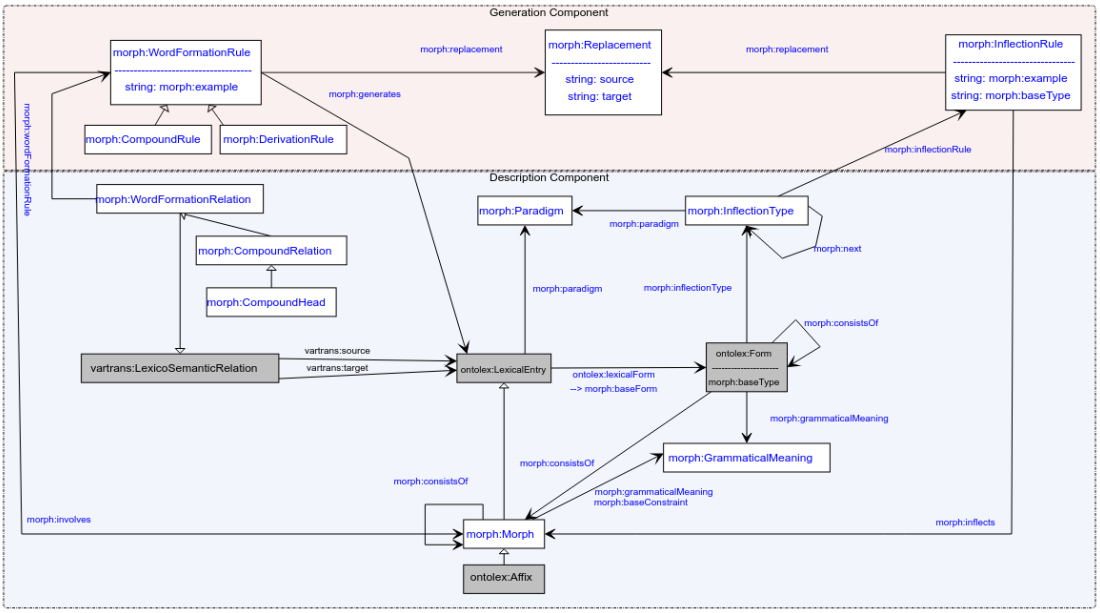


Figure 2: OntoLex-Morph vocabulary, draft version 4.13, April 2022

a concrete primitive element of morphological analysis. Morph is both a subclass of `ontolex:LexicalEntry` and a superclass of `ontolex:Affix`. Certain morphemes such as root, stem and zero morphs may not be affixes, but they are valid morphs; on the other hand, prefix, infix and suffix are being defined as subclasses of `ontolex:Affix` in the LexInfo vocabulary (Cimiano et al., 2011). Note that earlier versions of the vocabulary defined various subclasses of morph:Morph. As this classification is partially redundant with a number of designated classes for morpheme types in LexInfo, we suggest to, instead, extend LexInfo, accordingly.

With respect to intensional morphology, morphs are the core elements of the `WordFormationRules` and `InflectionRules` that `involve` them, resp., `inflect` accordingly. Both kinds of rules are illustrated by an `example` (a string, for descriptive morphology) or by a `replacement`, i.e., a pair of regular expressions that define preconditions and results of a replacement, similar to the `s///` operator in Perl.

In agglutinating languages, inflection may be represented by a *sequence* of morphemes, where morphs (resp., their rules) follow a specific order, e.g., gender before number morphemes. This is modelled by a series of `InflectionTypes` connected by the `next` property. Inflection types are part of a paradigm, but also, directly linked with forms generated by the associated inflection rules. As a necessary pre-condition for generating them, `Paradigm` can also be directly linked with a lexical entry. In existing grammars, rules do not always use the canonical form of a lexical entry as the basis, and in these cases, the corresponding base form can be marked by `baseForm`. Furthermore, lexical entries can group together forms generated from multiple base forms, and if so, the base form and the associated inflection rule can be marked for a `baseType` (a system-specific identifier). A form can then consist of other forms or morphs, and it can be characterized by a `GrammaticalMeaning`, i.e., a feature bundle of LexInfo properties or other annotations, e.g., gloss labels (the latter are not defined by OntoLex-Morph). Likewise, a grammatical meaning can also characterize a morph, e.g., the inflectional features expressed or part of speech and morphosyntactic features of a derived word. Furthermore, grammatical meanings can formulate `baseConstraints` in derivation, e.g., restrictions on the part of speech of the base that a particular affix can be used with.

Whereas inflection operates on forms, derivation is a *lexical* process. So, a `WordFormationRule` generates a `ontolex:LexicalEntry`, either by means of a `DerivationRule` or a `CompoundRule`. Whereas a word formation rule formulates or illustrates a general pattern, the lexico-semantic relation between two concrete lexical entries (the base and the derivation or a constituent word and the compound) is modelled as a `WordFormationRelation`. In compounding, the head can be marked by the subclass `CompoundHead`, if no head is explicitly marked, one can either use `CompoundRelation` or the decomp module of OntoLex (Cimiano et al., 2016).

## 3. Selected Use Cases

We illustrate inflection, word formation and morphological generation with three examples from computational linguistics and computational philology.

### 3.1. Inflection in Modern Greek

LEXIS (Anagnostopoulou et al., 2000) is a computational lexicon for Modern Greek intended for use in

NLP applications and modelled according to the PAROLE/SIMPLE model (Parole Consortium, 1996).
The basic unit in LEXIS is the Morphological Unit (MU), a single word with its assigned part of speech tag, which corresponds to the traditional notion of "lemma". Each MU is linked to Graphical Morphological Units (GMu), which correspond to orthographic variants of the lexical entry (e.g., "τραίνο" and "τρένο" [*train*]). Inflectional information is attached at the GMu level, in the form of an inflectional paradigm and a number of stems, each of which takes a number. The inflectional paradigm (GInP) is like an abstract "inflectional table", where each row (corresponding to an abstract/prototypical inflected wordform) is the combination of a numbered stem, a specific ending (suffix) and a bundle of grammatical features (e.g., case, number, person, tense, etc.). Full wordforms are not included in LEXIS; in principle, they should be produced with a generation algorithm that exploits coindexing information in the entries of stems and inflectional paradigms.
For instance, the lemma "άνθρωπος" [*person*] is a common noun with two stems, the stems "άνθρωπ-" and "ανθρώπ-". These can be represented in Ontolex-Morph as `ontolex:Form` and related to the lemma via the `morph:baseForm` property. Each of them takes a number as a value for the property `morph:baseType`.

```
<anthropos>
    a ontolex:Word, morph:Morph ;
    rdfs:label "άνθρωπος"@el, "person"@en;
    lexinfo:partOfSpeech lexinfo:noun ;
    morph:paradigm <efyvos_paradigm> ;
    morph:baseForm [
        a ontolex:Form ;
        ontolex:writtenRep "άνθρωπ"@el ;
        morph:baseType "1" ] ;
    morph:baseForm [
        a ontolex:Form ;
        ontolex:writtenRep "ανθρώπ"@el ;
        morph:baseType "2" ] .
```

Following the Ontolex-Morph model, each prototypical word form can be represented as a `morph:InflectionRule` which takes for the property `morph:baseType` a literal in the form of a number and a property `morph:replacement`, which combines together a `morph:source` and a `morph:target`, the latter being the ending and the former assumed to be derived from the `morph:baseType`.

```
<inflRule_MaSgGe1>
    a morph:InflectionRule ;
    morph:baseType "1" ;
    morph:replacement [
        a morph:Replacement ;
        morph:source "$" ;
        morph:target "ου"@el ] .
```

The above `InflectionRule` can therefore generate the wordform:

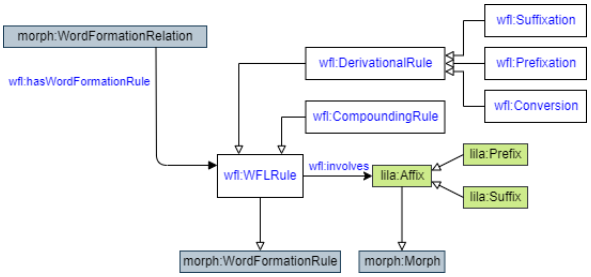


Figure 3: Architecture of the WFL ontology.

```
<anthropou1_form> a ontolex:Form ;
    ontolex:writtenRep "άνθρωπου"@el .
```

## 3.2. Word Formation in Latin

Word Formation Latin (Litta and Passarotti, 2019, WFL) is a derivational lexicon of Latin characterised by a step-to-step morphotactic approach: each lexeme is connected to the lexeme from which it is directly derived (if any) via a word formation rule. This resource has been recently modelled in an ontology (Pellegrini et al., 2021) in order to include it into the LiLa Knowledge Base[1] of interoperable linguistic resources for Latin (Passarotti et al., 2020).
The proposed modelling is fully compatible with the architecture of OntoLex-Morph as described above, and integrated with it. At the moment, this is done by specifying subclass relations, after the final release of OntoLex-Morph, we might directly use OntoLex-Morph vocabulary. So far, OntoLex-Morph and the WFL ontology were developed in parallel, but with mutual influences on each other. For example, Fig. 3 illustrates the distinction between relations and rules, as it is applied both in WFL and in OntoLex-Morph. Each `ontolex:LexicalEntry` of the WFL ontology is linked to the one(s) it derives from and/or to the ones that derive from it by means of a specific instance of the class `morph:WordFormationRelation`. In turn, each `morph:WordFormationRelation` is linked through the property `wfl:hasWordFormationRule` to a specific `wfl:WFLRule`. Rules are then arranged in a hierarchy of subclasses that reflects the distinction made in WFL between derivation (prefixation, suffixation, conversion) and compounding. Rules are also connected with the `lila:Affix` they display (if any) by means of the property `wfl:involves`.
For instance, there is a `morph:WordFormationRelation` between *felix* 'happy' and *felicitas* 'happiness', that instatiates the specific `wfl:WFLrule` creating the latter from the former. This rule belongs to the class of suffixal rules creating deadjectival nouns, which is a subclass of `wfl:Suffixation`. The rule is also stated to involve the suffix *-tas*:

```
:li_103068 a ontolex:LexicalEntry ;
```

[1] `https://lila-erc.eu`.

```
    rdfs:label "felix" .

:li_103063 a ontolex:LexicalEntry ;
    rdfs:label "felicitas" ;

:r18023_li_103068_li_103063 a
  morph:WordFormationRelation ;
    vartrans:source :li_103068 ;
    vartrans:target :li_103063 ;
    wfl:hasWordFormationRule
:Derivation_Suffix_li_103068_To_li_103063.

:Derivation_Suffix_li_103068_To_li_103063
  a wfl:AdjectiveToNoun ;
  rdfs:label
"felix To felicitas involves -tas/tat" ;
  wfl:involves
<http://lila-erc.eu/data/id/suffix/24> .

<http://lila-erc.eu/data/id/suffix/24> a
  lila:Suffix ;
    rdfs:label "-tas/tat" .

wfl:AdjectiveToNoun rdfs:subClassOf
  wfl:Suffixation .
```

It is useful to spend a few words on the treatment of compounding in the WFL ontology. As can be seen below, compounds are modelled in the same way as other morphologically complex words, except that there are two relations: the one between the compound and its first constituent on the one hand, the one between the compound and its second constituent on the other hand. Both relations point to the same rule. The order of constituents is coded via a datatype property `wfl:positionInWFR`. This choice is motivated by the fact that the class `morph:WordFormationRelation` is a sub-class of `vartrans:LexicalRelation`, from which it inherits the requirement of having exactly one source and one target.

```
:li_88060 a ontolex:LexicalEntry ;
    rdfs:label "ager" .

:li_94916 a ontolex:LexicalEntry ;
    rdfs:label "colo" .

:li_88174 a ontolex:LexicalEntry ;
    rdfs:label "agricola" .

:r8833_li_88060_li_88174 a
  morph:WordFormationRelation ;
  rdfs:label "ager > agricola" ;
  wfl:positionInWFR 1 ;
  wfl:hasWordFormationRule
:Compounding_li_88060_li_94916_To_li_88174;
  vartrans:source :li_88060 ;
  vartrans:target :li_88174 .

:r8833_li_94916_li_88174 a
  morph:WordFormationRelation ;
    rdfs:label "colo > agricola" ;
  wfl:positionInWFR 2 ;
  wfl:hasWordFormationRule
:Compounding_li_88060_li_94916_To_li_88174;
    vartrans:source :li_94916 ;
    vartrans:target :li_88174 .
```

It has been mentioned above that compounding can also be modelled using the vocabulary of the Decomposition module of OntoLex. However, this option is not adequate to model WFL data. First, it is not desirable to be forced to use different vocabularies for word formation processes that are treated homogeneously in WFL – namely, OntoLex-Decomp for compounding and OntoLex-Morph for derivation and conversion. Second, OntoLex-Decomp does not allow to reify the relations between the lexical entries involved in compounds, but these relations are needed in order to provide a connection to the compounding rules present in WFL, as can be seen in the listing above.

This motivates the choice of giving the possibility of modelling compounding also using OntoLex-Morph, alongside OntoLex-Decomp. The choice between the two is left to the data creator, as it crucially depends on the nature and organization of the data themselves: if compounds are simply split into their different constituents, then OntoLex-Decomp will suffice; if additional information is provided and/or compounding is treated by means of full-fledged relations – as happens not only in WFL, but also in other important resources tackling derivational morphology, e.g. DeriNet 2.0 (Vidra et al., 2019) – then it will be possible (or even necessary) to resort to Morph.

### 3.3. Generation for German

Chiarcos et al. (accepted) recently described the application of OntoLex-Morph to convert and link various morphological resources for German. While the focus of this work was primarily on the encoding and integration of different types of morphological resources on a unified basis, the capacity to merge is a trivial (and intended) side-effect of RDF conversion and was the general purpose and original motivation of OntoLex-Morph (Klimek et al., 2019).

In this section, we focus on morphological generation rules resulting from converting an FST grammar, as this aspect was only superficially touched by Chiarcos et al. (accepted): We transform a German finite state transducer into OntoLex-Morph, the Stuttgart FST library with the SMOR grammar (Schmid, 2005) and the Morphisto lexicon (Zielinski et al., 2009). In order to replicate complete finite state transducers in OntoLex-Morph, we made use of the `morph:InflectionType` concept. In this conversion, every state is represented by an independent inflection type, and transitions between states are modelled by means of `morph:next`. For generation, we use a simple path traversal over these inflection types to retrieve a sequence of replacements. In this case, however, the traversal is not conducted as part of the ac-

tual generation process, but with the goal of achieving optimal run-time performance. Instead sequences of Perl-style replacements are compiled out, where regular expressions and capturing groups are used to emulate the effect of replacement operations associated with state transitions in the underlying transducer. The resulting sequences of replacement operations can subsequently be executed in any programming language that supports regular expressions. So, instead of doing morphological generation directly, they, instead, bootstrap a morphological generator from OntoLex-Morph. The operation needed to create a morphological generator from OntoLex-Morph inflection rules is a single SPARQL query that traverses the sequence of inflection types and collects a series of replacement operations as defined in the inflection rules:

```
SELECT DISTINCT ?itype ?transformation
WHERE {
 { SELECT ?a ?end ?pathid
   (GROUP_CONCAT(?repl; separator=";")
    AS ?transformation)
   WHERE {
    ?a a morph:InflectionType.
    ?a morph:next* ?b.
    ?b morph:inflectionRule ?rule.
    ?rule morph:replacement ?repl.
    FILTER(?repl != "s/$//")
    ?b morph:next* ?end.
    MINUS { ?end a morph:InflectionType;
                 morph:next [] }
  } GROUP BY ?a ?end
 }
 BIND(?a as ?itype)
}
```

This query uses a nested `SELECT` to aggregate from one inflection type to the sequence of following inflection types.[2] The result of this aggregation query, then, is a sequence of replacements with regular expressions that can be directly executed with Perl, or Sed, or transformed with minimal overhead to any other programming language that support Perl-style regular expressions (e.g., Java, Python, ... or even SPARQL). The actual generator is therefore a thin wrapper around this query to create a replacement script, whereas the OntoLex lexicon is not *directly* used for the transformation. The replacement script then reads lexical entries with their base forms (or, if these are not available, canonical forms), part of speech information (represented using the LexInfo property `lexinfo:partOfSpeech`) and paradigms. The replacement script assumes the presence of special symbols that trigger generation rules (e.g., `<Sg>` for singular number, etc.) In the current implementation, these are automatically created from an existing OLiA Annotation Model for the Morphisto morphology (Chiarcos, 2010).

[2]The query is slightly simplified, we omit the creation of pathids.

Take the inflection of *zufällig* adj. 'random', this is marked as `lexinfo:partOfSpeech lexinfo:adjective`, so that one of the compatible annotation model concepts is `:ADJ_Pos_Fem_Nom_Sg`. As this carries `olias:hasTagEndingWith "<+ADJ><Pos><Fem><Nom><Sg>"`, and the base form is `zufällig`, this is concatenated into the input string `zufällig<+ADJ><Pos><Fem><Nom><Sg>` (which is paired with information about the associated lexinfo features). The associated paradigm in Morphisto is `:paradigm%23Adj%2B` from which we get to the inflection type `:type%23Adj%2B`. For this inflection type, the above query retrieved (among other possible paths) the replacement series illustrated in Fig. 4 (with replacement results for our input word added as comment).

While this works sufficiently well, Chiarcos et al. refrained from modelling morphophonological operations by means of this technology.[3] Instead, special symbols intended for resolution with two-level rules were simply omitted. The SPARQL query thus adds two additional replacements: The first marks forms that contain unresolved tags as being heuristic (insertion of initial $\star$), and the second removes all tags. The result string is $\star$`zufällige` and (except for the unimplemented lower case rule), this is actually the correct form. However, morphophonological rules do at times have a profound impact on the surface form of an inflected word, e.g., the insertion of epenthetic vowels or assimilation effects. As a result, a certain number of hypothetical forms predicted by such replacements are indeed, ungrammatical. But even if they are grammatical, they need to be marked in the data. We thus suggest to introduce the novel class hypothetical form. This is necessary for Morphisto because the capturing of morphophonological alternations may imperfect. So, any form produced by the application of rules to a base or canonical form is hypothetical unless confirmed by external evidence from a dictionary or a corpus.

For inflection, the Morphisto generator produced over 400,000 triples for 41,100 hypothetical forms. For evaluation, we evaluate the generated hypothetical form by parsing them with the Ubuntu 20.4 package 'fst', and the Morphisto/SMOR grammar. Out of 25,859 different written representations of the generated hypothetical forms (excluding those identical to base forms), our generator achieved a precision of 78.5% against SMOR/Morphisto,[4] i.e., 20,308 of 5,551 written representations of hypothetical forms (exclud-

[3]Technically, this would have been possible, but the application of a concept named 'inflection type' to assimilation rules would contradict linguistic intuitions associated with the term 'inflection type'.

[4]We do not calculate recall, as our conversion only encompasses the inflection component of of SMOR, and neither the derivation nor compounding rules that it also provides.

```
# from OntoLex-Morph
s/$/<FB>/;              # zufällig<+ADJ><Pos><Fem><Nom><Sg>
s/<+ADJ><Pos>//;        # zufällig<Fem><Nom><Sg>
s/<Fem><Nom><Sg>/e/;    # zufällige
s/$/<Low#>/;            # zufällige<Low#>

# remove special symbols that trigger morphophonological
# replacements
s/^\\(.*<\\)/\\*\\1/;   # *zufällige<Low#>
s/<[^>]*>//g;           # *zufällige
```

Figure 4: Automatically generated sequence of replacement operations retrieved from the OntoLex-Morph edition of SMOR/Morphisto, using the word *zufällig* 'random' with grammatical features as sample input

ing those identical to the base form) could be successfully parsed. As for the remaining 21.5%, these can be attributed to the insufficient support for morphophonological rules in OntoLex-Morph as well as invalid combinations of alternative base forms and inflection rules that are filtered out in SMOR in subsequent processing steps.

It is to be noted that vanilla morphological generation from OntoLex-Morph is a baseline functionality that has advantages in portability and sustainability, but that it lacks optimizations of FST, e.g., in disambiguation strategies and filtering conditions performed at the second level of two-level morphologies.

## 4. Discussion

The goal of this paper was to demonstrate to what extent the OntoLex-Morph vocabulary in its most recent edition can be used for modelling existing lexical resources concerned with or designed for computational morphological analysis and generation. This complements the work of Klimek et al. (2019) who discussed applications of OntoLex-Morph for descriptive morphological analysis in the realm of digital lexicography with a more technically oriented perspective. In particular, we aimed to evaluate its applicability to broad band-width of use cases in this domain, illustrated here for three representative resources.

### 4.1. Achievements

Providing morphological datasets as OntoLex and in RDF provides the natural benefits of linkability, in this paper, we thus focus on the *coverage* of OntoLex-Morph for representative use cases, focusing on language technology resources for inflectional morphology (for Greek), derivation and compounding (for Latin) and the general usability for morphological generation (for German). By focusing on existing resources in three different languages, we also expect a certain degree of heterogeneity in the requirements.

**Linkability and (Re-)Usability** Overall, using OntoLex and OntoLex-Morph for machine-readable dictionaries and morphological resources has the great advantage that these can be trivially linked, merged and integrated. This is a general characteristic of RDF and LLOD technology and to establish a community standard that facilitates such integration operations over legacy as well as digital-born data has been the initial motivation for developing OntoLex and OntoLex-Morph. Unsurprisingly, this has been repeatedly confirmed since, e.g., for lexical resources and knowledge graphs (McCrae et al., 2011), lexical resources with other lexical resources (Eckle-Kohler et al., 2015), lexical and morphological resources (Racioppa and Declerck, 2019) and morphological resources with other morphological resources (Chiarcos et al., accepted). We thus consider the benefit of *linkability* for morphological resources to be sufficiently established by earlier research – as well as the benefits that this entails with respect to representation and modelling (graphs can represent any linguistic data structure), structural and conceptual interoperability (generic data structures, shared vocabularies, uniform access protocol), federation (querying over distributed data), dynamicity (access remote resources at query time) and the availability of a mature technical ecosystem (Chiarcos et al., 2013; Cimiano et al., 2020). But these benefits are inherent to LLOD and not specific to OntoLex-Morph, so we did not specifically evaluate them.

**Applicability** Overall, we found that the OntoLex-Morph vocabulary was applicable to the resources addressed in this paper with relative ease. Although we encountered a number of borderline cases in which the current modelling leaves up either challenges or desiderata (see below), the typical cases could be represented in OntoLex-Morph, for inflection (Sect. 3.1), for word formation (Sect. 3.2) and for morphological generation in general (Sect. 3.3). We used the experiences we made while applying the OntoLex-Morph vocabulary on novel data and questions that were raised in the process to refine and clarify the current model draft.[5]

**Rule-based generation** In Sect. 3.3, we described how OntoLex-Morph resources can be used to bootstrap replacement scripts that emulate finite state trans-

[5] https://github.com/ontolex/morph/blob/master/draft.md

ducers by means of regular expressions. This is only a baseline functionality as aspects of morphophonology have not been addressed, but only "deep" morphology, but it was nevertheless successful in achieving a considerable degree of precision with a formalism (Perl-style regular expressions) that can be easily ported into any programming language, whereas the original FST grammar depended on a 2005 library (Schmid, 2005).

## 4.2. Challenges

**Variation in inflection** Another challenge which we are focusing on as part of the development of OntoLex-Morph is the representation of variants. This occurs, for instance, when more than one form realises the same cell in an inflection table for a given paradigm; this is also known as *overabundance* (Thornton, 2019). This can be due to dialectal, diachronic or simply orthographic variation. It is more common to have such variants in the case of languages without a standardised orthography and especially historical languages such as Old English. Indeed, it is not difficult to find examples in the latter, e.g., the first person preterite indicative form of the verb *cuman* 'to come' is often listed as both *cwom* and *com*. Overabundance is also widely attested in Latin data, where especially interesting are cases of lexemes that display variation between forms that belong to different inflection classes, for instance LAVO 'wash', that can be inflected according to either the 1st (e.g. PRS.ACT.INF *lavare*) or 3rd (e.g. PRS.ACT.INF *lavere*) conjugation. We are thus clearly dealing with morphological (rather than simply orthographic) variation. A current challenge is to find a systematic way of dealing with these cases that is compatible with the generative component of OntoLex-Morph. A related problem is suppletion, i.e., cases in which different forms of the same lexical entry are formed from different etymological roots. This is the case of the Old English verb *wesan* 'to be' whose infinitive represents one underlying root, whereas its indicative present singular forms are based on *two* other roots (*eom* 1.sg. '(I) am'; *bist* 2.sg. '(you) are'). This pattern is also preserved in modern English, and with once-regular morphological processes getting increasingly intransparent over time, has even expanded to form novel pairs of 'irregular' forms that appear to operate with different stems, e.g., in verbs like *bring* and *think*, whose nasal complement was lost in the past forms *brought* and *thought* after Germanic *-kt-* shifted to Old English *-ht-*. The same pattern is also observed in modern Greek, where alternative wordforms for the same grammatical meaning co-exist. Alternatives may be associated with alternative endings, e.g. πατέρ-ες and πατερ-άδες or alternative stems, as in the example listed in section 3.1, i.e. άνθρωπ-ου and ανθρώπ-ου. One of the forms may be marked as a dialectal, archaic, more formal, or colloquial variant but there are also cases where the two forms are just alternatives; such a case is that of contracting verbs, e.g. αγαπ-άω and αγαπ-ώ.

**Phonological processes** As mentioned in Sect. 3.3, only rules concerned with 'deep morphology' have been formalized, but not morphophonological processes that deal with phonological processes like assimilation or apophony, i.e., the second level in classical two-level morphologies. A particular problem here is that, at least in word formation in Latin, these are not fully predictable, and this prevents the simple juxtaposition of formative elements from generating the actual surface form of derivatives.

**Markers of morphological variation** When modelling linguistic variation at the morphological level, we are faced with the need for attributing markers (labels of style, dating, dialect, etc.) to wordforms, in the same way that traditional dictionaries assign them to lemmas. That is, as we have archaic, older, dialectal, formal lemmas, we also have inflectional variants that can be marked. For instance, in the example (Sect. 3.1), the form άνθρωπου is used in a more informal context compared to ανθρώπου. In modern Greek, a lot of dual wordforms originate from "katharevoussa"[6]. It remains an open question whether and how these markers would be modelled within the morph module in a uniform and generic way, and specifically in inflection rules so that a mechanism could be triggered to copy these markers (together with grammatical features) to the generated written forms as well, while keeping the model simple.

At the moment, we would consider such markers to be beyond the scope of OntoLex-Morph. It is, of course, necessary for successfully generating context-adequate forms, but we would see the individual attributes and features more in the general scope of the LexInfo vocabulary. Indeed, LexInfo provides a rudimentary vocabulary, e.g., with `lexinfo:register` and values such as `lexinfo:dialectRegister`, with `lexinfo:temporalQualifier` and values such as `lexinfo:archaicForm`, or with `lexinfo:dating` and values such as `lexinfo:old`. Neither of these terms fits katharevousa directly, but, in fact, a language-specific instance of `lexinfo:Register` or `lexinfo:TemporalQualifier` could also be created – unless the data providers decide to live with the imprecision of standard LexInfo terminology. However, what is important with regard to OntoLex-Morph is that it must provide the necessary prerequisites for adding such markers to morphologically relevant data structures, (morphological rules, lemmas, forms, etc.), i.e., they must be concepts, not properties. And, indeed, this is the case already. But even in this case, it would be desirable if the OntoLex-Morph vocabulary would eventually be accompanied by best practice recommendations for the assignment of markers and provenance.

[6]Katharevoussa is an archaic form of Greek constructed on the basis of the Attic dialect and used in formal settings; although its use is fading, it is still encountered in older texts.

### 4.3. State of Modelling

The OntoLex-Morph diagram has changed significantly since (Klimek et al., 2019), but only few vocabulary elements have changed their definition.[7] We thus consider the vocabulary stable, and revisions are now limited to cases when a change in the vocabulary meets the needs of *multiple* data providers or potential users. Selected suggested revisions include the revision of inflection type and the extension of LexInfo.

**Inflection type** An aspect that is still under discussion, as it can pose non-trivial problems when modelling data with this module, concerns the class `morph:InflectionType`. Since it was intended to account for the different slots available for values of different morphosyntactic properties in agglutinative languages, such problems emerge especially in fusional languages like Greek and Latin, where there are no such slots and the different values are expressed cumulatively by means of the same affix.

In Latin – like in many other languages – inflection rules are sensitive to inflection class distinctions: for instance, the rule to obtain the PRS.ACT.IND.2SG from the infinitive of 1st conjugation verbs (e.g. *amare* → *amas* 'to/you love') is different than the one of 3rd conjugation verbs (e.g. *dicere* → *dicis* 'to/you say'). Inflection classes can easily be coded as instances of the class `morph:Paradigm`. However, given this state of affairs it could be useful to have a property linking each `morph:Paradigm` to all the `morph:InflectionRules` it consists of, without having to go through `morph:InflectionType` as required in the current draft.

As the inflection type class has been created for agglutinating, not inflecting languages, it is unsurprising that it seems to be unnecessary here, and could be replaced by a direct link to inflection rule. At the same time, we suggested a novel application of inflection type to encode finite states, and it was mostly terminological issues that kept us from modelling morphophonological processes with 'inflection type', so that we suggested to model the order of morphemes as a sequence of inflection rules, instead, as their naming is less confusing.

A possible revision that would cater all three requirements would be to eliminate inflection type completely, i.e., to transfer all its properties to inflection rule, to connect grammatical meaning with inflection rule, and to encode the information that inflection type was originally meant for (position of a morphological 'slot' and its characteristics) as part of `GrammaticalMeaning`. This modelling, however, needs to be evaluated for its application to agglutinative languages and the original intended application of inflection type to represent morphological 'slots'.

**LexInfo** A number of suggested additions to LexInfo have been mentioned throughout this paper. This includes the introduction of additional subclasses of `ontolex:LexicalEntry` and `morph:Morph` to complement the classes `lexinfo:Suffix`, `lexinfo:Prefix` and `lexinfo:Infix` that LexInfo currently defines as subclasses of `ontolex:Affix`. In addition to subclasses of `ontolex:Affix`, we would require `lexinfo:RootMorph` and `lexinfo:StemMorph` as subclasses of `morph:Morph`, resp., `ontolex:LexicalEntry`.

A possible addition to LexInfo is in subproperties and object values of `ontolex:usage`, where morphological resources call for introducing object values such as `lexinfo:hypotheticalForm` (or, `lexinfo:nonattestedForm`), `lexinfo:reconstructedForm` and `lexinfo:incorrectForm`, which can be modelled in analogy to the properties `lexinfo:register`, and `lexinfo:domain` by means of a property `lexinfo:evidence`.

## 5. Summary and Outlook

In this paper, we described the recent extension of OntoLex-Morph with respect to computational morphology, and in particular, vocabulary elements necessary for describing morphological generation by means of rules, forms and morphs. This paper complements our earlier work on OntoLex-Morph (Klimek et al., 2019) that took a stronger focus on requirements from lexicography and the language sciences, and with the recent extensions, the overall structure of the vocabulary has been considerably extended. Taking the results of both papers together, we cover two major strands of use cases for an OntoLex Morphology module, so that after more than five years of development within the W3C Community Group Ontology-Lexica, the OntoLex-Morph vocabulary can now be considered relatively mature and stable.

Despite the advanced state of affairs after five years of development in this community, there are some limitations as pointed out in Sect. 4 that we plan to address in the next months. After having demonstrated that we cover requirements from both lexicography and language technology, we will work on consolidating the OntoLex-Morph vocabulary in order to prepare its final publication, probably in 2023. The primary goal of this paper and our presentation is two-fold: On the one hand, it documents the recent extensions, and on the other hand, it aims to elicit feedback from reviewers and audience to take into account before publishing it as a W3C vocabulary in the form of a community report of the W3C Ontology-Lexica Community Group.

[7] The most significant change in the overall model is that we now define `morph:Morph` as a subclass of lexical entry rather than as an independent concept, so that the existing `ontolex:Affix` class can now be interpreted as a subclass of `morph:Morph` and that the redundancy between `ontolex:Affix` and `morph:AffixMorph` is eliminated.

## 6. Acknowledgements

The research described in this paper was conducted in the context of the Cost Action CA18209 *Nexus Linguarum. European network for Web-centred linguistic data science*, with partial support from the Research Group *LiODi. Linked Open Dictionaries* (funded by the German Federal Ministry of Education and Science, BMBF, 2015–2022), from the ERC Horizon 2020 research and innovation programme via the projects Prêt-à-LLOD (grant agreement no. 825182), and Linking Latin (grant agreement no. 769994) and from "APOLLONIS: Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation" (MIS 5002738), (co-financed by Greece and the EU). Moreover, the authors would like to thank Bettina Klimek and Maxim Ionov for coordinating the development of OntoLex-Morph in the last years, as well as all OntoLex-Morph contributors.

## 7. Bibliographical References

Anagnostopoulou, D., Desipri, E., Labropoulou, P., Mantzari, E., and Gavrilidou, M. (2000). Lexis - Lexicographical Infrastructure: Systematising the Data. In *Proceedings of the International Workshop on Computational Lexicography and Multimedia Dictionaries (COMPLEX 2000)*, Patras, Greece.

Bosque-Gil, J. and Gracia, J. (2019). The Ontolex Lemon Lexicography Module. Final Community Group Report. Technical report, W3C.

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., Fäth, C., and Ionov, M. (accepted). Unifying Morphology Resources with Ontolex-Morph. a Case Study in German. In *13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France, June.

Chiarcos, C. (2010). Towards robust multi-tool tagging. an OWL/DL-based approach. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Cimiano, P., McCrae, J., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. Technical report, W3C.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data*. Springer.

Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. (2015). lemonUby – a large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6(4):371–378.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., and Chiarcos, C. (2019). Challenges for the Representation of Morphology in Ontology Lexicons. *Proceedings of eLex*, pages 570–591.

Litta, E. and Passarotti, M. (2019). (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, et al., editors, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, December.

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

Parole Consortium. (1996). Morphosyntactic specifications : Language Specific Instantiations. Technical report, LE-PAROLE report.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Pellegrini, M., Litta, E., Passarotti, M., Mambrini, F., and Moretti, G. (2021). The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 101–109.

Racioppa, S. and Declerck, T. (2019). Enriching Open Multilingual Wordnets with Morphological Features. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*.

Schmid, H. (2005). A Programming Language for Finite State Transducers. *Finite-State Methods and Natural Language Processing FSMNLP 2005*, page 50.

Thornton, A. M. (2019). Overabundance: A Canonical Typology. In *Competition in Inflection and Word-Formation*, pages 223–258. Springer.

Vidra, J., Žabokrtskỳ, Z., Ševčíková, M., and Kyjánek, L. (2019). Derinet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89.

Zielinski, A., Simon, C., and Wittl, T. (2009). Morphisto: Service-Oriented Open Source Morphology for German. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 64–75. Springer.

# Author Index