Enhancing Interoperability for Under-Resourced Languages: A Case Study on Linking Lithuanian-English Data in the Cybersecurity Domain

Christian Chiarcos
University of Augsburg, Germany
christian.chiarcos@uni-a.de

Maxim Ionov
University of Cologne, Germany
mionov@uni-koeln.de

Andrius Utka
Vytautas Magnus University, Lithuania
andrius.utka@vdu.lt

Sigita Rackevičienė
Mykolas Romeris University, Lithuania
sigita.rackeviciene@mruni.eu

Relevant UniDive working groups: WG2, WG4, WG3

1 Introduction

The paper presents ongoing work on converting bilingual (Lithuanian-English) textual datasets on cybersecurity (a terminology resource and corpora) into a linkable format and linking them to each other, as well as integrating them into the Linked Language Open Data (LLOD) Cloud. This is motivated by two key factors: the under-resourced state of the Lithuanian language and the ever-growing relevance of cybersecurity and its terminology in contemporary discourse.

Despite notable progress in the data field, Lithuania still lacks language resources for the development of state-of-the-art technologies. Existing resources are in diverse formats, making them difficult to discover and use. The conversion of Lithuanian language datasets into a linkable (RDF) format, which enables their integration into the global linguistic data eco-system and makes them easily discoverable and interoperable, and thus contributes to a more diverse and inclusive linguistic landscape.

The cybersecurity domain has been chosen for the linking case study as it is one of the most rapidly evolving domains with great demands for Lithuanian terminology development. The collection of bilingual (English-Lithuanian) data, their structuring, and making them globally interoperable and discoverable are essential for development and dissemination of Lithuanian terminology and integrating it in the overall linguistic eco-system.

2 Datasets

The datasets chosen for our case study are of two types:

(1) bilingual English-Lithuanian corpora: a 1.4M token parallel corpus in TMX, accompanied by a comparable corpus (4M words), both also with morphological annotations in the Vert TSV format (a tab-separated format akin to commonly used CoNLL-TSV formats);

(2) a bilingual terminological dataset in the TermBase eXchange format (TBX), an XML-based international standard for exchange of terminological information (ISO/TC 37/SC 3, 2019).

The corpora contain bilingual cybersecurity texts produced in the time period of 2010-2021 and representing various discourses: legal, administrative-informative, academic and media. The corpora are available in the CLARIN-LT repository (Utka et al., 2022b,a).

The bilingual terminological dataset is exported from the Lithuanian-English Cybersecurity Termbase which is developed on the basis of the data extracted from the above-described corpora. The current TBX version contains 233 cybersecurity concepts with their corresponding Lithuanian and English terminological designations, definitions and context examples. Every entry is composed of the following levels and data categories:

Concept level categories

- Concept id
- Subject field (subdomain and sub-subdomain to which the concept belongs)

Language level categories

- Term group consisting of (1) Term and (2) TermNote (indication of frequency of a term in the corpora) for up to three synonymous terms denoting the same concept in each language.
- Description group 1: (1) Definition; (2) Definition source
- Description group 2: (1) Context example; (2) Context example source

- Description group 3: A list of other synonymous terms (if more than 3)

The Lithuanian-English Cybersecurity Termbase (Lithuanian title: Lietuvių-anglų kalbų kibernetinio saugumo terminų bazė) is located on the Terminologue platform (Utka et al., 2023a). The TBX file is also available in the CLARIN-LT repository (Utka et al., 2023b).

3 TBX conversion

To provide the means for linking between the dataset and the corpora and to integrate the data further with the external resources, we convert it to Linked Data. To do so, we use the OntoLex vocabulary (McCrae et al., 2017), a *de facto* standard for modelling lexical resources as Linguistic Linked Open Data. Various methods for such a conversion have been proposed (Cimiano et al., 2015; Montiel-Ponsoda et al., 2015; Di Buono et al., 2020) as well as studies on the mismatches between representations and how to overcome them (Bellandi et al., 2023; Martín-Chozas and Declerck, 2022).

To be able to carefully address all potential issues and idiosyncrasies, we implement the converter for this particular dataset from scratch. One such idiosyncrasy is a subject field that simultaneously encodes the subject field of a term and the hierarchy of that subject field. To make this compatible with OntoLex-Lemon, we created a SKOS[1] dataset that models each subject as a top concept of `skos:ConceptScheme` with `skos:narrower` and `skos:broader` properties between them. Each concept is defined as `skos:Concept` and belongs to `skos:ConceptScheme`. Each synonymous term of a concept corresponds to a separate `skos:LexicalEntry`. Both definitions and usage examples are modelled using SKOS properties (`skos:definition` and `skos:example`, respectively) pointing to blank nodes that combine a value with the source:

```
:eid-6 a skos:Concept ; skos:definition [
  rdf:value "<...>"@lt ;
  dct:source "KS LT PALYGINAMASIS" ] .
```

This is consistent with the current recommendations from the Terminology module discussions and is subject to change with the introduction of the module.[2]

[1] https://www.w3.org/2009/08/skos-reference/skos.html

[2] https://www.w3.org/community/ontolex/wiki/Terminology#Open_Issues

4 TMX and TSV conversion

The Translation Memory eXchange (TMX) format is an XML-based standard for storing and exchanging translation memories. Aside from textual metadata, the current TMX 1.4b specification defines translation units (`<tu>`) which group together one or more translation unit variants (`<tuv>`) per language. Each of these can contain one or more segments (`<seg>`, e.g., sentences) with plain text data. In addition, the Vert format provides linguistic annotations (parts of speech, morphosyntactic features, lemma) for both Lithuanian and English text in different files. We combine common vocabularies to expose this information as RDF:

Web Annotation (Ciccarese et al., 2013) for doing standoff annotations over web documents
NLP Interchange Format (Hellmann et al., 2013, NIF) for annotating strings in a web document, and
CoNLL-RDF (Chiarcos and Fäth, 2017) for the linguistic annotation of NIF words and sentences.

We provide URIs for `<tuv>` and `<tu>` as Web Annotation annotation elements that point into the original TMX data by means of an XPath Selector. Each annotation created for a `<tu>` element then provides its string value as a NIF "context". Every sentence (`<seg>`) within the `<tu>` corresponds to a NIF sentence as generated from the Vert annotation and is linked as a NIF substring. With the OntoLex-FrAC vocabulary, each lexical entry is now assigned the relevant `<tu>` elements (resp., their Web Annotation URIs) by means of `frac:attestation`.

5 Outlook

The data structures defined in the two preceding sections allow us now to link the terminology repository and the corpus data — and thus, to access. to query, and to process all resources mentioned before in a unified fashion — or to enrich it further with external knowledge graphs and lexical resources. To the best of our knowledge, this combination of vocabularies has not been proposed before for multilingual terminology data and translation memories, and we would like to discuss our proposal and a number of challenges associated with it, e.g., in the treatment of multi-word expressions, in the context of UniDive. Furthermore, we believe this to be a suitable technology and a proof-of-principle implementaion that could inspire the UniDive lexicon-corpus interface.

6 Acknowledgements

The work described in this paper was conducted in the context of the COST Action *Nexus Linguarum. European network for Web-centred linguistic data science* (CA18209, 2019-2024). We would like to thank three anonymous reviewers for support and feedback and their encouragement to discuss this line of research in the context of UniDive.

References

Andrea Bellandi, Giorgio Maria Di Nunzio, Silvia Piccini, and Federica Vezzani. 2023. The importance of being interoperable: Theoretical and practical implications in converting tbx to ontolex-lemon. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 646–651.

Christian Chiarcos and Christian Fäth. 2017. Conll-rdf: Linked corpora done in an nlp-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Paolo Ciccarese, Stian Soiland-Reyes, and Tim Clark. 2013. Web annotation as a first-class object. *IEEE Internet Computing*, 17(6):71–75.

Philipp Cimiano, John P McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517.

Maria Pia Di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. Terme-a-llod: Simplifying the conversion and hosting of terminological resources as linked data. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28–35.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

ISO/TC 37/SC 3. 2019. ISO 30042:2019 Management of terminology resources. TermBase eXchange (TBX). Technical report, ISO.

Patricia Martín-Chozas and Thierry Declerck. 2022. Representing multilingual terminologies with ontolex-lemon. In *Proceedings of the 1st International Conference on Multilingual digital terminology today. Design, representation formats and management systems*.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Elena Montiel-Ponsoda, Julia Bosque-Gil, Jorge Gracia, and Daniel Vila-Suero. 2015. Towards the integration of multilingual terminologies: an example of a linked data prototype. In *Terminology and Artificial Intelligence (TAI)*, pages 205—206.

Andrius Utka, Sigita Rackevičienė, Agnė Bielinskienė, Marius Laurinaitis, Liudmila Mockienė, and Aivaras Rokas. 2023a. Lithuanian-english cybersecurity termbase. Terminologue.

Andrius Utka, Sigita Rackevičienė, Agnė Bielinskienė, Marius Laurinaitis, Liudmila Mockienė, and Aivaras Rokas. 2023b. Lithuanian-english cybersecurity termbase v.0.1. CLARIN-LT digital library in the Republic of Lithuania.

Andrius Utka, Sigita Rackevičienė, Aivaras Rokas, Agnė Bielinskienė, Liudmila Mockienė, and Marius Laurinaitis. 2022a. English-lithuanian comparable cybersecurity corpus - DVITAS. CLARIN-LT digital library in the Republic of Lithuania.

Andrius Utka, Sigita Rackevičienė, Aivaras Rokas, Agnė Bielinskienė, Liudmila Mockienė, and Marius Laurinaitis. 2022b. English-lithuanian parallel cybersecurity corpus - DVITAS. CLARIN-LT digital library in the Republic of Lithuania.

Appendix: Example

We provide a minimalistic example for illustration. The representation of lexical data in OntoLex-Lemon (from TBX) and of one-word-per-line annotations in NIF/CoNLL-RDF is established practice. The modelling of parallel data (from TMX) and the linking between different corpus sources and lexical resources are innovative aspects of our work.

TMX Modelling

Consider the following TMX excerpt

```
<tu creationdate="20210610T095316Z" creationid="LF
    Aligner 4.21">
 <tuv xml:lang="EN">
   <seg>They are the cornerstone for achieving the
        digital single market.</seg>
 </tuv>
 <tuv xml:lang="LT">
  <seg>Jos yra skaitmeninės bendrosios rinkos
       kūrimo pagrindas;</seg>
 </tuv>
</tu>
```

The custom datatype `:TranslationSet` represents translation units as annotations in RDF, using the the Web Annotation vocabulary (`oa:`) to point into the original XML document by an XPath. The objects `source:str_48` and `source:str_47` are the English and Lithuanian text of the translation unit, translation unit variants are defined by reference to the same source strings:

```
:tu_24 a oa:Annotation, :TranslationSet;
 oa:hasTarget source:str_48, source:str_47.
source:str_48
  oa:hasSource <https://.../EUR-Lex_001.tmx> ;
 oa:hasSelector [
   a oa:XPathSelector;
   rdf:value "/tmx/body/tu[24]/tuv[2]" ] ].
:tuv_48 a oa:Annotation;
  oa:hasTarget source:str_48 .
```

We annotate these objects with their their string value as `nif:Contexts`:

```
:tuv_48 a :NIFStringAnnotation ;
  oa:hasBody :tuv_48_str .
:tuv_48_str a nif:Context ;
  nif:isString "Jos yra skaitmeninės ...;"@lt.
```

Modelling One-Word-Per-Line Annotations

The morphosyntactically annotated (Vert) files use a CoNLL-style format extended with SGML:

```
<s>
Jos         jis  Pgfsgn         jis-p
yra         būti Vgmp3---n--ni- būti-v (etc.)
```

We use the CoNLL-RDF library to convert this to a minimal NIF sub-vocabulary:

```
:s24_0 a nif:Sentence .
:s24_1 a nif:Word ; nif:nextWord :s24_2 ;
       conll:WORD "Jos" ;   conll:LEMMA "jis";
       conll:POS "Pgfsgn";  conll:TERM "jis-p";
       conll:HEAD :s24_0 .
```

The `nif:Context` of the translation unit variant is a NIF string object, as is the sentence produced by CoNLL-RDF, so, they are connected by means of a `nif:subString` property:

```
:s24_0 nif:subString :tuv_48_str.
```

Modelling and Linking TBX Data

We illustrate TBX modelling for the term *DoS ataka* as modelled by our converter:

```
:eid-6 a skos:Concept .

:DoS+ataka-lt a ontolex:LexicalEntry ;
    ontolex:canonicalForm :DoS+ataka-lt_form ;
    ontolex:sense :DoS+ataka-lt_sense .

:DoS+ataka-lt_form a ontolex:Form ;
    ontolex:writtenRep "DoS ataka"@lt .

:DoS+ataka-lt_sense a ontolex:LexicalSense ;
    ontolex:isSenseOf :DoS+ataka-lt ;
    ontolex:reference :eid-6 ;
```

Using the OntoLex-FrAC vocabulary, we can now create a link from the OntoLex entry for *DoS ataka* to a `tuv` URI from the corpus as defined above.

```
:DoS+ataka-lt frac:attestation
  [ frac:locus :tuv_157 ].
```

Information Integration

The key advantage of our approach is that it enables the conjoint querying of the three different types of resources (parallel corpus, annotated corpus, terminological/lexical data) in a unified fashion.

If all data is merged into a single RDF graph, the following SPARQL query pattern retrieves attestations of Lithuanian *DoS ataka* from a corpus:

```
?lith_entry
  ontolex:canonicalForm/writtenRep "DoS ataka"@lt;
  frac:attestation/frac:locus ?lith_tuv.
?lith_tuv oa:hasBody/nif:isString ?lith_example.
```

With lexical data and the parallel corpus, we can now retrieve Lithuanian sentences that translate English sentences that contain the English translation of Lithuanian *DoS ataka*. That is, English resources are acting as a pivot between Lithuanian terms and Lithuanian texts. This can be helpful for a morphologically rich language such as Lithuanian where term identification routines are harder to develop than for English.

```
?lith_entry
  ontolex:canonicalForm/writtenRep "DoS ataka"@lt;
  ontolex:sense/ontolex:reference ?term.
?en_entry ontolex:sense/ontolex:reference ?term;
  frac:attestation/frac:locus ?en_tuv.
?tu oa:hasTarget/oa:hasBody ?en_tuv ;
    a :TranslationSet ;
    oa:hasTarget/oa:hasBody ?lith_tuv.
?lith_tuv oa:hasBody/nif:isString ?lith_example.
  FILTER(lang_matches(?lith_example,"lt"))
```

In a similar way, we can access annotations, e.g., to identify the exact word that matches our search term.