CURLICAT

Curated Multilingual Language Resources for CEF.AT

Agreement number: INEA/CEF/ICT/A2019/1926831

Action No: 2019-EU-IA-0034

CEF programme

Deliverable 4

Terminology Enrichment

Curated Multilingual Language resources for CEF.AT

Version 1.0

2022-11-30

Document Information

Activity:	Activity 4: Terminology Enrichment
Deliverable number:	D4
Deliverable title:	Terminology Annotated in the Corpora
Indicative submission date:	2022-11-30
Actual submission date of deliverable:	2022-11-30
Main Author(s):	Tamás Váradi, Svetla Koeva, Radovan Garabík, Andraž Repar, Bartłomiej Nitoń, Vanja Štefanec, Elena Irimia
Participants:	Bence Nyéki, László János Laki, Zijian Győző Yang, Simon Krek, Maciej Ogrodniczuk, Piotr Pęzik, Marko Tadić, Radu Ion, Vasile Păis
Version:	V1.0

History of versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description / Approval Level
V0.1	2022-30-11	Completed	IBL, MTANYTI, JULS SAV, JSI, RAKAI, UZ, ICS	Tamás Váradi, Svetla Koeva, Radovan Garabík, Andraž Repar, Bartłomiej Nitoń, Vanja Štefanec, Elena Irimia, Bence Nyéki, László János Laki, Zijian Győző Yang, Simon Krek, Maciej Ogrodniczuk, Piotr Pęzik, Marko Tadić, Radu Ion, Vasile Păis	

EXECUTIVE SUMMARY

The Activity 4 of the project ***Curated Multilingual Language resources for CEF.AT*** (CURLICAT) aims at enriching the documents in the seven monolingual corpora (for Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) with (1) IATE terms and (2) domain-specific terminology. It also concentrates on the identification of words and multi-word expressions that meet the requirements for domain-specific terms.

The result of this activity is the compilation of the set of seven monolingual corpora annotated with IATE and domain-specific single- and multi-word terms. Adding the annotation of relevant tokens (words) with IATE term identifiers, the corpora are enriched with additional information, i.e., how IATE terms appear in sentences from representative domains and how texts from different languages can be connected through equivalent IATE terms. The seven corpora are also enriched with additional information by annotating relevant sentences with domain-specific terms, i.e., the distribution of domain-specific terms in representative domains. In addition, the annotated sentences provide contextual information for the terminology.

In addition, a terminology dataset containing domain-specific terms for the seven languages is delivered.

1. Introduction **5**

2. Data format **7**

3. Enriching the Bulgarian corpus with terms **8**
3.1. Enriching the Bulgarian corpus with IATE terms 8
3.2. Domain-specific terminology enrichment for Bulgarian 8

4. Enriching the Croatian corpus with terms **11**
4.1. Enriching the Croatian corpus with IATE terms 11
4.2. Domain-specific terminology enrichment for Croatian 11

5. Enriching the Hungarian corpus with terms **13**
5.1. Enriching the Hungarian corpus with IATE terms 13
5.2. Domain-specific terminology enrichment for Hungarian 13

6. Enriching the Polish corpus with terms **16**
6.1. Enriching the Polish corpus with IATE terms 16
6.2. Domain-specific terminology enrichment for Polish 16

7. Enriching the Romanian corpus with terms **18**
7.1. Enriching the Romanian corpus with IATE terms 18
7.2. Domain-specific terminology enrichment for Romanian 18

8. Enriching the Slovak corpus with terms **20**
8.1. Enriching the Slovak corpus with IATE terms 20
8.2. Domain-specific terminology enrichment for Slovak 20

9. Enriching the Slovenian corpus with terms **22**
9.1. Enriching the Slovenian corpus with IATE terms 22
9.2. Domain-specific terminology enrichment for Slovenian 22

10. Activity 4 at a glance **24**

11. Summary of results and conclusion **24**

12. Bibliographical references **26**

1. Introduction

The Activity 4 of the project ***Curated Multilingual Language resources for CEF.AT*** includes two tasks: 1. Enrichment of the monolingual corpora with IATE terms, and 2. Domain-specific terminology enrichment.

Task 1: Enrichment of the monolingual corpora with IATE terms

The goal of this task is to enrich monolingual corpora with terms from IATE (Inter-Active Terminology for Europe – the EU's interinstitutional terminology database)[1]. IATE is a highly representative and widely used resource that includes terminology databases that were built and maintained independently by the translation services of the various EU institutions.

The following filtering criteria are available for IATE, which may be obtained in TBX (TermBase eXchange, ISO 30042:2019) or CSV (Comma-separated values) format: language code (one or more EU languages), domain (one or more domains with their subdomains), collections (by institution, keyword, and creation date), term type (term, abbreviation, short form, phrase, formula, lookup), evaluation (preferred, admitted, obsolete, deprecated, proposed), reliability (marked with 1 to 4 stars depending on the degree of reliability).

IATE terms variants are related within one and the same entry, for example, *special representative* has the following equivalents: in Bulgarian – *специален представител*, *специален представител на ЕС*; in Croatian – *posebni predstavnik, posebni predstavnik EU-a*; in Hungarian – *az EU különleges képviselője, különleges képviselő*, in Polish – *specjalny przedstawiciel Unii Europejskiej, SPUE, specjalny przedstawiciel, specjalny przedstawiciel UE*, in Romanian – *reprezentant special, reprezentant special al Uniunii Europene, RSUE*; in Slovak – *osobitný zástupca, osobitný zástupca EÚ, OZEÚ*; in Slovenian – *posebni predstavnik, posebni predstavnik EU, PPEU*.

IATE concepts are associated with thematic domains, namely with the 21 subject categories based on EuroVoc[2], the multidisciplinary thesaurus that covers EU activities. The EuroVoc domains are: 04 – Politics, 08 – International relations, 10 – European union, 12 – Law, 16 – Economics, 20 – Trade, 24 – Finance, 28 – Social questions, 32 – Education and communication, 36 – Science, 40 – Business and competition, 44 – Employment and working condition, 48 – Transport, 52 – Environment, 56 – Agriculture, forestry and fisheries, 60 – Agri-foodstuffs, 64 – Production, technology and research, 66 – Energy, 68 – Industry,, 72 – Geography,, 76 – International organization. The IATE takes a concept-based approach and has a separate entry for every concept covered. Moreover, terms with broader sense are related with those expressing more narrow sense, for example, the term *secondary education* is broader than the terms *upper secondary education* or *lower secondary education..* The term *secondary education* is related to the EuroVoc term *level of education* and through it to the EuroVoc domain Education and communication.

Different methods for the enrichment of the corpora with IATE terms were used in different languages, based on the adjustment and implementation of previously available

1 https://termcoord.eu/iate/the-new-iate/

2https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

technologies and tools. An important prerequisite for successful annotation with IATE terms is the use of part-of-speech tagging and lemmatization. Single- and multi-word term detection and annotation rely on techniques already available for different languages.
We use existing IATE terms to annotate relevant sentences that contain those terms. By doing so, we enrich the corpus with additional information, i.e., how IATE terms appear in sentences from representative domains and how texts from different languages are related through equivalent IATE terms. In addition, the annotated sentences provide contextual information for the terminology. Since IATE multi-word terms are annotated as single entities, the language modeling for domain-specific texts can be more precise and therefore works more effectively.
The enrichment of seven corpora with IATE terms allows the implementation of different methods for text classification according to the EuroVoc thematic domains and can be used e.g. in multilingual clustering of documents.

Task 2: Domain-specific terminology enrichment

This task aims for high quality term-recognition and term-annotation serving the facilitation of automatic translation.
The term-recognition and term-annotation is performed via different automatic text analysis methods in order to identify words and multi-word expressions fulfilling the criteria for terminological units. The partners used different language-specific tools for part of speech tagging, lemmatization and term recognition to provide the correct annotation of domain-specific terms (combining pattern matching and statistical approaches).
The domain-specific terminology is identified and annotated in 15 thematic domains – the domains considered in the CURLICAT project are: Culture, Economy, Education, Health, Nature, Politics, Science, Social issues, Finance, Industry, Trade, Energy, European Union, Religion, Law, General (some partners identified words and multiword expressions with high level of specificity, belonging to the general lexica). Some of the domains coincide with the EuroVoc domains: Economy, Education, Politics, Science, Social issues, Finance, Industry, Trade, European Union, Law. However, term recognition approaches differ from the use of predefined IATE terms, and our aim is to identify terms that are different from them. Moreover, when a term from domain-specific terminology matches an IATE term in a given domain, this match disambiguates the IATE terms in case of ambiguity (one IATE term can belong to multiple domains).
Terminology integration is used at various phases of machine translation training and at various stages of the machine translation process. The subject of domain adaptation, a key concept of customization in machine translation, is directly related to terminology usage. To meet the requirements of constantly growing and changing terminology in everyday domains, machine translation uses resources that contain up-to-date terminology. The CURLICAT project provides corpus enrichment with both IATE and domain-specific terminology, supplemented by a dataset of terminology for seven languages in fifteen domains. This offers options for the integration of terminology into the training data and for checking for term consistency and completeness in the training data.
The result offers monolingual corpora in seven languages annotated with IATE terms and domain-specific single- and multi-word terms. Through the IATE terms IDs equivalent terms in different languages are linked. Through the IATE terms domain IDs and domain-specific terms IDs documents representing equivalent thematic domains are related. In addition, a

terminology dataset, embracing domain-specific terms for the seven languages is developed.

2. Data format

The Curlicat delivers the data in the CoNLL-U Plus format. Each language specific subcorpus observes the same format, which was deliberately modeled after the CoNLL-U format by including four additional columns. The first ten (1 to 10) columns keep their CoNLL-U values, while the last 2 columns are specific to terminology enrichment task, namely:

- 13th column – CURLICAT:IATE: the annotation of a IATE term by its IATE ID, '_' otherwise;
- 14th column -- CURLICAT**:**DOMAINTERM: the annotation of a domain-specific term by its ID, '_' otherwise.

The CURLICAT IATE ID contains IATE identification number and the identification number of the related Eurovoc domain or term separated by a dash; if more than one EuroVoc domains or terms are related, IATE terms are encoded as different concepts, separated by semicolons. For example, the term *secondary education* is related with the ID 915163-08, where 915163 is the IATE ID and 08 is the code of the EuroVoc domain International relations. The term *employee* is related with the following IDs: 1567863-2826 where 1567863 is the IATE ID and 2826 stands for the EuroVoc term *social affairs,* part of the domain Social questions; 1567863-4006 where 4006 stands for for the EuroVoc term *business organization*, part of the domain Business and communication; 1567863-4026 where 4026 stands for the EuroVoc term *accounting,* part of the domain Business and communication; 1567863-44 where 44 stands for the domain Employment and working conditions, and so on. The links to terms belonging to one and the same domain are separated with commas, as we can expect that a term has the same sense in one and the same domain and the links to different domains are separated with a semicolon. For example, the term *employee* (namely, its translation equivalents in the seven languages) is annotated as follows: 1:1567863-2826;2:1567863-4006,4026;3:1567863-44.

The CURLICAT domain-specific ID contains two-digit prefix pointing to the thematic domain and five-digit unique number, for example, 0155555, where 01 is the domain code, and 55555 is the term ID. The Curlicat domain codes are: 01 – Culture, 02 – Economy, 03 – Education, 04 – Health, 05 – Nature, 06 – Politics, 07 – Science, 08 – Social issues, 09 – Finance, 10 – Industry, 11 – Trade, 12 – Energy, 13 – European Union, 14 – Religion, 15 – Law, 16 – General[3].

There might be cases in which a term is used in different domains (with a different or the same sense). The annotations of the polysemous terms used in a domain different from the domains of the term are separated by a comma., for example: 1:0112345,2:0212345.

3 The *General* is a pseudo-domain for unspecific texts; the extracted "terms" are therefore statistically significantly more frequent words or multiword expressions and as such they are not included in the terminology annotation.

Each IATE and domain-specific term within a sentence is numerically marked from 1 to X, where X is the number of terms within a sentence. The separator between the enumeration and the term is the colon character. Multi-word terms are marked by repeating the number of the term within the sentence until the end of the multiterm. For example, 1:0682148 and 1 are the annotation for the term *Външното министерство* (Foreign Ministry), which belongs to the domain of Politics:

```
1 Външното      външен        ADJ  Asdn        Definite=Def|Degree=Pos|Gender=Neut|Number=Sing
        29      amod    _       _       0       B-NP    _       1:0682148
2 министерство министерство NOUN NCNson      Definite=Ind|Gender=Neut|Number=Sing
        27      nsubj   _       _       0       I-NP    _       1
```

The term database is delivered in two formats: TBX (TermBase eXchange) and CSV (where the delimiter is |). For example:

E_ID|L_CODE|T_TERM|T_TYPE|E_DOMAIN
22222|ro|statului portului|Term|02
44444|ro|Monitorul Oficial|Term|01
00270|pl|zawodowy nauczyciel|Term|03
11305|sk|štátny tajomník|Term|02

where E_ID stands for the entry ID, L_CODE – for language code, T_TERM for the term, T_TYPE for term type, E_DOMAIN – for the domain of the entry.
If a term belongs to 2 or more domains, the domains' codes are separated by comma. For example:
33333|ro|ordinul|Term|01,02

3. Enriching the Bulgarian corpus with terms

3.1. Enriching the Bulgarian corpus with IATE terms

For IATE term annotation, a dedicated instrument called TextAnnotator was used (previously implemented in the scope of the MARCELL project) (Koeva et al. 2020) with an updated IATE database (version from 2022-03-31). The TextAnnotator calls dictionaries of terms and finds occurrences of these terms in the documents. Both the documents and the dictionaries are represented in the CoNLL-U format. For example, the Bulgarian translation equivalent of the IATE term *Doctors Without Borders* has the following dictionary entry:

IATE-128861:1236,2841,4016	M	_	_
Лекари	N	лекар	NCMpo
без	R	без	R
граници	N	граница	NCFpo
#	M	#	M

The annotation tool matches sequences of lemmas and part-of-speech tags of dictionary entries and lemmas and part-of-speech tags of tokens. The matching procedure is based on a hash table indexing. The algorithm gives a priority to the longest length classes, which ensures the selection of longest matches. When a match is found, the corresponding tokens

in the document are annotated and the processing continues from the end index of the match.
46,592 IATE terms are used for Bulgarian and only terms with good reliability were selected. From the obtained dataset 23,647 IATE terms were annotated in the documents. The annotation takes into account that several terms can be related with one and the same IATE ID (synonyms) and one term can be related with different IDs (polysemy). There are also IATE terms in Bulgarian which describe concepts specific for other languages. Such terms were excluded from the annotation.

3.2. Domain-specific terminology enrichment for Bulgarian

The term recognition is performed via automatic text analysis methods in order to identify words and multiword expressions fulfilling the criteria for terms. The focus texts (texts from the Bulgarian Curlicat corpus) and the reference texts (texts from literature and news that are supposed not to contain terms) are tagged for part-of-speech and lemmatised. This ensures that each mult-iword term in the focus texts can be matched against the following linguistic filters (N, AN, AAN, NRN, ANRN, NRAN, ANRAN, NN, ANN, NAN, where A is adjective, N noun, R – preposition) and that frequencies can be calculated correctly when terms are used in different word forms. For each sequence of part-of-speech tags in the focus texts matching one of the linguistic filters and for each adjective from the reference corpus the following information was indexed: the number of texts in which they occur and the number of all occurrences. Multi-word term candidates that contain indexed reference adjectives are eliminated.
To compare the number of occurrences of term candidates in the focus texts with the number of their occurrences in the reference corpus TF-IDF and Log Likelihood algorithms are implemented. The threshold for TF-IDF is set to 0.02. We use the union of the results from Tf-IDF and Log Likelihood. To increase the results the Dice algorithm is applied, which identifies terms similar to those already recognised (with a threshold set to 0.85). Altogether 15 078 unique terms were extracted which allows for 4 007 894 annotations. The number of validated and extracted terms according to the domain is presented in Table 3.2.1:

Domain	*Extracted terms*	*% of terms extracted*
Science	4357	21.87
Culture	2887	14.49
Health	2493	12.51
Politics	2094	10.51
Industry	1705	8.56
Finance	1564	7.85
Economy	1533	7.69
EU	922	4.63
Education	864	4.34
Trade	828	4.16
Energy	679	3.41

Table 3.2.1. Number of automatically extracted terms by domain for Bulgarian

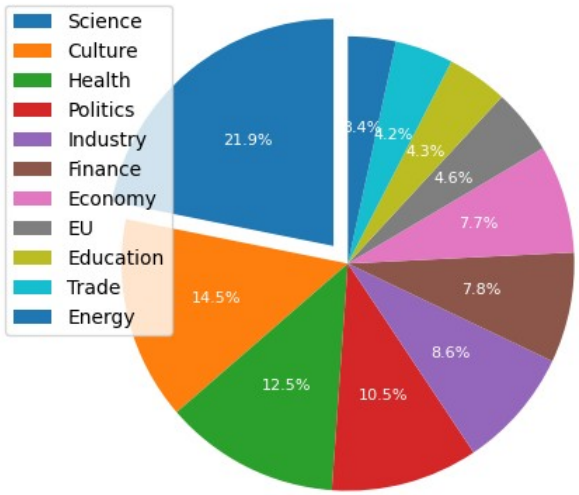

Figure 3.2.1. Number of automatically extracted terms by domain for Bulgarian

4. Enriching the Croatian corpus with terms

4.1. Enriching the Croatian corpus with IATE terms

For the annotation of both IATE and domain-specific terms, the same annotation tool was used, which was originally developed during and for the purpose of MARCELL project. This tool performs pattern-maching against lemmatized sequences of text and is capable of recognizing overlapping terms of different lengths. The tool itself is completely terminology-type-agnostic and relies on the external data provided in a JSON file.
The annotation of IATE terms in Croatian corpus was performed using the export from IATE database from 2022-02-07, comprised of 37,024 terms within 29,119 concept entries. Out of this number of terms in the export, after filtering out those of type "formula" and keeping those of types "fullForm", "shortForm", "phrase" and "abbreviation", the total of 36,553 terms was selected and fed into the annotation tool.
The corpus annotation resulted in 3,206,097 annotated occurences of 9,420 individual IATE terms.

4.2. Domain-specific terminology enrichment for Croatian

The extraction of domain-specific terminology from the corpus was performed using the implementation of the Simple maths keyword extraction method (Kilgarriff 2009), created by the Slovak partner, using the smoothing parameter N=10 and score > 3. The resulting list of extracted terms was used as-is, without any filtering.
The number of extracted terms and their percentual ratio according to the domain is shown in Table 4.2.1.

Domain	*Extracted terms*	*% of terms extracted*
Culture	0	0
Economy	358	29.56
Health	765	63.17
Science	88	7.27

Table 4.2.1. Number of automatically extracted terms by domain for Croatian

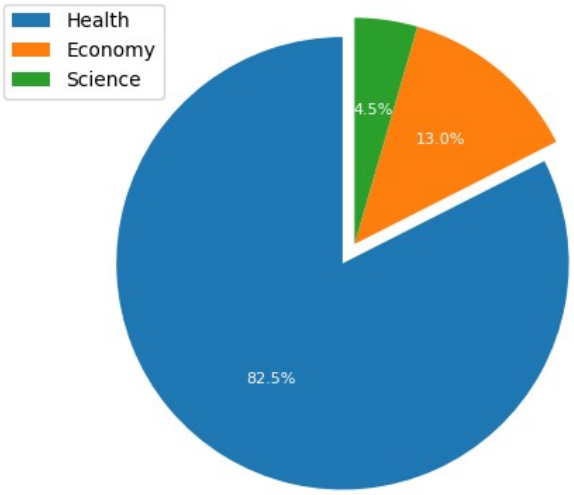

Figure 4.2.1. Number of automatically extracted terms by domain for Croatian

5. Enriching the Hungarian corpus with terms

5.1. Enriching the Hungarian corpus with IATE terms

The Hungarian corpus uses the same annotation method as in MARCELL, with an updated IATE database (version from 21 July 2022). The matching was performed with a string matching algorithm that uses the longest match for multi word terms. As the annotation of IATE terms include overlapping or embedded terms, a special algorithm was developed to extract the longest matching terms. The results are given in column 13 of the Hungarian CURLICAT corpus.

Some basic statistics of the matched terms:

Domain	*One word*	*Multi-word*	*Total*
Culture	1 379 038	7 720	1 386 758
Economy	155 373	7 886	163 259
Science	1 880 961	20 923	1 901 884
Social	2 223 936	58 453	2 282 389
Total	5 639 308	94 982	5 734 290

Table. 5.1.1. Distribution of single and multi-word IATE terms per domains in the Hungarian corpus

5.2. Domain-specific terminology enrichment for Hungarian

We used SketchEngine (Kilgariff et al. 2014), a well-known corpus query and corpus management system for term recognition, and emTerm (Simon et al. 2020) for term annotation. Our workflow consists of the following steps:
1. The Hungarian Curlicat corpus was split into four subcorpora, each standing for a single domain (Culture, Economy, Science and Social issues).
2. These corpora were uploaded one-by-one to OneClickTerms, an interface giving easy access to the advanced terminology extraction functionality of Sketch Engine. We ran OneClickTerms using the default settings, except for two minor changes: the number of results was set to 3000 and the minimum frequency was set to 1 for both single-word and multi-word expressions. Behind the scenes, the following term recognition pipeline was executed:

- Part-of-speech tagging and lemmatization of the domain-specific subcorpus.
- Providing a list of term candidates based on a Hungarian term grammar.
- Determining the relative frequency of each term candidate.
- Comparing the obtained relative frequency of each term candidate to the relative frequency of the same phrase in general language. In our case, this is represented by

the huTenTen corpus (Jakubíček et al. 2013). The comparison method is called Simple maths (Kilgariff 2009), resulting in a keyness score for each candidate.

- Sorting the candidates in descending order according to their keyness score.

As a result, OneClickTerms returned one table for the single-word terms and one for the multiword-ones, for each domain-specific subcorpus.

3. We performed some basic data cleaning on these tables, e.g. removing the names of large publishing houses and deleting e-mail addresses. The frequency of these were salient in our focus corpora compared to the reference corpus, but they clearly shouldn't be regarded as terms.

4. The tables were merged and transformed to the format specified by the Curlicat documentation. Our resulting database is sorted first according to domain IDs and second according to the alphabetical order of terms. If a term appeared in multiple tables (that is, in multiple domains), we kept these occurrences distinct. For example, the term versenyzői egyensúly 'competitive equilibrium' appears as a term in the Economy domain as well as in the Social issues domain. We see this step as a benefit since it aids domain disambiguation.

5. Term annotation was performed by the emTerm module of the e-magyar text processing system (Váradi et al. 2018), using the newly created term database as emTerm's termlist input.

Domain	*Extracted terms*	*% of terms extracted*
Economy	5824	25.24
Social issues	5802	25.14
Science	5737	24.86
Culture	5714	24.76

Table 5.2.1. Number of automatically extracted terms by domain for Hungarian

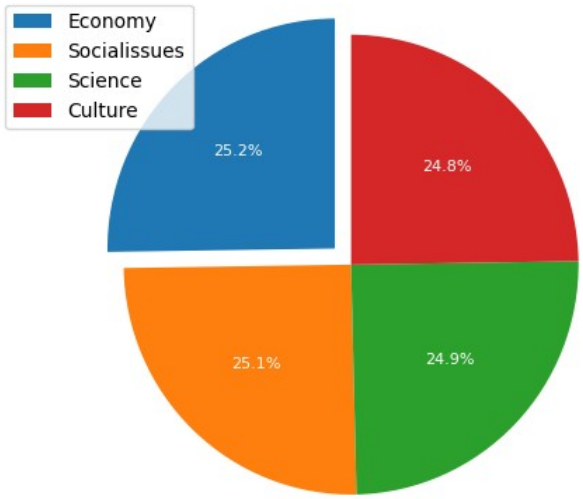

Figure 5.2.1. Number of automatically extracted terms by domain for Hungarian

6. Enriching the Polish corpus with terms

6.1. Enriching the Polish corpus with IATE terms

The Polish corpus uses the same annotation method as in the MARCELL project (Váradi et al. 2020), with an updated IATE database (version from 2021-12-22). Annotation is made using a base string matching algorithm between IATE terms and lemmatized text. Matching is performed on sequences of lemmas using the longest match algorithm for the terms belonging to the same IATE entry.
For the purpose of annotation we used IATE version in TBX format and annotated only terms:

- with type equal to term (fullForm), phrase, short form or formula;
- with reliability code over 6 (3 and 4 stars);
- excluding terms with evaluation (administrativeStatus) proposed or obsolete.

Above conditions result in 82,790 IATE terms in Polish which corresponds to 69,623 individual IATE entries to annotate.

6.2. Domain-specific terminology enrichment for Polish

We used the Polish corpus to extract keywords (single terms and bigrams) from the subcorpora of different CURLICAT domains, using the Simple maths keyword extraction method (Kilgarriff 2009), with empirically selected smoothing parameter N=10 and score > 3. We used the whole Polish corpus as the reference corpus, and a subcorpus specific to the given domain as the focus one. We removed from extracted terms ones containing prepositions, conjunctions, pronouns, particles and words not known to the Morfeusz2 morphological analyser (Woliński 2014; http://morfeusz.sgjp.pl/en).
Similarly to what was performed for IATE terms, annotation is carried out using a base string matching algorithm between terms and lemmatized text. Matching is performed on sequences of lemmata using the longest match algorithm for the terms belonging to the same domain.
The number of extracted terms is shown in Table 6.2.1.

Domain	*Extracted terms*	*% of terms extracted*
Health	838	23.52
Nature	632	27.70
Politics	235	10.30
Education	214	9.38
Science	151	6.62
Economy	111	4.86
Culture	63	2.76
Social issues	38	1.67

Table 6.2.1. Number of automatically extracted terms by domain for Polish

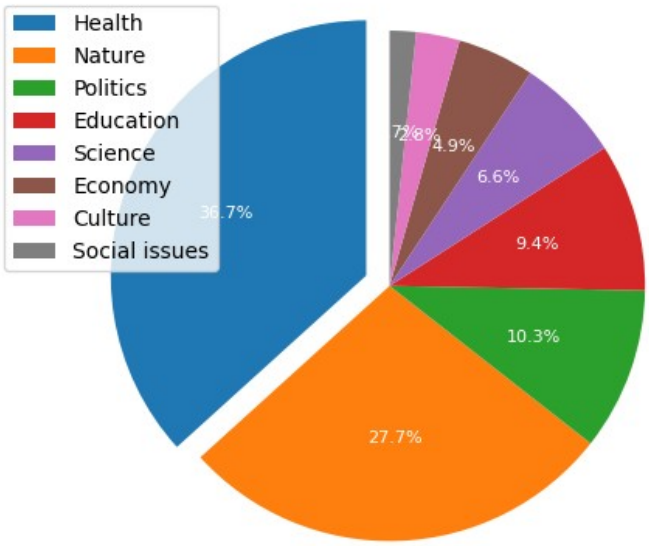

Figure 6.2.1. Number of automatically extracted terms by domain for Polish

7. Enriching the Romanian corpus with terms

7.1. Enriching the Romanian corpus with IATE terms

The Romanian corpus was enriched with IATE terms using an improved version (Avram et al., 2022) of the annotation tool developed in the MARCELL project. It makes use of NER-like matching on words and lemmata. It improves over the previous implementation described in (Coman et al., 2019) and is integrated in our RELATE platform (Păiș et al., 2020). The annotation made use of the latest IATE version available in December 2021 and all the terms were used for the annotation process. Results are given in column 13 ("CURLICAT:IATE") of the CoNLL-U Plus format.

7.2. Domain-specific terminology enrichment for Romanian

Romanian terminology identification was done using a statistical measure of word association, which we chose to be the Dice coefficient[4]. This coefficient was experimentally determined to be better than the Pointwise Mutual Information score[5]. The algorithm works as follows:

1. From the POS tagged and lemmatized Romanian corpus, we extracted all of the NOUN or ADJECTIVE continuous sequences of two words;
2. We have filtered these sequences of words to contain at least one Romanian word, as attested by our inflected word forms lexicon of more than 1M word forms[6];
3. Using the words' lemmas, for a pair (a,b), we counted its frequency as $f(a,b)$, the frequency of the pair ¿[7] and the frequency of the pair $(¿,b)$ and used these values for the Dice score of the pair (a,b):

$$D(a,b)=\frac{2f(a,b)}{f¿¿}$$

4. Retaining all pairs of at least *N* occurrences ($N=2$), we manually evaluated the output of the algorithm for each domain of the Romanian corpus;
5. We have applied the same algorithm for three word sequences (a,b,c) by concatenating the first two members and the last two members and considering them a single lexical unit. Thus, we have applied steps 1 through 4 above for pairs $(a+b,c)$ and $(a,b+c)$ where the "+" operator stands for string concatenation.

All identified terms were sorted descendingly according to their Dice score and only terms with a score bigger than 0.01 were taken into account for manual validation. The number of validated and extracted terms according to the domain is presented in Table 7.2.1:

4 https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient

5 https://en.wikipedia.org/wiki/Pointwise_mutual_information

6 https://github.com/racai-ai/Rodna/blob/master/data/resources/tbl.wordform.ro

7 "*a*" followed by any other word.

Domain	*Extracted terms*	*% of terms extracted*
Health	1446	22.1%
Science	1385	21.2%
Nature	1185	18.1%
Politics	1156	17.7%
Economy	631	9.7%
Culture	366	5.6%
Education	360	5.5%

Table 7.2.1. Number of automatically extracted terms by domain for Romanian

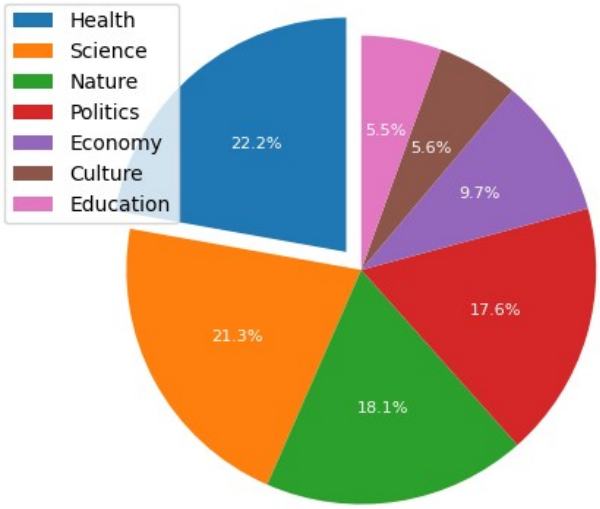

Figure 7.2.1. Number of validated automatically extracted terms by domain for Romanian

After manually validating the automatically extracted domain-specific terms, the actual annotation of the corpus was realized using the same tool (Avram et al., 2022) used for IATE annotation. Final results are given in column 14 ("CURLICAT:DOMAINTERM") of the CoNLL-U Plus format, using annotations similar to the IATE annotations.

8. Enriching the Slovak corpus with terms

8.1. Enriching the Slovak corpus with IATE terms

Slovak corpus uses the same annotation method as in MARCELL, with an updated IATE database (version from 21 July 2022, 64947 individual IATE entries). We included all the Slovak IATE terms, but without the *multiple languages* and *Latin* entries (such entries are generally used translingually in several domains). A simple filtering has been applied to the IATE terms, based on surface features of the terms according to manual assessment of the accuracy of the annotation (Garabík & Levická 2022). We used sequence matching of multiword IATE terms on lemmatized entries, using the longest match for ambiguous annotation. As the sequence matching is performed on sequences of lemmas, a new, improved lemmatization has been implemented (morphology database has been extended to 114 thousand lemmas) and we investigated improvements by replacing the MorphoDiTa lemmatizer by spaCy using multilingual BERT (Garabík & Mitana 2022).

8.2. Domain-specific terminology enrichment for Slovak

We used the Slovak corpus to extract keywords (single terms and bigrams) from the subcorpora of different CURICAT domains, using the *Simple maths* keyword extraction method (Kilgarriff 2009), with empirically selected smoothing parameter N=10 and *score* > 3. We used the whole Slovak corpus as the *reference* corpus, and a subcorpus specific to the given domain as the *focus* one.
The number of extracted terms and their percentual ratio according to the domain is shown in Table 8.2.1; the *General* domain predictably did not yield terminological units, but common words and fixed collocations instead. Therefore we excluded the extracted terms from the *General* domain subcorpus from terminology enrichment markup.

Domain	*Extracted terms*	*% of terms extracted*
Health	3927	32.37
Law	2384	19.65
Education	1510	12.45
Religion	1264	10.42
General †	1228	10.12
Nature	1052	8.67
Politics	497	4.10
Economy	239	1.97
Science	30	0.25

Table 8.2.1. Number of automatically extracted terms by domain for Slovak
† the terms extracted in the *General* domain are not used in the terminology annotation

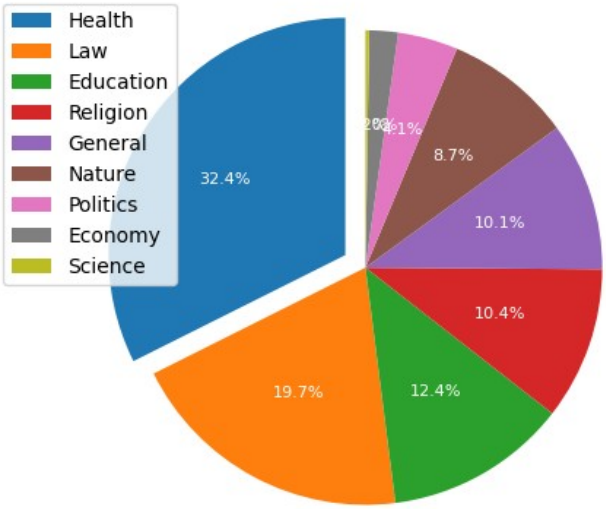

Figure 8.2.1. Number of automatically extracted terms by domain for Slovak

9. Enriching the Slovenian corpus with terms

9.1. Enriching the Slovenian corpus with IATE terms

The Slovenian corpus was enriched with IATE terms using the annotation tool developed in the MARCELL project. We used sequence matching of multiword IATE terms on lemmatized entries, using the longest match for ambiguous annotation. The annotation accuracy was further improved due to the use of the latest version of the Classla[8] morphosyntactic parser for Slovene. The annotation made use of the latest IATE version available in August 2022 and all the terms (79,683 entries) were used for the annotation process. Results are given in column 13 ("CURLICAT:IATE") of the CoNLL-U Plus format.

9.2. Domain-specific terminology enrichment for Slovenian

We used the Slovenian corpus to extract keywords (single terms and bigrams) from the subcorpora of different CURICAT domains, using the *Simple maths* keyword extraction method (Kilgarriff 2009), with empirically selected smoothing parameter N=10 and *score* > 2. We used the whole Slovenian corpus as the *reference* corpus, and a subcorpus specific to the given domain as the *focus* one. Additionally, we filtered out all bi-grams that start with a word that is shorter than 3 characters to remove wrongly identified candidates starting with auxiliary verbs.
The number of extracted terms and their percentual ratio according to the domain is shown in Table 9.2.1.

Domain	*Extracted terms*	*% of terms extracted*
Health	1448	23.52
Finance	1251	20.32
Culture	1029	16.72
Economy	995	16.16
Politics	738	11.99
Education	695	11.29

Table 9.2.1. Number of automatically extracted terms by domain for Slovenian

8 https://github.com/clarinsi/classla

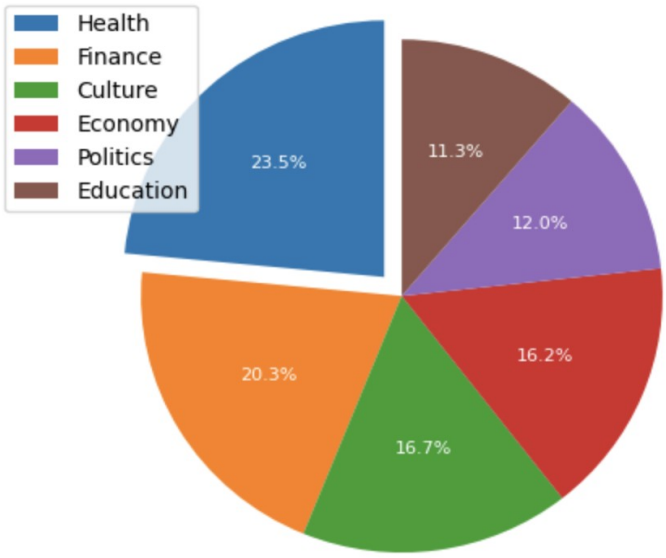

Figure 9.2.1. Number of automatically extracted terms by domain for Slovenian

10. Activity 4 at a glance

Activity 4 produced seven monolingual corpora annotated with IATE and domain-specific single and multiword terms.

The corpora are enriched with additional information by annotating relevant sentences with IATE and domain-specific terms, indicating the distribution of terms in representative domains. On the other hand, the annotated sentences provide contextual information for the terminology.

The table below provides statistical information for the number of sentences, number of words, number of annotated IATE terms and number of annotated domain-specific terms within the seven corpora.

	Number of sentences	*Number of words*	*Number of annotated IATE terms*	*Number of annotated domain-specific terms*
Bulgarian	2 158 765	35 319 695	8 731 145	4 269 020
Croatian	2 123 658	42 252 218	3 206 097	1 073 895
Hungarian	2 756 706	61 196 946	5 734 290	5 367 286
Polish	2 421 154	59 301 782	20 283 841	4 081 176
Romanian	3 557 812	94 925 454	18 807 026	802 254
Slovak	4 805 676	66 891 145	5 744 473	7 404 508
Slovenian	2 003 626	43 481 563	1 217 897	6 575 649
Total	**19 827 397**	**403 368 803**	**63 724 769**	**29 573 788**

Table 10. Information for terminology enrichment of monolingual corpora

The terms database for all languages is available for download from the CURLICAT website (in both CSV and TBX format): https://curlicat-project.eu/deliverables.html

11. Summary of results and conclusion

The translation of specialized knowledge and information is linked to the development of terminological resources since the terminology primarily seeks to achieve consistency, clarity, and understandable and user-friendly content. The annotation of IATE terms (provided and approved by experts) may help to ensure the accuracy and consistency of translation, as well as contribute to effective knowledge transfer.

Still, terminology resources for less resourced languages are not freely available, or if there are some of them, they are usually limited in quantity and of unknown origin. Providing an

annotation with domain-specific terminology might also contribute to the accuracy and consistency of the translation.

To contribute in solving challenges in terminology interpretation we enrich the documents in the seven monolingual corpora (for Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian) with (1) IATE terms and (2) domain-specific terminology.

12. Bibliographical references

Avram, A.M., Paiș, V., Tufiș, D. (2022). Terminology Annotation Using a String Matching Algorithm. In: Proceedings of the 17th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing.

Coman, A., Mitrofan, M., Tufiș, D. (2019). Automatic Identification and Classification of Legal Terms in Romanian Law Texts. In: International Conference on Linguistic Resources and Tools for Natural Language Processing.

Garabík R., Mitana D. (2022). Accuracy of Slovak Language Lemmatization and MSD Tagging – MorphoDiTa and SpaCy. In: LLOD Approaches for Language Data Research And Management, Abstract Book, Mykolo Romerio universitetas, Vilnius, pp. 75–78.

Garabík, R., Levická, J. (2022). Naïve Terminological Annotation of Legal Texts in Slovak - Can It Be Useful? In: Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 48(1), pp. 27-44.

Kilgarriff A. (2009). Simple maths for keywords. In: Proceedings of Corpus Linguistics Conference CL2009, Mahlberg, M., González-Díaz, V. & Smith, C. (eds.), University of Liverpool, UK.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In Lexicography 1. pp. 7–36.

Koeva, Sv., N. Obreshkov, M. Yalamov. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis (eds.): Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 6988-6994.

Păiș, V., Tufiș, D., Ion, R. (2020). A Processing Platform Relating Data and Tools for Romanian Language. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.) Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020), pp. 81–88, Marseille, France, European Language Resource Association (ELRA).

Simon, E., Indig, B., Kalivoda, Á., Mittelholcz, I., Sass, B., Vadász, N. (2020). Újabb fejlemények az e-magyar háza táján. In MSZNY 2020, XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 29–42.

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B. (2018) E-magyar – A Digital Language Processing System. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 1307–1312.

Váradi T., Koeva K., Yamalov M., Tadić M., Sass B., Nitoń B., Ogrodniczuk M., Pęzik P., Barbu Mititelu V., Ion R., Irimia E., Mitrofan M., Păiș V., Tufiș D., Garabík R., Krek K., Repar A., Rihtar M., Brank J. (2020). The MARCELL legislative corpus. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.) Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020), pp. 3761–3768, Marseille, France. European Language Resources Association (ELRA).

Woliński M. (2014). Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2014), pp. 1106–1111, Reykjavík, Iceland. European Language Resource Association (ELRA).

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. (2013). The TenTen corpus family. In 7th International Corpus Linguistics Conference. pp. 125–127.