CURLICAT

Curated Multilingual Language Resources for CEF.AT

Колекция от многоезикови ресурси за CEF.AT

Светла Коева
Институт за български език, Българска академия на науките
24 ноември 2022
Трети национален семинар за споделяне на езикови ресурси

Curated multilingual resources for CEF.AT

Колекция от многоезикови ресурси за CEF.AT

CURLICAT

Колекция от многоезикови ресурси за CEF.AT

- **Цел**

  Да се съберат подбрани едноезикови данни на седем езика (**български, хърватски, унгарски, полски, румънски, словашки** и **словенски**) в предварително дефинирани тематични области, които са от значение за Европейските инфраструктури за цифрови услуги (DSI), наричани още градивни елементи, например електронно издаване на документи (eDelivery), електронна идентификация (eID), електронен подпис (eSignature) и **електронен превод** (**eTranslation**).

CURLICAT

Колекция от многоезикови ресурси за CEF.AT

- Първоначалният източник на данни са националните корпуси за **български, хърватски, унгарски, полски, румънски, словашки** и **словенски.**
- Изисквания кум ресурсите:
  - **Големина**: най-малко 2 милиона изречения за всеки език (общо най-малко 14 милиона изречения или 140 милиона думи)
  - **Състав**: балансирани по обем, доколкото е възможно, документи от тематични области като: наука, култура, здравеопазване, икономика и финанси
  - **Лицензи**: със свободни права, така че да могат да се разпространяват посредством платформата ELRC-SHARE

CURLICAT

Колекция от многоезикови ресурси за CEF.AT

- **Предназначение**
- Предоставянето на седем едноезикови колекции от данни с голям размер цели да подпомогне усъвършенстването на платформата за автоматичен превод на Механизма за свързване на Европа. Това ще позволи на потребителите от различни европейски държави да получат достъп до информация на език, който разбират добре.

CURLICAT

Дейност 1. Съставяне на корпусите

- **Задача 1.1. Подбор на данни от националните корпуси**
  - **Тематични области в CURLICAT: Европейски съюз, енергетика, здравеопазване, икономика, индустрия, култура, наука, образование, политика**, право, природа, религия, социални въпроси, **търговия, финанси**
  - **Използвани източници**
    - Достъпни текстови колекции в интернет
    - Ръчен подбор чрез сърфиране в интернет
    - Автоматично обхождане и извличане на документи от подбрани източници в интернет

CURLICAT

Дейност 1. Съставяне на корпусите

- **Българският национален корпус** е създаден в Института за български език
- Състои се от текстове на български и 47 паралелни корпуса с различна големина
- Българската част съдържа 1.2 милиарда думи
- Материалите в Корпуса отразяват състоянието на българския език (предимно в неговата писмена форма) от средата на XX в. (1945 г.) до наши дни
- Предоставя възможност за извличане на специализирани или общи подкорпуси по определени критерии (тематика, автор, година / период на издаване, източник и др.)

Дейност 1. Съставяне на корпусите

ОБЛАСТИ	ИЗРЕЧЕНИЯ	ДУМИ
НАУКА	989443	10278581
ПОЛИТИКА	728230	9735420
КУЛТУРА	542929	6019423
ЕВРОПЕЙСКИ СЪЮЗ	215707	3360901
ИКОНОМИКА	175474	2656557
ОБРАЗОВАНИЕ	131481	2262071
ФИНАНСИ	259371	4710613
ИНДУСТРИЯ	32939	443701
ЗДРАВЕОПАЗВАНЕ	51755	608575
ЕНЕРГИЯ	6892	127110
ТЪРГОВИЯ	4290	71548
ОБЩО	**3,138,511**	**40,274,500**

Таблица 1. Българският корпус преди процедурите за отстраняване на неподходящи данни

Дейност 2. Разширяване на корпусите и изясняване на авторските права

- **Задача 2.1. Идентифициране на небалансирани тематични области**

ОБЛАСТИ		ИЗРЕЧЕНИЯ	ДУМИ
НАУКА		989443	10278581
	ЗДРАВЕ	51755	608575
ПОЛИТИКА		728230	9735420
КУЛТУРА		542929	6019423
ЕВРОПЕЙСКИ СЪЮЗ		215707	3360901
ИКОНОМИКА		175474	2656557
	ИНДУСТРИЯ	32939	443701
	ЕНЕРГИЯ	6892	127110
	ТЪРГОВИЯ	4290	71548
ОБРАЗОВАНИЕ		131481	2262071
ФИНАНСИ		259371	4710613
Общо		**3,138,511**	**40,274,500**

Таблица 2. Българският корпус: относително балансиране

CURLICAT

Дейност 2. Разширяване на корпусите и изясняване на авторските права

- **Задача 2.2. Изясняване на авторските права, подбор на притежатели на данни**

- Признание (за автора) 
- Споделяне на споделеното (запазване на лиценза при разпространение) 
- Некомерсиална употреба 
- *Без производни (без промени) 

CURLICAT

Дейност 2. Разширяване на корпусите и изясняване на авторските права

- **Задача 2.3. Подбор на допълнителни данни**
- Интернет (автоматично обхождане и изтегляне): 7172 документа
- Българският портал за отворена наука: 481 документа
- Договори с притежатели на данни за използване на техните документи с лиценз, който разрешава промяна: 3662 документа

CURLICAT

Дейност 5. Хармонизация на метаданните

- **Задача 5.1: Формулиране на обща (CURLICAT) схема на метаданните**

Задължителни	Незадължителни	Специфични
• *Идентификатор* • *Език* • ***Лиценз*** • *Дата на публикуване* • *Заглавие на документа* • *Тип на документа* • *Източник* • ***Домейн*** • ***Брой_изречения|думи|пунктуация|токъни***	• *Автор* • *Тип на източника* • *Интернет адрес* • *Стил* • ***Поддомейн***	• *За български* • ***Класификация към EuroVoc*** • *Дата на изтегляне* • *Линк към лиценза в източника* • *Брой_параграфи*

CURLICAT

Дейност 5. Хармонизация на метаданните

- **Задача 5.2: Преобразуване на разнообразни схеми с метаданни към общата схема**
  - Източници с много богата структура на метаданните, като Българския национален корпус
  - Източници с плитка структура на метаданните, като някои публични хранилища с отворени данни
  - Източници без структура на метаданните, но с възможност за извличане на стойности на метаданните, например уебсайтове, блогове и др.

CURLICAT

Дейност 5. Хармонизация на метаданните

- **Задача 5.3. Преобразуване на метаданните, асоциирани към документите, в избрания формат**
  - Ограничен брой категории и стойности на метаданните остават непроменени, тъй като съвпадат с приетия формат
  - Някои оригинални категории и стойности на метаданните могат директно да се преобразуват към категориите и стойностите на CURLICAT
  - Някои стойности на метаданните се извличат автоматично от документите

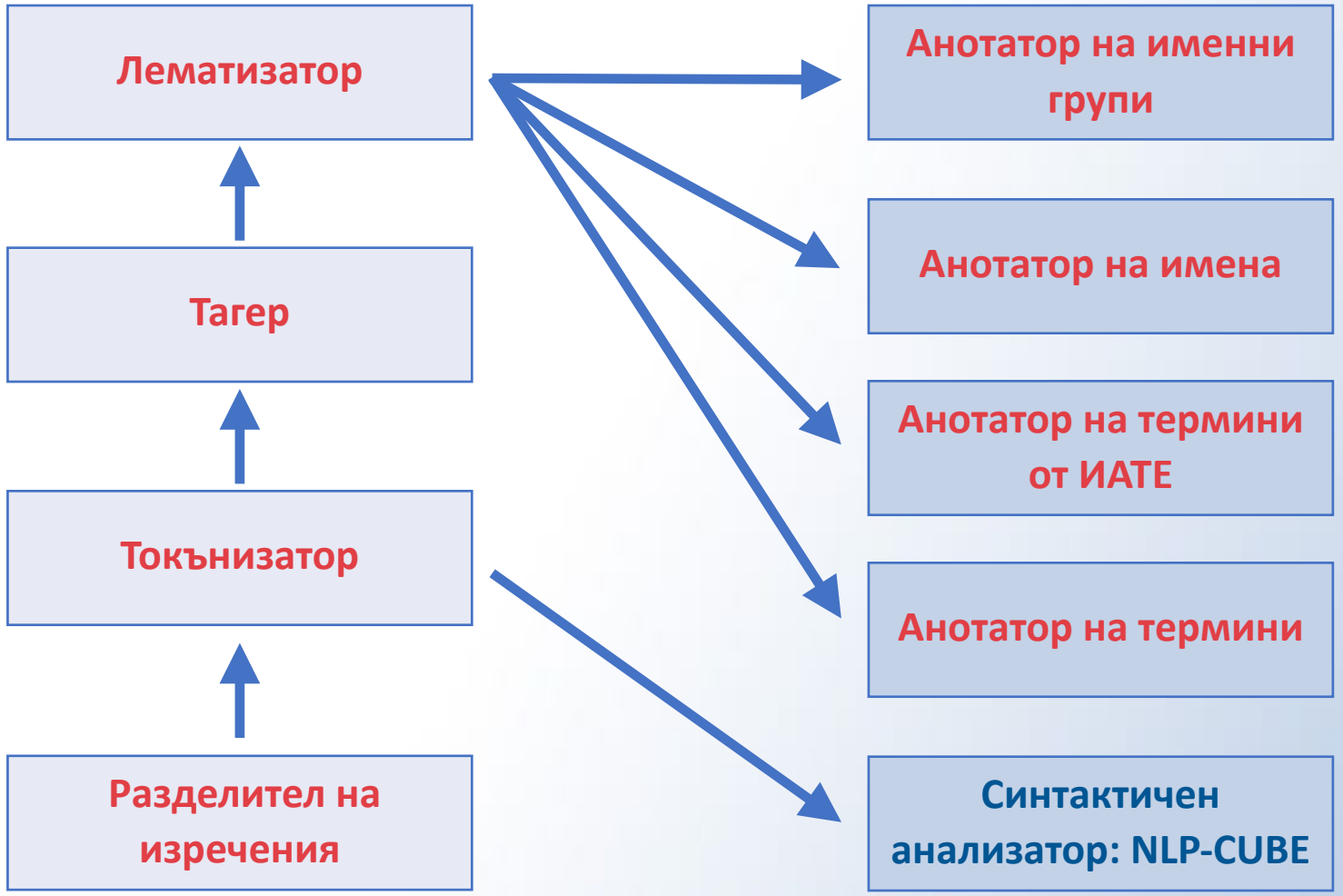
CURLICAT

Дейност 1.2. Анотация

- **Система от програми за обработка на документите за всеки език**
  - Морфосинтактична
    - Токън (поредица от символи)
    - Лема (основна форма)
    - Част на речта
    - Граматични характеристики
  - Синтактични зависимости
  - Имена (на хора, организации и географски обекти)
  - Частичен синтактичен анализ (именни групи)
  - Термини (от IATE: Interactive Terminology for Europe – Интерактивна терминология за Европа, и допълнително извлечени)

CURLICAT

Дейност 1.2. Уеббазирани услуги на Секцията по компютърна лингвистика



CURLICAT

Дейност 1.2. Анотация

- CoNLL-U Plus формат (Computational Natural Language Learning)
  - От 1 до 10 колона са стандартните за CoNLL стпйности

ID FORM LEMMA UDPOS XPOS FEATS HEAD DEPRELDEPS MISC

  - От 11 до 14 колона са стойности, специфични за CURLICAT

11 колона CURLICAT:NE – анотация на имена (на хора, организации, локации) във формат BIO (Beginning-In-Out : Начало-Продължение-Липса)

12 колона CURLICAT:NP – анотация на именни групи (плитък синтактичен анализ) във формат BIO

13 колона CURLICAT:IATE – анотация на термините от ИАТЕ с пореден номер в рамките на изречение ('_' в противен случай)

14 колона CURLICAT:DOMAINTERM – анотация на допълнителни термини с пореден номер в рамките на изречение ('_' в противен случай)

CURLICAT

Дейност 1.2. Анотация

ID	FORM	LEMMA	UDPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
# sent_id = bg-2021111307860-p2s21									
# text = Предметът на разработката се фокусира върху специфичния инструментариум за нейното изследване и оценяване.									
1	Предметът	предмет	NOUN	NCMslm	Definite=Def|Gender=Masc|Number=Sing	5	nsubj	_	_
2	на	на	ADP	R	_	3	case	_	_
3	разработката	разработка	NOUN	NCFsdf	Definite=Def|Gender=Fem|Number=Sing	1	nmod	_	_
4	се	се	PART	T	Case=Acc|PronType=Prs|Reflex=Yes	5	expl	_	_
5	фокусира	фокусирам	VERB	VLITse2	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
6	върху	върху	ADP	R	_	8	case	_	_
7	специфичния	специфичен	ADJ	Ashm	Definite=Def|Degree=Pos|Gender=Masc|Number=Sing	8	amod	_	_
8	инструментариум	инструментариум.	NOUN	NCMsom	Definite=Ind|Gender=Masc|Number=Sing	5	iobj	_	_
9	за	за	ADP	R	_	11	case	_	_
10	нейното	мой	PRON	PPYsdne3f	Definite=Def|Gender=Neut|Number=Sing|Person=3|Poss=Yes|PronType=Prs	11	det	_	_
11	изследване	изследване	NOUN	NCNson.	Definite=Ind|Gender=Neut|Number=Sing	8	nmod	_	_
12	и	и	CCONJ	CC	_	6	cc	_	_
13	оценяване	оценяване	NOUN	NCNson	_	_	_	_	_
14	.	.	PUNCT	U	_	_	_	_	_

CURLICAT

Дейност 3. Анонимизация

- **Задача 3.1. Идентифициране и изясняване на изискванията за анонимизация**
- **Задача 3.2. Разработване на решения за анонимизация**
  - Програмата МАПА (Многоезикова анотация за публичната администрация) за анонимизация на имена и други чувствителни данни (https://mapa-project.eu/)
  - Допълване на резултатите от МАПА с идентифициране на допълнителни имена на български и заместване с произволни имена
  - Допълване на резултатите от МАПА с идентифициране на дати, телефонни номера и други идентификатори със специфично изписване на български
- **Т3.3. Анонимизация на събраните данни**

CURLICAT

Дейност 3. Анотация

ID	FORM	LEMMA	UDPOS	CURLICAT:NE	CURLICAT:NP	CURLICAT:IATE	CURLICAT:TERM
# sent_id = bg-06455-p31s2							
# text = Министерският съвет изпълнява всяка функция, която не е дадена на друг държавен или местен орган.							
1	Министерският	министерски	ADJ	**B-ORG**	**B-NP**	_	_
2	съвет	съвет	NOUN.	**I-ORG**	**I-NP**	1:878414-100	_

CURLICAT

Дейност 4. Терминологично обогатяване

- **Задача 4.1. Обогатяване на корпусите с термини от IATE (Interactive Terminology for Europe – Интерактивна терминология за Европа))**

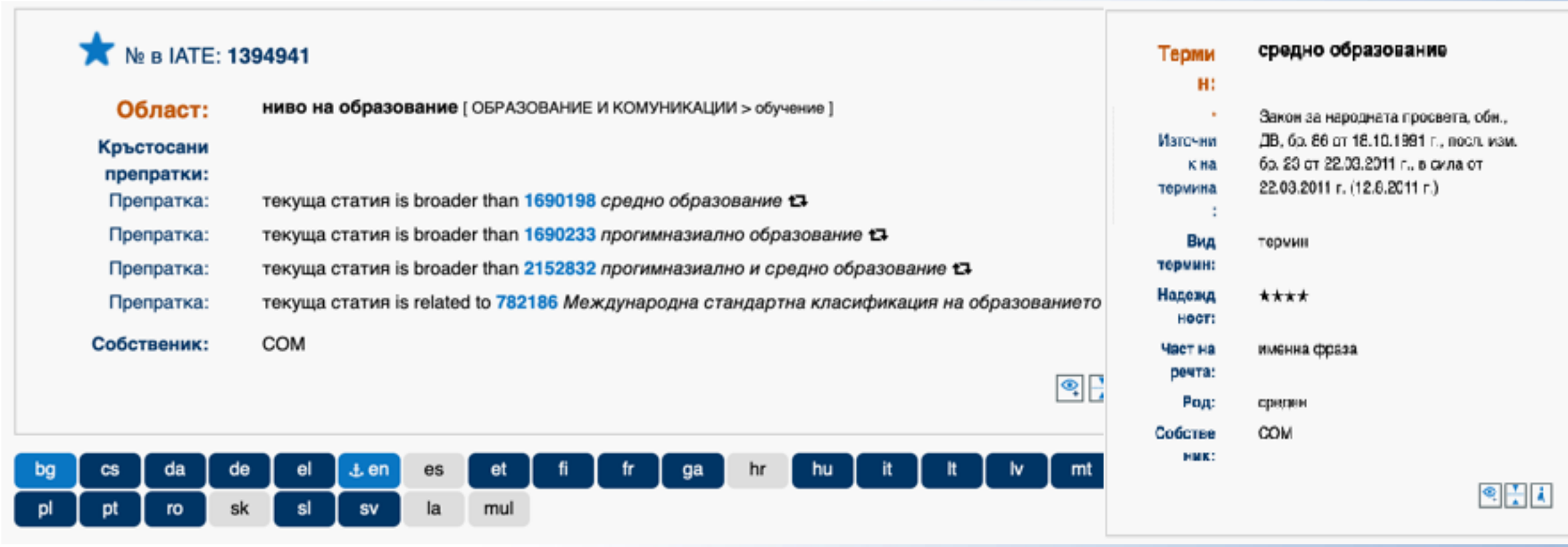
- Над 37 000 уникални термина за български
- Инструмент за анотация, който дава приоритет на най-дългите съвпадения
- Термините са представени в CoNLL формат

<P>	X	<P>	X
<S>	X	<S>	X
IATE-84-1011	M	IATE-84-1011	M
компетентност	N	компетентност	NCFsof
на	R	на	R
държавите	N	държава	NCFpd
членки	N	членка	NCFpo
</S>	X	</S>	X
</P>	X	</P>	X

CURLICAT

Дейност 4. Терминологично обогатяване

- **Задача 4.1. Обогатяване на корпусите с термини от ИАТЕ**

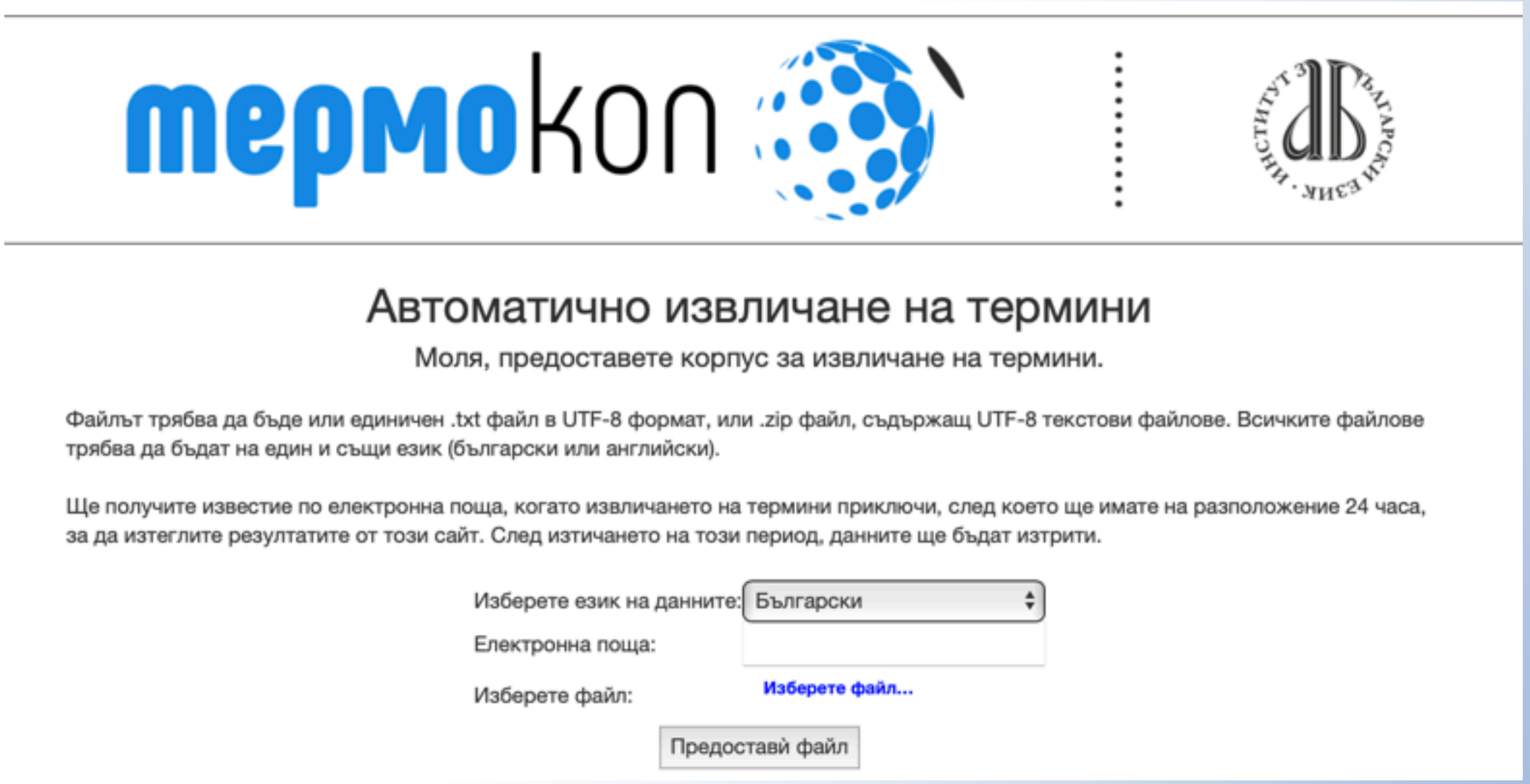


Дейност 4. Анотация

ID FORM	LEMMA	UDPOS	CURLICAT:NE	CURLICAT:NP	CURLICAT:IATE	CURLICAT:TERM
# sent_id = bg-07309-p5s2						
# text = Получава средно образование в Белград, след което завършва психология в Белградския университет.						
1 Получава	получавам	VERB	O	O	_	_
2 средно	среден	ADJ	O	B-NP	**1:1394941-3211**	_
3 образование	образование	NOUN	O	I-NP	1	_
4 в	в	ADP	O	O	_	_
5 Белград	Белград	NOUN	B-LOC	O	**2:925710-72**	_

CURLICAT

Дейност 4. Терминологично обогатяване



CURLICAT

Дейност 4. Анотация

ID FORM	LEMMA	UDPOS	CURLICAT:NE	CURLICAT:NP	CURLICAT:IATE	CURLICAT:TERM

sent_id = bg-2011060313995-p17s2

text = Така например при вирусно заболяване в организма се разпространяват множество вирусни частици, в болшинството от случаите изградени и от протеини.

4 вирусно	вирусен	ADJ	O	B-NP	_	**1:0130217**
5 заболяване	заболяване	NOUN	O	I-NP	_	**1**
6 в	в	ADP	O	O	_	_
7 организма	организъм	NOUN	O	O	1**:1442772-3606**	_

CURLICAT

Резултати

- Седем едноезикови корпуса (за български, полски, румънски, словашки, словенски, унгарски и хърватски) с големина над 2 милиона изречения и над 20 милиона думи, съдържащи текстове със свободни за разпространение и промяна авторски права (общо над 14 милиона изречения и над 140 милиона думи) от подбрани тематични области
- Текстовете са снабдени с богати и хармонизирани между отделните езици метаданни
- Текстовете са обогатени с подробна лингвистична информация: морфологична (част на речта и граматични характеристики), синтактична (синтактични зависимости, именни групи), терминологична (термините от ИАТЕ и от подбрани тематични области), лексикална (имена за хора, организации и локации)
- Седемте корпуса се разпространяват посредством платформата ELRC-SHARE (Координация на езиковите ресурси в Европа СПОДЕЛЯНЕ)

CURLICAT

Консорциум

- Институтът за български език „Проф. Л. Андрейчин“, Българска академия на науките
- Институтът по компютърни науки, Полска академия на науките
- Институтът за изкуствен интелект, Румънска академия на науките
- Институтът „Йожеф Стефан“, Словения
- Институтът по лингвистика „Людовик Щур“, Словашка академия на науките
- **Изследователският институт по лингвистика, Унгарска академия на науките**
- Университетът в Загреб, Хърватия

CURLICAT

Благодаря за вниманието!

Финансиране: Изпълнителна агенция за иновации и мрежи. Механизъм за свързване на Европа. Сектор „Телекомуникации“, INEA/CEF/ICT/A2019/1926831. No: 2019-EU-IA-0034

This action received funding from the Connecting Europe Facility Telecommunications Programme.
Agreement number: INEA/CEF/ICT/A2019/1926831. Action No: 2019-EU-IA-0034