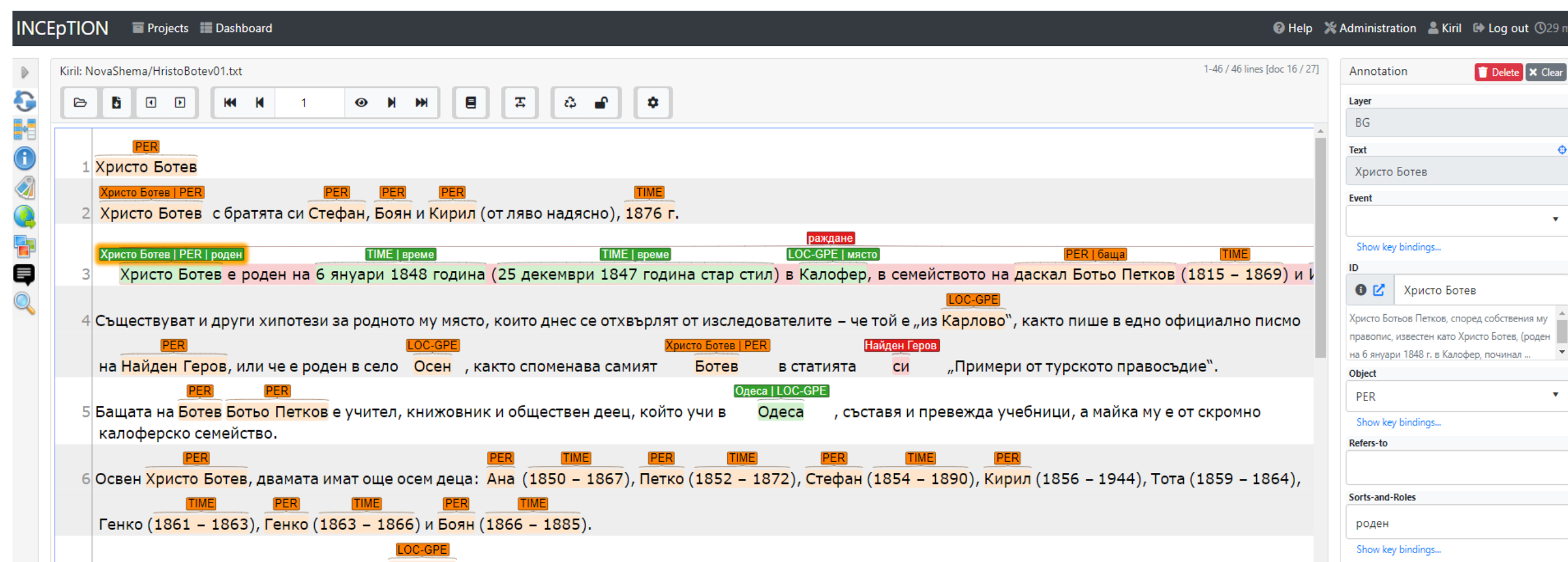


## 1. Overview

- The main reason for the construction of the **BEC corpus** is to have appropriate data for training Named Entity Recognition (NER), Named Entity Linking (NEL) and Event Recognition models
- Such models would help us in the extraction of structured knowledge from domain texts in the area of Social Sciences and Humanities (SSH)
- The extracted structured knowledge will be ultimately used for the creation of a Bulgaria-centric Knowledge Graph
- The corpus comprises a wide variety of domain texts:
  - historical texts from different periods of Bulgarian history;
  - cultural artefacts like church icons;
  - scientific publications;
  - archival documents;
  - encyclopedic articles from Bulgarian Wikipedia.
- In the initial annotation of the corpus we concentrated on:
  - a rich set of Named Entities,
  - some general concepts and events that happen to be frequent

## 2. BEC Specifics and Annotation Schema

- In order to control and predict the structure of the extracted knowledge, the annotation scheme followed the philosophy of **CIDOC-CRM** - ontology which has been widely used in the area of GLAM (Galleries, Libraries, Archives, and Museums) and Humanities
- In addition, we used event descriptions from **FrameNet**, and locally adjusted the scheme to our data
- For the creation of the corpus we relied on the **INCEpTION** annotation tool (<https://inception-project.github.io/>)



- Several Annotation Levels:** Named Entity Annotation - PER, LOC, LOC-GPE, ORG, TIME; Events and Roles – Event definition and participants roles: Event(Role01, Role02, ...). Kappa: between 0.87 and 1.0
- Events** are selected on the basis of ontology CIDOC-CRM and extended with frames from FrameNet: between 0.87 and 0.91
- Named Entities** are annotated with identifiers from Bulgarian DBpedia

## 3. Named Entities Types

Label	Description
DOC	Various texts, including documents, excluding juridical documents – see JUR
EVT	Named events like Second World Wars
JUR	Juridical documents: laws, regulations, etc.
LOC	Locations/places — natural or man-made like mountains, lakes, etc., geopolitical units are excluded – see LOC-GPE
LOC-GPE	Geopolitical units (countries, regions, cities, cantons, etc.)
MSC	Miscellaneous names that not included in the other categories
MSR	Measurements with expressed quantity
ORG	Organizations of any kind
PER	People (existing in reality or fictional ones)
PER-GPE	Nationalities (Bulgarian), the birth place, or the place where people live
PER-GRP	Groups of people that cannot be described as PER-GPE or PER-LOC (Slavs, etc.)
PER-LOC	People that are related to geographical region, but not PER-GPE
PRO	Products — tangible and intangible (DOC and JUR excluded)
REF	Bibliographical references, citations of them, links.
SUM	Amounts of money — a subclass of MSR
TIME	Time points or periods

## 4. Event Types

Event	Roles
Donation	<b>donor</b> (person or organization) <b>recipient</b> (person or organization) <b>theme</b> (object) <b>mediator</b> (person or organization, it could be fund) <b>period-of-iterations</b> (time: the length of time from when the event denoted by the target began to be repeated to when it stopped) <b>goal</b> (situation: the goal for which the donor gives the theme to the recipient) <b>time</b> <b>place</b>
Giving-Birth	<b>brought-into-life</b> (the new born person) <b>parents</b> (the mother and father expressed together, for example “his parents” or “Penka and Toncho Ivanovi”) <b>mother</b> <b>father</b> <b>place</b> (the birth place — usually the name of a city, country or hospital) <b>time</b> (the time of birth — usually it’s a date, but can include hours, or it’s just month and year)
Moving-in-Place	<b>agent</b> (a person) or theme (another type of object) <b>coagent</b> (another person or group of people the agent is moving with) <b>move-from</b> (the place from which the agent or the time moves) <b>move-to</b> (the place where the agent or the theme moves to) <b>time/beginning/end/duration</b> <b>purpose</b> (a situation or another event which causes the moving) <b>goal</b> (a situation/event to be achieved with the moving)

## 5. From Annotation to RDF

Mapping between **AS** to **CIDOC-CRM** defines the conversion to RDF:

E86 Leaving = leaving

E39 Actor (**E21 Person** = **PER** **E74 Group** = **ORG/LOC-GPE**)

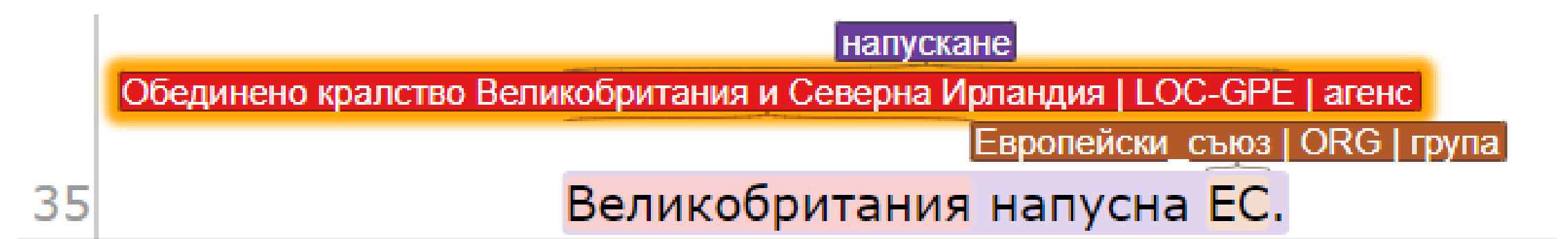
**P145 separated** (left by) = **agent** domain: **E86 Leaving** range: **E39 Actor**

#Text=Великобритания напусна ЕС. **Great Britain left EU.**

35-1 2091-2105 Великобритания **LOC-GPE**[118]

[http://bg.dbpedia.org/resource/Обединено\\_кралство](http://bg.dbpedia.org/resource/Обединено_кралство)[118]

**agent**[118] **leaving**[118]



[http://bg.dbpedia.org/resource/Обединено\\_кралство](http://bg.dbpedia.org/resource/Обединено_кралство) rdf:type

[http://www.cidoc-crm.org/cidoc-crm/E74\\_Group](http://www.cidoc-crm.org/cidoc-crm/E74_Group) .

dbpedia:Обединено\_кралство a cidoc:E74\_Group .

dbpedia:Европейски\_съюз a cidoc:E74\_Group .

dbpedia:Излизане\_на\_ВБ\_от\_ЕС a cidoc: E86\_Leaving .

dbpedia:Излизане\_на\_ВБ\_от\_ЕС cidoc: P145\_separated

dbpedia:Обединено\_кралство .

dbpedia:Излизане\_на\_ВБ\_от\_ЕС cidoc: P146\_separated\_from dbpedia: ЕС .

## 6. Experiments and Results

The models that were initially trained are implementations in **spaCy**, **Flair NLP** and **Hugging Face**

Entity	Examples	NER-TBP			BiLSTM		
		P	R	F1	P	R	F1
TIME	495	80.87	71.51	75.90	70.00	81.01	74.81
LOC	197	68.53	58.95	63.38	71.60	61.42	66.12
LOC-GPE	641	80.99	84.43	82.67	84.32	91.42	87.72
PER	858	79.61	83.05	81.29	85.26	84.97	85.11
ORG	300	64.47	67.77	66.08	67.55	68.00	67.77
JUR	17	34.09	36.59	35.29	53.85	41.18	46.67
EVT	126	68.60	50.43	58.13	76.79	68.25	72.27
PERS	134	82.46	59.49	69.12	85.34	73.88	79.20
DOC	35	29.79	24.56	26.92	23.08	25.71	24.32
PRO	112	24.24	12.50	16.49	36.51	20.54	26.29
SUM	195	50.00	10.34	17.14	65.78	63.08	64.40
All	3114	74.85	71.38	73.08	76.29	76.69	76.49