



The Treatment of Named Entities in the Bulgarian Event Corpus

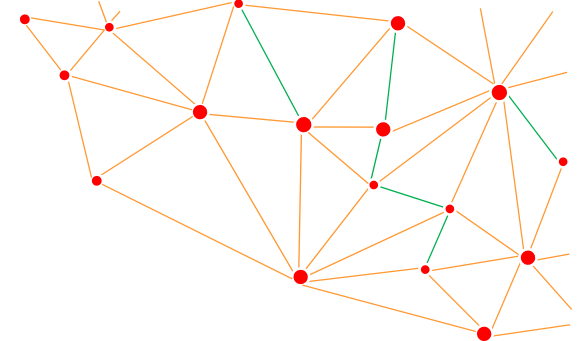
Preslava Georgieva, Petya Osenova, Kiril Simov

Institute of Information and Communication Technologies, BAS

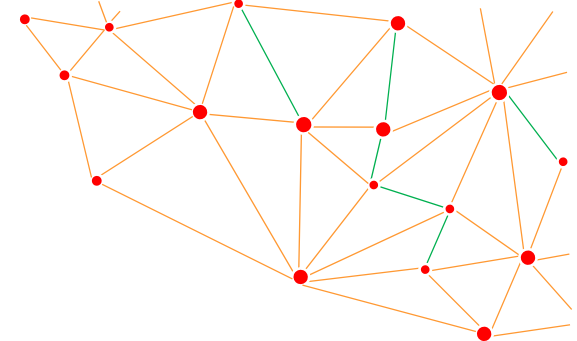


Plan of the Talk

- Bulgarian Event Corpus (BEC): An overview
- Annotation Process within BEC
- Challenges on NEs Annotation Level
- Focus on Appositions – Definition and Types
- Approaches to the Annotation of Appositions
 - Related to a Person (PER)
 - Related to a Non-person (LOC and GPE)
- Conclusions and Future Work



Bulgarian Event Corpus (BEC): An Overview (1)



- Field – Humanities and Social Sciences
- Focus – semantic annotation on **3 levels**:
 1. **Named entities (NEs) and terms**
 2. **Events and roles**
 3. **URL-linking**

}

INCEpTION

}

Link_To_URL
- Data sources:
 1. From the **partners** – research papers, biographical descriptions, archive documents and more
 2. From **Wikipedia** – articles on important geopolitical and administrative entities, organizations and big historical events
- Dataset: over **325** annotated and curated files

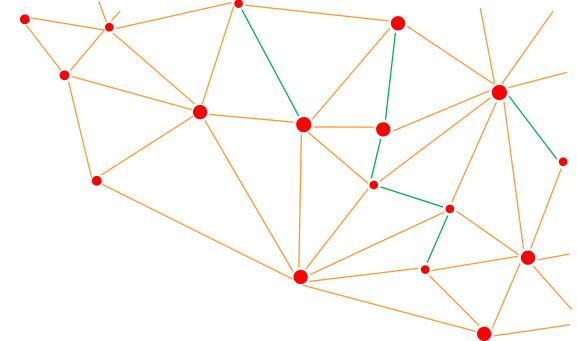
Bulgarian Event Corpus (BEC) (2)

Annotation scheme (AS) – based on **CDOC-CRM ontology** and **FrameNet**

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

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

Annotation Process within BEC (1)



- Three levels of annotation:

1. NEs and terms annotation level (16 types of NEs)

19 През юли 1861 г. Игнатиев пътува за Цариград като делегат на Александър II, носейки поздрав на новия турски султан Абдул-Азис.

In [July 1861]TIME, [Ignatiev]PER traveled to [Tsargrad]LOC-GPE as a [delegate]term of [Alexander II]PER, offering greetings to the new Turkish [Sultan]term [Abdul-Aziz]PER.

2. Events and roles annotation level (38 types of Events)

19 През юли 1861 г. Игнатиев пътува за Цариград като делегат на Александър II, носейки поздрав на новия турски султан Абдул-Азис.

In [[July 1861]TIME]time], [[Ignatiev]PER]agent] traveled to [[Tsargrad]LOC-GPE]move-to] [as a [delegate]term of [Alexander II]PER]purpose], offering [greetings to the new Turkish [Sultan]term [Abdul-Aziz]PER]goal].

Annotation Process within BEC (2)

3. URL-linking annotation level

Link_To_URL - ANTONYUGOV.TSVCEPF

Действия Каталог Помощ

Изречение 7

7 Югов е роден на 5 август 1904 година в Ругуновец (Карасуле), тогава в Османската империя.

8 Сестрин син е на Иванчо Карасулията, войвода на ВМОРО и ВМОК.

9 След Балканските войни семейството му се преселва в Дедеагач, а после се установява в Гюмюрджина, като и двата града тогава са в България.

10 След Първата световна война, когато Гюмюрджина е отнета от България, семейството на Югов се мести в Пловдив.

11 През 1919 година завършва прогимназия, след което работи в тютюневата промишленост.

12 От 1921 г. е член на Българския комунистически младежки съюз, а от 1928 година – на БКП.

13 До 1933 г. участва в дейността на нелегалната БКП в Пловдивско, както и в прокомунистическата Вътрешна македонска революционна организация (обединена), като през 1933 – 1934 година е секретар на нейния Централен комитет с псевдоним Рашко.

14 По това време михайловистите издават смъртна присъда на Антон Югов и той е принуден да ходи с охранител, като това е Иван Козарев, бъдещият първи партизанин на България.

15 През октомври 1934 г. Антон Югов заедно с група комунисти нелегално заминава да учи в Съветския съюз.

Refer to (in document) 0

Reference	Count	Current num	URL
1933 - 1934	0	0	
1941	0	0	
Антон Югов	29	0	
БКП	1	0	

Category ne Refers to Антон Югов

Tag PER

Text Югов

URL Антон_Югов

Югов е роден на 5 август 1904 година в Ругуновец (Карасуле), тогава в Османската империя.

Изречение 7 Документ 6

Token_id	Token	Ne_1	Ev_1	Ro_11
84	Югов	P<	p<	p<
85	е		а	т<
86	роден		ж	р
87	на		д	
88	5	T<	а	в<
89	август	I	н	р
90	1904	M	е	е
91	година	E		м
92	в			
93	Ругуновец	L<	м<	т<
94	(O	е	е
95	Карасуле	C	с	м
96)	-	т	а
97	,		о	
98	тогава		п	в<
99	в		о	
100	Османската	L<	л	м<
101	империя	O	о	я
102	.			

Елемент	Референция	URL	Текст	F_token	L
КОМЕНТАР					
НАИМЕНОВАНИ СЪЩНОСТИ					
PER	Антон Югов	Антон_Югов	Югов	84	8
TIME			5 август 1904 година	88	9
LOC-GPE		Ругуновец	Ругуновец (Карасуле)	93	9
LOC-GPE		Османска_империя	Османската империя	100	1
СЪБИТИЯ					
раждане			Югов е роден на 5 август	84	9
роден	Антон Югов	Антон_Югов	Югов	84	8
тригер			е роден	85	8
време			5 август 1904 година	88	9
място		Ругуновец	Ругуновец (Карасуле)	93	9
местоположение			Ругуновец (Карасуле), то	93	1
тема		Ругуновец	Ругуновец (Карасуле)	93	9
време	1904 г.		тогава	98	9
място		Османска_империя	Османската империя	100	1

Annotation Process within BEC (3)

3. URL-linking annotation level

1. Name

- Петко Славейков
- Петко Рачов Славейков
- Иван Славейков
- Иван Петков Славейков

2. URL

https://bg.wikipedia.org/wiki/Петко_Славейков

Петко Рачов Славейков е български поет, публицист, фолклорист и политик. Той е сред водачите на Либералната партия след Освобождението, пр?

3. Names

- Име
- Петко Славейков
- Петко Рачов Славейков

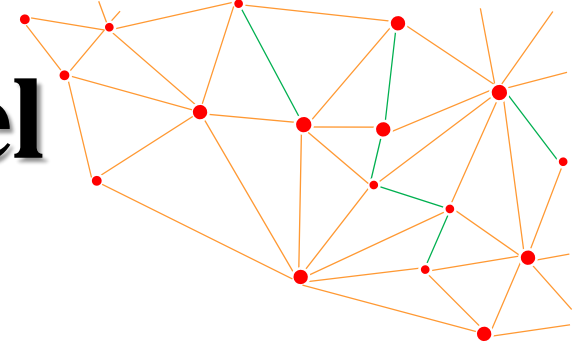
4. Links

- Link
- https://bg.wikipedia.org/wiki/Файл:2010-03-10_Petko_Slaveikov_Bonn_Duis
- https://bg.wikipedia.org/wiki/Иван_Карастоянов
- https://bg.wikipedia.org/wiki/17_ноември
- <https://bg.wikipedia.org/wiki/1827>
- <https://bg.wikipedia.org/wiki/Търново>

5. Properties

Наименование	Стойност
Class:	български поет и политик
Class:	Портрет от 1884 г. на Иван Кар
Роден	17 ноември 1827 г.Търново, Ос
Починал	1 юли 1895 г.София, Княжество
Възраст	(67 г.)
Погребан	Централни софийски гробище
Националност	България
Работил в	поет • публицист • политик
Class:	Политика
Партия	Либерална партия(1879 – 1895)
Class:	МВР
министър	(1880 – 1881; 1884 – 1885)
Class:	МНП
(упр.) министър	(1880)
Class:	НС
председател	(1880)
Class:	Депутат
Class:	Семейство
Съпруга	Ирина Райкова
Деца	Иван Славейков, Христо Славейков

Challenges on NEs Annotation Level



- **Ambiguity:**

1. **One label for two or more different entities.**

- the *same proper noun* is used for *different geopolitical and administrative entities* (exp. Levski **Station**, Levski **Peak**, Levski **Municipality**)



2. **Two or more different labels for the same element.**

- different *natural or administrative entities named after famous historical figures*
Botev Peak --- Hristo **Botev**, **Elin Pelin** Station --- **Elin Pelin** (person)

Focus on Appositions – Definition and Types (1)

- **Apposition** – A noun phrase (NP) that designates another noun and is not connected to it with a preposition (Huddleston, R., Pullum, G. K., 2002; Брезински, Ст., 2000)
- **Classification of appositives:**
 - Complexity:
 - Non-complex appositives

LOC
На връх Ботев са изградени

Botev **Peak**

PER
хофмаршала на княз Фердинанд

Prince Ferdinand

- Complex appositives

термин ORG PER
уважавания и известен архитект от Виенската академия за изящни изкуства Фердинанд Фелнер

the respected and famous architect from the Academy of Fine Arts in Vienna Ferdinand Fellner

Focus on Appositions – Definition and Types (2)

Type of the head word:

- Related to a **person**

PER – *refers to real or fictional people who are represented in the text by their names, pseudonyms, initials*

- Related to a **non-person**

GPE – *refers to geopolitical entities*

LOC – *refers to man-made and natural sites that are not geopolitical entities*

Reasons:

inanimate entities → *could not be determined by many classifiers*

people → *have different roles through their life*

Focus on Appositions – Definition and Types (3)

Stability of the role attributed (for PER)

Permanent – related to the *place of origin* (виенчанин ‘Viennese’), *ethnicity* (славянин ‘Slavic’)

Temporary – related to *nationality* (българин ‘Bulgarian’), *community to which the person belong* (работник ‘worker’), *place they inhabit* (планинец ‘mountaineer’, островинтянин ‘islander’), *profession* (архитект ‘architect’), *status in society and kinship relations* (крал ‘king’, чичо ‘uncle’, свекърва/тъща ‘mother-in-law’), etc.

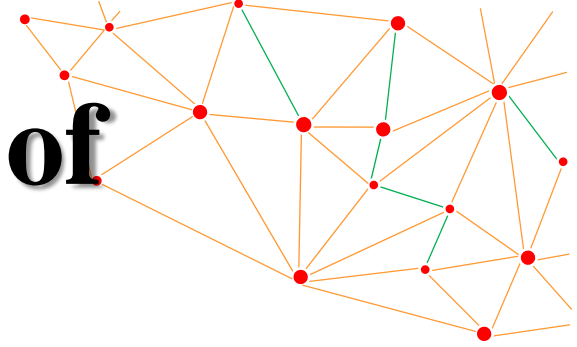
Our Approach to the Annotation of Appositions in the Earlier Version of the AS

- **Maximum segment** – when the noun phrase (NP) and the proper noun are marked
- **Minimum segment** – when only the proper noun of the entity is marked

<i>Maximum segment</i>	<i>Minimum segment</i>
	
<i>[prof. Penchev]_{PER}</i> <i>[the poet Petya Dubarova]_{PER}</i>	<i>prof. [Penchev]_{PER}</i> <i>the poet [Petya Dubarova]_{PER}</i>

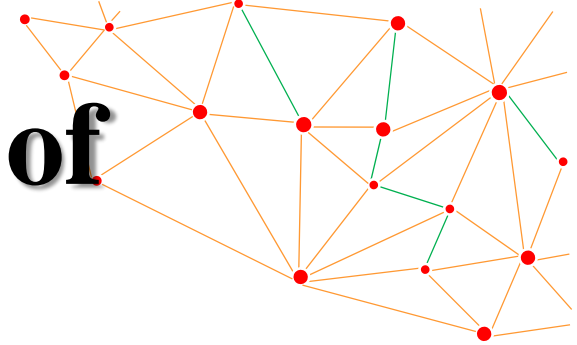
→ *Not very high results regarding annotators agreement*

Other Approaches to the Annotation of Appositions for PER



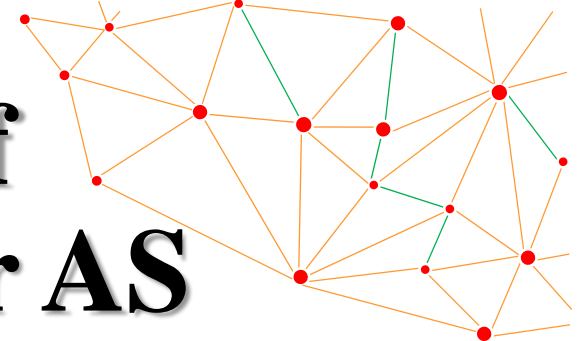
1. The appositive nouns are ***not part of the NE*** and are ***not annotated***
 - **for English** – MUC-7 Named Entity Task Definition, ver. 3.5 (1997); Co-reference Guidelines for English OntoNotes, ver. 7.0 (2007);
 - **for Bulgarian, Czech, Polish and Russian** – The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages;
2. The appositive nouns are ***not part of the NE***, but are ***annotated with different annotation tags***
 - **for English** – ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6. (2008);
 - **for Spanish** (historical texts) – TEI-friendly annotation scheme for medieval named entities: a case on a Spanish medieval corpus;

Other Approaches to the Annotation of Appositions for LOC and GPE



1. The appositive nouns *are part of the NE*
 - **for English** – MUC-7 Named Entity Task Definition, ver. 3.5 (1997); ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6. (2008); ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6. (2008);
 - **for Bulgarian, Czech, Polish and Russian** – The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages.
2. The appositive nouns are *not part of the NE*, but *are annotated with different annotation tags*

Our Approaches to the Annotation of Appositions in the new version of our AS



- For PER: *Not part of the NE*, but *annotated with different annotation tags*

архитект Виктор Румпелмайер

PER-GPE

MSR | 10 000 фенове
10 000 от тях са пловдивчани

PER-LOC

балканджийка, опазила

PER-GRP

различни разбирания – либерали, радикали и революционери.

- For LOC and GPE:

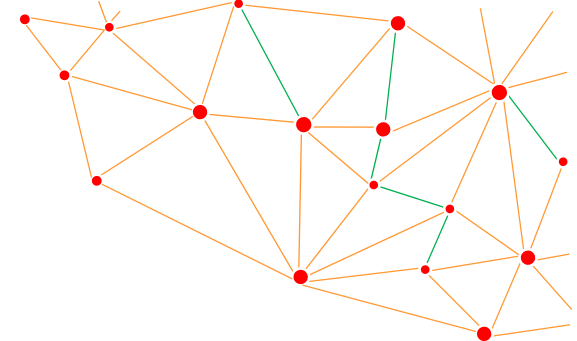
Part of the NE

в района на връх Ботев

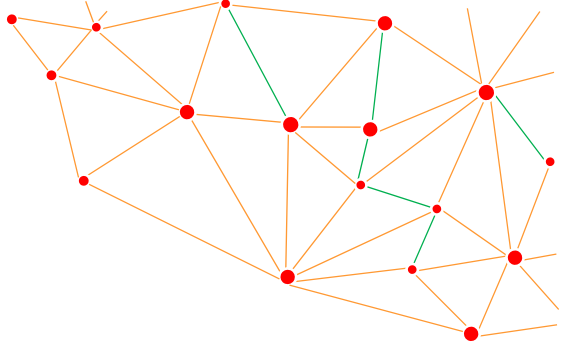

Бежанци на гара Свиленград

В община Левски са изградени

Conclusions and Future Work



- To decide how to treat the appositives defining other **non-person** head words – **ORG, PRO, DOC**, etc.
- To deal with complex appositions
- To make the AS more detailed towards the annotation of appositives related to a **person**
 - *PER-TITLE, PER-KIN*, etc.
- To perform experiments for Named Entity Recognition in order to justify the changes that we have done in our AS



Thank you for your attention!