Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4436–4443
Marseille, 11–16 May 2020

Introducing RONEC - the Romanian Named Entity Corpus

Stefan Daniel Dumitrescu, Andrei-Marius Avram[1]
University Politehnica of Bucharest[1]
Bucharest, Romania
dumitrescu.stefan@gmail.com, avram.andreimarius@gmail.com

Abstract

We present RONEC - the Named Entity Corpus for the Romanian language. The corpus contains over 26000 entities in 5000 annotated sentences, belonging to 16 distinct classes. The sentences have been extracted from a copy-right free newspaper, covering several styles. This corpus represents the first initiative in the Romanian language space specifically targeted for named entity recognition. It is available as BRAT and CoNLL-U Plus (in Multi-Word Expression and IOB formats) text downloads, and it is free to use and extend at github.com/dumitrescustefan/ronec .

Keywords: Named Entity Corpus, NER, Romanian, CoNLL-U Plus format, BRAT, open-source

1. Introduction

Language resources are an essential component in entire R&D domains. From the humble but vast repositories of monolingual texts that are used by the newest language modeling approaches like BERT[1] and GPT[2], to parallel corpora that allows our machine translation systems to inch closer to human performance, to the more specialized resources like WordNets[3] that encode semantic relations between nodes, these resources are necessary for the general advancement of Natural Language Processing, which eventually evolves into real apps and services we are (already) taking for granted.

We introduce **RONEC** - the **RO**manian **N**amed **E**ntity **C**orpus[4], a free, open-source resource that contains annotated named entities in copy-right free text.

A named entity corpus is generally used for Named Entity Recognition (NER): the identification of entities in text such as names of persons, locations, companies, dates, quantities, monetary values, etc. This information would be very useful for any number of applications: from a general information extraction system down to task-specific apps such as identifying monetary values in invoices or product and company references in customer reviews.

We motivate *the need* for this corpus primarily because, for Romanian, there is *no other such corpus.* This basic necessity has sharply arisen as we, while working on a different project, have found out there are no usable resources to help us in an Information Extraction task: we were unable to extract people, locations or dates/values. This constituted a major road-block, with the only solution being to create such a corpus ourselves. As the corpus was out-of-scope for this project, the work was done privately, outside the umbrella of any authors' affiliations - this is why we are able to distribute this corpus completely free[5].

The current landscape in Romania regarding language resources is relatively unchanged from the outline given by the META-NET[6] project over six years ago. The in-depth analysis performed in this European-wide Horizon2020-funded project revealed that the Romanian language falls in the "fragmentary support" category, just above the last, "weak/none" category (see the language/support matrix in (Rehm and Uszkoreit, 2013)). This is why, in 2019/2020, we are able to present the first Romanian NER resource.

2. Related Work

We note that, while fragmentary, there are a few related language resources available, but *none* that *specifically* target named entities:

2.1. ROCO corpus

ROCO[7] is a Romanian journalistic corpus that contains approx. 7.1M tokens. It is rich in proper names, numerals and named entities. The corpus has been automatically annotated at word-level with morphosyntactic information (MSD annotations).

2.2. ROMBAC corpus

Released in 2016, ROMBAC[8] is a Romanian text corpus containing 41M words divided in relatively equal domains like journalism, legalese, fiction, medicine, etc. Similarly to ROCO, it is automatically annotated at word level with MSD descriptors.

2.3. CoRoLa corpus

The much larger and recently released CoRoLa corpus[9] contains over 1B words, similarly automatically annotated.

[1] BERT, released in 2018, is the baseline for today's much more advanced systems.

[2] OpenAI's GPT-2 (Radford et al., 2019) is a very strong text generation model.

[3] RoWordNet (Dumitrescu et al., 2018) is a relatively recent resource in the Romanian language space.

[4] RONEC ISLRN: 723-333-596-623-8, available at `https://github.com/dumitrescustefan/ronec`

[5] Unfortunately, many Romanian language resources have been developed in different funded projects and carry stronger copyright licenses, including requiring potential users to print/ sign/scan/send copyright forms, a step that discourages the vast majority of people.

[6] META-NET website: `http://www.meta-net.eu/`

[7] ROCO ISLRN: 312-617-089-348-7, ELRA-W0085

[8] ROMBAC ISLRN: 162-192-982-061-0, ELRA-W0088

[9] CoRoLa available at: `http://corola.racai.ro/`

In all these corpora the named entities are not a separate category - the texts are morphologically and syntactically annotated and all proper nouns are marked as such - NP - without any other annotation or assigned category. Thus, these corpora cannot be used in a true NER sense. Furthermore, annotations were done automatically with a tokenizer/tagger/parser, and thus are of slightly lower quality than one would expect of a gold-standard corpus.

3. Corpus Description

The corpus, at its current version 1.0 is composed of **5127 sentences**, annotated with **16 classes**, for a total of **26377 annotated entities**. The 16 classes are: PERSON, NAT_REL_POL, ORG, GPE, LOC, FACILITY, PRODUCT, EVENT, LANGUAGE, WORK_OF_ART, DATETIME, PERIOD, MONEY, QUANTITY, NUMERIC_VALUE and ORDINAL.

It is based on copyright-free text extracted from Southeast European Times (SETimes) (Tyers and Alperen, 2010). The news portal has published[10] "news and views from Southeast Europe" in ten languages, including Romanian. SETimes has been used in the past for several annotated corpora, including parallel corpora for machine translation. For RONEC we have used a hand-picked[11] selection of sentences belonging to several categories (see table 1 for stylistic examples).

The corpus contains the standard diacritics in Romanian: letters **ș** and **ț** are written with a comma, not with a cedilla (like **ş** and **ţ**). In Romanian many older texts are written with cedillas instead of commas because full Unicode support in Windows came much later than the classic extended ASCII which only contained the cedilla letters.

The 16 classes are inspired by the OntoNotes5 corpus (Weischedel et al., 2013) as well as the ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13 (Consortium and others, 2005). We dropped 2 classes from OntoNote's 18 classes[12]. Each one will be presented in detail, in section 4. A summary of available classes with word counts for each is available in table 2.

The corpus is available in two formats: **BRAT** and **CoNLL-U Plus** (MWE and IOB styles).

3.1. BRAT format

As the corpus was developed in the BRAT [13] environment, it was natural to keep this format as-is. BRAT is an online environment for collaborative text annotation - a web-based tool where several people can mark words, sub-word pieces, multiple word expressions, can link them together by relations, etc. The back-end format is very simple: given a text file that contains raw sentences, in another text file every annotated entity is specified by the start/end character offset as well as the entity type, one per line. RONEC is exported in the BRAT format as ready-to-use in the BRAT annotator itself. The corpus is pre-split into sub-folders, and contains all the extra files such as the entity list, etc, needed to directly start an eventual edit/extension of the corpus.

Example (raw/untokenized) sentences:

```
Tot în cadrul etapei a 2-a, a avut loc
întâlnirea Vardar Skopje - S.C. Pick
Szeged, care s-a încheiat la egalitate,
24 - 24.
I s-a decernat Premiul Nobel pentru liter-
atură pe anul 1959.
```

Example annotation format:

```
T1 ORDINAL 21 26 a 2-a
T2 ORGANIZATION 50 63 Vardar Skopje
T3 ORGANIZATION 66 82 S.C. Pick Szeged
T4 NUMERIC_VALUE 116 118 24
T5 NUMERIC_VALUE 121 123 24
T6 DATETIME 175 184 anul 1959
```

3.2. CoNLL-U Plus format

The CoNLL-U Plus[14] format extends the standard CoNLL-U which is used to annotate sentences, and in which many corpora are found today. The CoNLL-U format annotates one word per line with 10 distinct "columns" (tab separated):

1. ID: word index;
2. FORM: unmodified word from the sentence;
3. LEMMA: the word's lemma;
4. UPOS: Universal part-of-speech tag;
5. XPOS: Language-specific part-of-speech tag;
6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension;
7. HEAD: Head of the current word, which is either a value of ID or zero;
8. DEPREL: Universal dependency relation to the HEAD or a defined language-specific subtype of one;
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs;
10. MISC: Miscellaneous annotations such as space after word.

The CoNLL-U Plus extends this format by allowing a variable number of columns, with the restriction that the columns are to be defined in the header. For RONEC, we define our CoNLL-U Plus format as the standard 10 columns **plus another extra column named RONEC:CLASS**. This column has the following format[15]:

- each named entity has a distinct id in the sentence, starting from 1; as an entity can span several words, all

[10] setimes.com has ended publication in March 2015

[11] We tried to select sentences so as to both maximize the amount of named entities while also keep a balanced domain coverage.

[12] Compared to OntoNotes we dropped its LAW class as it had almost no entity in our corpus, and compressed DATE and TIME into DATETIME, as surprisingly we found many cases where the distinction between DATE and TIME would be confusing for annotators. Furthermore, DATETIME entities will usually require further sub-processing to extract exact values, something which is out of scope for this corpus.

[13] BRAT Rapid Annotation Tool: `http://brat.nlplab.org/`

[14] CoNLL-U Plus format description available at: `http://universaldependencies.org/ext-format.html`

[15] based on the PARSEME:MWE multi-word expressions, see CUPT format here.

Style	Example sentence
Current news	În **două zile**, luptele de la **Fallujah** din **Irak** au provocat moartea a **105 persoane** și rănirea a peste alte **200**.
Historical news	**Jean-Claude Juncker**, premierul **luxemburghez** s-a născut în **9 decembrie 1954**.
Free time	**Turiștii** care doresc să-și petreacă vacanța într-un loc liniștit, frumos și cu minim de cheltuieli pot opta pentru spațiile special amenajate pentru **corturi** atât la munte, cât și la mare sau în **Delta Dunării**.
Sports	Tot în cadrul etapei **a 2-a**, a avut loc întâlnirea **Vardar Skopje** - **S.C. Pick Szeged**, care s-a încheiat la egalitate, **24** - **24**.
Juridical news pieces	Ordonanța Guvernului nr. **83 / 2004** pentru modificarea și completarea Legii nr. **57 / 2003** privind Codul fiscal prevede, la art. **253**, alineatul (**6**)...
Personal adverts (e.g. buying-selling)	**S.C. "Innuendo" S.R.L.** vinde en gros, prin intermediul depozitului propriu situat în incinta **Centrului Comercial "Euro 1"**...
Editorials (written sometimes in first person)	Pe **Valea Cernei** am ajuns, de această dată, pe drumul (**DN67D**) dinspre **Baia de Aramă**.

Table 1: Stylistic domains and examples (bold marks annotated entities). Translations are depicted in Appendix A.

Class	Total words	**Total entities**	Words per entity
PERSON	10251	5363	1.911
NAT_REL_POL	1353	1324	1.022
ORGANIZATION	9794	3410	2.872
GPE	4751	4180	1.137
LOC	2364	920	2.57
FACILITY	2510	1187	2.115
PRODUCT	2042	1331	1.534
EVENT	1341	425	3.155
LANGUAGE	98	97	1.01
WORK_OF_ART	863	248	3.48
DATETIME	7072	3003	2.355
PERIOD	1295	385	3.364
MONEY	2591	898	2.885
QUANTITY	769	360	2.136
NUMERIC_VALUE	2807	2714	1.034
ORDINAL	859	532	1.615
Total	50760	**26377**	2.137

Table 2: Corpus statistics: Each entity is marked with a class and can span one or more words

words that belong to it have the same id (no relation to word indexes)

- the first word belonging to an entity also contains its class (e.g. word "John" in entity "John Smith" will be marked as "1:PERSON")
- a non-entity word is marked with an asterisk *

Table 3 shows the CoNLL-U Plus format where for example "a 2-a" is an ORDINAL entity spanning 3 words. The first word "a" is marked in this last column as "1:ORDINAL" while the following words just with the id "1".

We also release the same CoNLL-U Plus format but the last column is encoded in the classic IOB format.

The CoNLL-U Plus format we provide was created as follows: (1) annotate the raw sentences using the NLP-Cube[16] tool for Romanian (it provides everything from tokenization to parsing, filling in all attributes in columns #1-#10; (2) align each token with the human-made entity annotations from the BRAT environment (the alignment is done automatically and is error-free) and fill in column #11.

4. RONEC Classes

For the English language, we found two "categories" of NER annotations to be more prominent: CoNLL- and ACE-style. Because CoNLL only annotates a few classes (depending on the corpora, starting from the basic three: PERSON, ORGANIZATION and LOCATION, up to seven), we chose to follow the ACE-style with 18 different classes. After analyzing the ACE guide we have settled on 16 final classes that seemed more appropriate for Romanian, seen in table 2.

In the following sub-sections we will describe each class in turn, with a few examples. Some examples have been left in Romanian while some have been translated in English for the reader's convenience. In the examples at the end of each class' description, translations in English are colored for easier reading.

4.1. PERSON

Persons, including fictive characters. We also mark common nouns that refer to a person (or several), including pronouns (us, them, they), but not articles (e.g. in "an individual" we don't mark "an"). Positions are not marked unless they directly refer to the person: "The presidential counselor has advised ... that a new counselor position is open.", here we mark "presidential counselor" because it refers to a person and not the "counselor" at the end of the sentence as it refers only to a position.

[16]NLP-Cube is a multilingual text preprocessing tool with SOTA-level accuracy, that exports directly in CoNLL format and is available at `https://github.com/adobe/NLP-Cube`

ID (#1)	FORM (#2)	LEMMA (#3)	UPOS (#4)	XPOS (#5)	HEAD (#7)	DEPREL (#8)	**RONEC:CLASS (#11)**
1	Tot	tot	ADV	Rp	3	advmod	*
2	în	în	ADP	Spsa	3	case	*
3	cadrul	cadru	NOUN	Ncmsry	10	obl	*
4	etapei	etapă	NOUN	Ncfsoy	3	nmod	*
5	**a**	al	DET	Tsfs	6	det	**1:ORDINAL**
6	**2**	2	NUM	Mc-p-d	4	nummod	**1**
7	**-a**	-a	DET	Tffs-y	6	det	**1**
8	,	,	PUNCT	COMMA	3	punct	*
9	a	avea	AUX	Va–3s	10	aux	*
10	avut	avea	VERB	Vmp–sm	0	root	*
11	loc	loc	NOUN	Ncms-n	10	fixed	*
12	întâlnirea	întâlnire	NOUN	Ncfsry	10	nsubj	*
13	**Vardar**	Vardar	PROPN	Np	12	nmod	**2:ORGANIZATION**
14	**Skopje**	Skopje	PROPN	Np	13	flat	**2**
15	-	-	PUNCT	DASH	13	punct	*
16	**S.C.**	s.c.	NOUN	Yn	13	conj	**3:ORGANIZATION**
17	**Pick**	Pick	PROPN	Np	13	flat	**3**
18	**Szeged**	Szeged	PROPN	Np	17	flat	**3**
19	,	,	PUNCT	COMMA	23	punct	*
20	care	care	PRON	Pw3–r	23	nsubj	*

Table 3: CoNLL-U Plus format for the first 20 tokens of sentence "*Tot în cadrul etapei a 2-a, a avut loc întâlnirea Vardar Skopje - S.C. Pick Szeged, care s-a încheiat la egalitate, 24 - 24.*" (bold marks entities). The format is a text file containing a token per line annotated with 11 tab-separated columns, with an empty line marking the start of a new sentence. Please note that only column #11 is human annotated (and the target of this work), the rest of the morpho-syntactic annotations have been automatically generated with NLP-Cube (Boroș et al., 2018).

Locul doi i-a revenit româncei **Otilia Aionesei**, o **elevă** de 17 ani.
The second place was won by **Otilia Aionesei**, a 17 year old **student**.

Ministrul bulgar **pentru afaceri europene, Meglena Kuneva** ...
The Bulgarian **Minister for European Affairs**, **Meglena Kuneva** ...[17]

4.2. NAT_REL_POL

These are nationalities or religious or political groups. We include words that indicate the nationality of a person, group or product/object. Generally words marked as NAT_REl_POL are adjectives.

---

avionul **american**
the **American** airplane

Grupul **olandez**
the **Dutch** group

Grecii își vor alege președintele.
The **Greeks** will elect their president.

---

[17]Note: in Romanian word ordering makes for two entities while in English it looks like just one.

4.3. ORGANIZATION

Companies, agencies, institutions, sports teams, groups of people. These entities must have an organizational structure. We only mark full organizational entities, not fragments, divisions or sub-structures.

Universitatea Politehnica București a decis ...
The **Politehnic University of Bucharest** has decided ...

Adobe Inc. a lansat un nou produs.
Adobe Inc. has launched a new product.

4.4. GPE

Geo-political entities: countries, counties, cities, villages. GPE entities have *all* of the following components: (1) a population, (2) a well-defined governing/organizing structure and (3) a physical location. GPE entities are not sub-entities (like a neighbourhood from a city).

Armin van Buuren s-a născut în **Leiden.**
Armin van Buuren was born in **Leiden.**

U.S.A. ramane indiferentă amenințărilor **Coreei de Nord.**
U.S.A. remains indifferent to **North Korea**'s threats.

4.5. LOC

Non-geo-political locations: mountains, seas, lakes, streets, neighbourhoods, addresses, continents, regions that are not GPEs. We include regions such as Middle East, "continents" like Central America or East Europe. Such regions include multiple countries, each with its own government and thus cannot be GPEs.

Pe **DN7 Petroșani–Obârșia Lotrului** carosabilul era umed, acoperit (cca 1 cm) cu zăpadă, iar de la Obârșia Lotrului la stațiunea Vidra, stratul de zăpadă era de 5-6 cm.
On **DN7 Petroșani–Obârșia Lotrului** the road was wet, covered (about 1cm) with snow, and from Obârșia Lotrului to Vidra resort the snow depth was around 5-6 cm.[18]

Produsele comercializate în **Europa de Est** au o calitate inferioară celor din **vest.**
Products sold in **East Europe** have a lower quality than those sold in the **west.**[19]

4.6. FACILITY

Buildings, airports, highways, bridges or other functional structures built by humans. Buildings or other structures which house people, such as homes, factories, stadiums, office buildings, prisons, museums, tunnels, train stations, etc., named or not. Everything that falls within the architectural and civil engineering domains should be labeled as a FACILITY. We do not mark structures composed of multiple (and distinct) sub-structures, like a named area that is composed of several buildings, or "micro"-structures such as an apartment (as it a unit of an apartment building). However, larger, named functional structures can still be marked (such as "terminal X" of an airport).

Autostrada A2 a intrat în reparații pe o bandă, însă pe **A1** nu au fost încă începute lucrările.
Repairs on one lane have commenced on the **A2 highway**, while on **A1** no works have started yet.

Aeroportul Henri Coandă ar putea sa fie extins cu un nou **terminal.**
Henri Coandă Airport could be extended with a new **terminal.**

4.7. PRODUCT

Objects, cars, food, items, anything that is a product, including software (such as Photoshop, Word, etc.). We don't mark services or processes. With very few exceptions (such as software products), PRODUCT entities have to have physical form, be directly man-made. We don't mark entities such as credit cards, written proofs, etc. We don't include the producer's name unless it's embedded in the name of the product.

Mașina cumpărată este o **Mazda.**
The bought **car** is a **Mazda.**

S-au cumpărat 5 **Ford Taurus** și 2 **autobuze** Volvo.
5 **Ford Taurus** and 2 Volvo **buses** have been acquired.[20]

4.8. EVENT

Named events: Storms (e.g.:"Sandy"), battles, wars, sports events, etc. We don't mark sports teams (they are ORGs), matches (e.g. "Steaua-Rapid" will be marked as two separate ORGs even if they refer to a football match between the two teams, but the match is not specific). Events have to be significant, with at least national impact, not local.

Războiul cel Mare, **Războiul Națiunilor**, denumit, în timpul celui de **Al Doilea Război Mondial**, **Primul Război Mondial**, a fost un conflict militar de dimensiuni mondiale.
The **Great War**, **War of the Nations**, as it was called during the **Second World War**, the **First World War** was a global-scale military conflict.

4.9. LANGUAGE

This class represents all languages.

Românii din România vorbesc **română.**
Romanians from Romania speak **Romanian.**[21]

În Moldova se vorbește **rusa** și **româna.**
In Moldavia they speak **Russian** and **Romanian.**

4.10. WORK_OF_ART

Books, songs, TV shows, pictures; everything that is a work of art/culture created by humans. We mark just their name. We don't mark laws.

Accesul la **Mona Lisa** a fost temporar interzis vizitatorilor.
Access to **Mona Lisa** was temporarily forbidden to visitors.

În această seară la **Vrei sa Fii Miliardar** vom avea un invitat special.
This evening in **Who Wants To Be A Millionaire** we will have a special guest.

---

[18]Note: "Obârșia Lotrului" and "Vidra resort" are cities or villages and are thus GPEs; only DN7 which is a national road designation is marked as LOC, including where exacly on DN7 (names of cities are used as markers for the road segment)

[19]Note: "west" refers to West Europe and thus we mark it as a LOC.

[20]Note: here we won't mark "Volvo" but will mark "Ford" as in one two-word entity "Ford Taurus" as it is embedded in the name.

[21]Note: we mark languages, not countries (which are GPEs) or the country's inhabitants (which are NAT_REL_POL)

4.11. DATETIME

Date and time values. We will mark full constructions, not parts, if they refer to the same moment (e.g. a comma separates two distinct DATETIME entities only if they refer to distinct moments). If we have a well specified period (e.g. "between 20-22 hours") we mark it as PERIOD, otherwise less well defined periods are marked as DATETIME (e.g.: "last summer", "September", "Wednesday", "three days"); Ages are marked as DATETIME as well. Prepositions are not included.

Te rog să vii aici în cel mult **o oră**, nu **mâine** sau **poimâine.**
Please come here in **one hour** at most, not **tomorrow** or the **next day.**

Actul s-a semnat la **orele 16.**
The paper was signed at **16 hours.**

August este **o lună** secetoasă.
August is a dry **month.**

Pe **data de 20 martie** între orele 20-22 va fi oprită alimentarea cu curent.
On the **20th of March**, between 20-22 hours, electricity will be cut-off.[22]

4.12. PERIOD

Periods/time intervals. Periods have to be very well marked in text. If a period is not like "a-b" then it is a DATETIME.

Spectacolul are loc între **1 și 3 Aprilie.**
The show takes place between **1 and 3 April.**

În prima jumătate a lunii iunie va avea loc evenimentul de două zile.
In the first half of June the two-day event will take place.[23]

4.13. MONEY

Money, monetary values, including units (e.g. USD, $, RON, lei, francs, pounds, Euro, etc.) written with number or letters. Entities that contain any monetary reference, including measuring units, will be marked as MONEY (e.g. 10$/sqm, 50 lei per hour). Words that are not clear values will not be marked, such as "an amount of money", "he received a coin".

Primarul a semnat un contract în valoare de **10 milioane lei noi**, echivalentul a aproape **2.6m EUR.**
The mayor signed a contract worth **10 million new lei**, equivalent of almost **2.6m EUR.**

---

[22]Note: "20-22 hours" is a PERIOD and not a DATETIME, this is why it is not marked here as such.

[23]Note: "the first half of June" while it is a period, because it is not clearly specified, it will be marked as DATETIME. Also "two-day" is a DATETIME because we don't know exactly which 2 days.

4.14. QUANTITY

Measurements, such as weight, distance, etc. Any type of quantity belongs in this class.

Conducătorul auto avea peste **1g/ml** alcool în sânge, fiind oprit deoarece a fost prins cu peste **120 km/h** în localitate.
The car driver had over **1g/ml** blood alcohol, and was stopped because he was caught speeding with over **120km/h** in the city.

4.15. NUMERIC_VALUE

Any numeric value (including phone numbers), written with letters or numbers or as percents, which *is not* MONEY, QUANTITY or ORDINAL.

Raportul **XII-2** arată **4 552** de investitori, iar structura de portofoliu este: cont curent **0,05%**, certificate de trezorerie **66,96%**, depozite bancare **13,53%**, obligațiuni municipale **19,46%.**
The **XII-2** report shows **4 552** investors, and the portfolio structure is: current account **0,05%**, treasury bonds **66,96%**, bank deposits **13,53%**, municipal bonds **19,46%.**

4.16. ORDINAL

The first, the second, last, 30th, etc.; An ordinal must imply an order relation between elements. For example, "second grade" does not involve a direct order relation; it indicates just a succession in grades in a school system.

Primul loc a fost ocupat de echipa Germaniei.
The first place was won by Germany's team.

---

5. Annotation Methodology

The corpus creation process involved a small number of people that have voluntarily joined the initiative, with the authors of this paper directing the work. Initially, we searched for NER resources in Romanian, and found none. Then we looked at English resources and read the in-depth ACE guide, out of which a 16-class draft evolved. We then identified a copy-right free text from which we hand-picked sentences to maximize the amount of entities while maintaining style balance. The annotation process was a trial-and-error, with cycles composed of annotation, discussing confusing entities, updating the annotation guide schematic and going through the corpus section again to correct entities following guide changes. The annotation process was done online, in BRAT[24]. The actual annotation involved 4 people, has taken about 6 months (as work was volunteer-based, we could not have reached for 100% time commitment from the people involved), and followed the steps:

[24]Please note that while the CoNLLU-Plus MWE format supports multi word entities that are not in a **continuous** sequence, as we performed the annotation in BRAT, we only annotated multi-word contiguous entities.

1. Each person would annotate the full corpus (this included the cycles of shaping up the annotation guide, and re-annotation). Inter-annotator agreement (ITA) at this point was relatively low, at 60-70%, especially for a number of classes.
2. We then automatically merged all annotations, with the following criterion: if 3 of the 4 annotators agreed on an entity (class&start-stop), then it would go unchanged; otherwise mark the entity (longest span) as CONFLICTED.
3. Two teams were created, each with two persons. Each team annotated the full corpus again, starting from the previous step. At this point, class-average ITA has risen to over 85%.
4. Next, the same automatic merging happened, this time entities remained unchanged if both annotations agreed.
5. Finally, one of the authors went through the full corpus one more time, correcting disagreements.

Notes regarding classes and inter-annotator agreements:

- ORGANIZATION, NAT_REL_POL, LANGUAGE or GPEs have the highest ITA, over 98%.
- DATETIME also has a high ITA, with some overlap with PERIOD: annotators could fall-back if they were not sure that an expression was a PERIOD and simply mark it as DATETIME.
- WORK_OF_ART and EVENTs have caused some problems because the scope could not be properly defined from just one sentence. For example, a fair in a city could be a local event, but could also be a national periodic event.
- MONEY, QUANTITY and ORDINAL all are more specific classes than NUMERIC_VALUE. So, in cases where a numeric value has a unit of measure by it, it should become a QUANTITY, not a NUMERIC_VALUE. However, this "specificity" has created some confusion between these classes, just like with DATETIME and PERIOD.
- The ORDINAL class is a bit ambiguous, because, even though it ranks "higher" than NUMERIC_VALUE, it is the least diverse, most of the entities following the same patterns.
- PRODUCT and FACILITY classes have the lowest ITA by far (less than 40% in the first annotation cycle, less than 70% in the second). We actually considered removing these classes from the annotation process, but to try to mimic the OntoNotes classes as much as possible we decided to keep them in. There were many cases where the annotators disagreed about the scope of words being facilities or products. Even in the ACE guidelines these two classes are not very well "documented" with examples of what is and what is not a PRODUCT or FACILITY. Considering that these classes are, in our opinion, of the lowest importance amongst all others, a lower ITA was accepted.

Finally, we would like to address the "semantic scope" of the entities - for example, for class PERSON, we do not annotate only proper nouns (NPs) but basically any reference to a person (e.g. through pronouns "she", job position titles, common nouns such as "father", etc.). We do this because we would like a high-coverage corpus, where entities are marked as more semantically-oriented rather than syntactically - in the same way ACE entities are more encompassing than CoNLL entities [25].

6. Conclusions

We have presented RONEC - the first Named Entity Corpus for the Romanian language. At its current version, in its 5127 sentences we have 26377 annotated entities in 16 different classes. The corpus is based on copy-right free text, and is released as open-source, free to use and extend. There is also an annotation guide that we will improve, and in time evolve into a full annotation document like the ACE Annotation Guidelines for Entities (Consortium and others, 2005).

We have released the corpus in two formats: CoNLL-U PLus (text-based tab-separated pre-tokenized and annotated format, in MWE and IOB flavours) and BRAT (another text-based, non-tokenized format which annotates spans of characters with classes).

We also release a spaCy[26] pretrained NER model on our GitHub repo for immediate usage.

7. Bibliographical References

Boroș, T., Dumitrescu, S. D., and Burtica, R. (2018). Nlp-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179.

Consortium, L. D. et al. (2005). Ace (automatic content extraction) english annotation guidelines for entities. *Version*, 5(6):2005–08.

Dumitrescu, S. D., Avram, A. M., Morogan, L., and Toma, S.-A. (2018). Rowordnet–a python api for the romanian wordnet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rehm, G. and Uszkoreit, H. (2013). *META-NET strategic research agenda for multilingual Europe 2020*. Springer.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

[25] RONEC contains a total of 1251 NPs for the class PERSON. The full list can be found at: `https://github.com/dumitrescustefan/ronec/blob/master/ronec/meta/person_proper_nouns.txt`

[26] spaCy is a well-known Python text processing API, offering an easy interface to everything from tokenization to parsing and NER.

Appendix

A Table Translations

This appendix contains the translations of the phrases in table 1 (in order of appearance):

- In **two days**, the **Fallujah** battles from **Iraq** caused the death of **105 people** and injured more than **200**.
- **Jean-Claude Juncker**, Prime Minister of **Luxembourg** was born on **December 9, 1954**.
- **Tourists** that want to spend their vacations in a quiet, beautiful, and with a minimum of expenses, can opt for spaces specially set up for **tents** either on the mountain, at the sea or in the **Danube Delta**.
- Also in **the second** stage, the **Vardar Skopje** - **S.C. Pick Szeged** meeting took place, which ended on equal footing, **24-24**.
- Ordinance of the Government no. **83 / 2004** for amending and supplementing Law no. **57 / 2003** regarding the Fiscal Code stipulates, at art. **253**, paragraph (**6**)...
- **SC "Innuendo" S.R.L.** sells in bulk, through its own warehouse located inside the **"Euro 1" Shopping Center** ...
- On **Cerna Valley** we arrived, this time, on the road (**DN67D**) from **Baia de Aramă**.

B BRAT to CoNLLU-Plus conversion

The corpus was annotated using BRAT, which does not take into account tokenization: it simply marks character spans with a class. For this reason, the conversion was done by directly tokenizing the text with NLP-Cube and then aligning the resulting tokens to their respective class. There weren't any cases where tokenization produced a token that belonged to 2 different classes.

C SpaCy integration

At the time of writing the model is not yet integrated natively into spaCy's repo to be able to be used directly. However, we provide a pre-trained model available here[27]. The only extra step is to download this model locally.
Here is how to use spaCy & RONEC:

```
import spacy

nlp = spacy.load(<model>)
doc = nlp("Popescu Ion a fost la Cluj.")

for ent in doc.ents:
    print(ent.text, ent.start_char,
          ent.end_char, ent.label)
```

and we should see an output like:

```
Popescu Ion 0 11 PERSON
Cluj 22 26 GPE
```

[27] https://github.com/dumitrescustefan/ronec/tree/master/spacy