# Slav-NER: the 3rd Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic languages

**Jakub Piskorski,[1] Bogdan Babych,[2] Zara Kancheva,[3] Olga Kanishcheva,[4] Maria Lebedeva,[5] Michał Marcińczuk,[6] Preslav Nakov,[7] Petya Osenova,[3] Lidia Pivovarova,[8] Senja Pollak,[9] Pavel Přibáň,[10] Ivaylo Radev,[3] Marko Robnik-Šikonja,[11] Vasyl Starko,[12] Josef Steinberger,[10] Roman Yangarber[8]**

[1]Polish Academy of Sciences, Warsaw, Poland jpiskorski@gmail.com
[2]Heidelberg University, Germany bogdan.babych@iued.uni-heidelberg.de
[3]IICT-BAS, Bulgarian Academy of Sciences, Bulgaria {petya,zara,radev}@bultreebank.org
[4]Kharkiv Polytechnic Institute, Ukraine kanichshevaolga@gmail.com
[5]Pushkin State Russian Language Institute, Moscow, Russia mylebedeva@pushkin.institute
[6]Wrocław University of Science and Technology, Poland michal.marcinczuk@pwr.edu.pl
[7]Qatar Computing Research Institute, HBKU pnakov@hbku.edu.qa
[8]University of Helsinki, Finland first.last@helsinki.fi
[9]Jozef Stefan Institute, Slovenia senja.pollak@ijs.si
[10]University of West Bohemia, Czech Republic {pribanp,jstein}@kiv.zcu.cz
[11]University of Ljubljana, Slovenia Marko.RobnikSikonja@fri.uni-lj.si
[12]Ukrainian Catholic University, Lviv, Ukraine vstarko@gmail.com

## Abstract

This paper describes Slav-NER: the 3rd Multilingual Named Entity Challenge in Slavic languages. The tasks involve recognizing mentions of named entities in Web documents, normalization of the names, and cross-lingual linking. The Challenge covers six languages and five entity types, and is organized as part of the $8^{th}$ Balto-Slavic Natural Language Processing Workshop, co-located with the EACL 2021 Conference. Ten teams participated in the competition. Performance for the named entity recognition task reached 90% F-measure, much higher than reported in the first edition of the Challenge. Seven teams covered all six languages. Detailed evaluation information is available on the shared task web page.

## 1 Introduction

Analyzing named entities (NEs) in Slavic languages poses a challenging problem, due to the rich inflection and derivation, free word order, and other morphological and syntactic phenomena exhibited in these languages (Przepiórkowski, 2007; Piskorski et al., 2009). Encouraging research on detection and normalization of NEs—and on the closely related problem of cross-lingual, cross-document *entity linking*—is of paramount importance for improving multilingual and cross-lingual information access in these languages.

This paper describes the $3^{rd}$ Shared Task on multilingual NE recognition (NER), which aims at addressing these problems in a systematic way. The shared task was organized in the context of the $8^{th}$ BSNLP: Balto-Slavic Natural Language Processing Workshop, co-located with the EACL 2021 conference. The task covers six languages—Bulgarian, Czech, Polish, Russian, Slovene and Ukrainian—and five types of NE: person, location, organization, product, and event. The input text collection consists of documents collected from the Web, each collection centered on a certain "focal" event. The rationale of such a setup is to foster the development of "end-to-end" NER and cross-lingual entity linking solutions, which are not tailored to specific, narrow domains. This paper also serves as an introduction and a guide for researchers wishing to explore these problems using the training and test data, which are released to the public.[1]

The paper is organized as follows. Section 2 reviews prior work. Section 3 describes the task; Section 4 describes the annotation of the dataset. The evaluation methodology is introduced in Section 5. Participant systems are described in Section 6, and the results obtained by these systems are presented in Section 7. We present the conclusions and lessons learned in Section 8.

## 2 Prior Work

The work described here builds on the $1^{st}$ and $2^{nd}$ Shared Task on Multilingual Named Entity Recognition, Normalization and cross-lingual Match-

[1] bsnlp.cs.helsinki.fi/shared_task.html

ing for Slavic Languages, (Piskorski et al., 2017, 2019), which, to the best of our knowledge, are the first attempts at such shared tasks covering multiple Slavic languages.

High-quality recognition and analysis of NEs is an essential step not only for information access, such as document retrieval and clustering, but it also constitutes a fundamental processing step in a wide range of NLP pipelines built for higher-level analysis of text, such as Information Extraction, see, e.g. (Huttunen et al., 2002). Other NER-related shared tasks have been organized previously. The first *non-English* monolingual NER evaluations—covering Chinese, Japanese, Spanish, and Arabic—were held in the context of the Message Understanding Conferences (MUCs) (Chinchor, 1998) and the ACE Programme (Doddington et al., 2004). The first *multilingual* NER shared task, which covered several European languages, including Spanish, German, and Dutch, was organized in the context of the CoNLL conferences (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The NE types covered in these campaigns were similar to the NE types covered in our Challenge. Worth mentioning in this context is Entity Discovery and Linking (EDL) (Ji et al., 2014, 2015), a track of the NIST Text Analysis Conferences (TAC). EDL aimed to extract entity mentions from a collection of documents in multiple languages (English, Chinese, and Spanish), and to partition the entities into cross-document equivalence classes, by either linking mentions to a knowledge base or directly clustering them. An important difference between EDL and our task is that EDL required linking entities to a pre-existing knowledge base.

Related to cross-lingual NE recognition is NE transliteration, i.e., linking NEs across languages that use different scripts. A series of NE Transliteration Shared Tasks were organized as a part of NEWS—Named Entity Workshops—(Duan et al., 2016), focusing mostly on Indian and Asian languages. In 2010, the NEWS Workshop included a shared task on Transliteration Mining (Kumaran et al., 2010), i.e., mining of names from parallel corpora. This task included corpora in English, Chinese, Tamil, Russian, and Arabic.

Research on NE focusing on Slavic languages includes tools for NE recognition for Croatian (Karan et al., 2013; Ljubešić et al., 2013), NE recognition in Croatian tweets (Baksa et al., 2017), a manually annotated NE corpus for Croatian (Agić and Ljubešić, 2014), tools for NE recognition in Slovene (Štajner et al., 2013; Ljubešić et al., 2013), a Czech corpus of 11K annotated NEs (Ševčíková et al., 2007), NER tools for Czech (Konkol and Konopík, 2013), tools and resources for fine-grained annotation of NEs in the National Corpus of Polish (Waszczuk et al., 2010; Savary and Piskorski, 2011), NER shared tasks for Polish organized under the umbrella of POLEVAL[2] evaluation campaigns (Ogrodniczuk and Łukasz Kobyliński, 2018, 2020). and a recent shared task on NE Recognition in Russian (Starostin et al., 2016).

## 3 Task Description

The data for this edition of the shared task consists of sets of documents in six Slavic languages: Bulgarian, Czech, Polish, Russian, Slovene and Ukrainian. To accommodate entity linking, each set of documents is chosen to revolve around one certain entity—e.g., a person, an organization or an event. The documents were obtained from the Web, by posing a keyword query to a search engine or publicly available crawled data repositories, and extracting the textual content from the respective sources.

The task is to recognize, classify, and "normalize" all named-entity mentions in each of the documents, and to link across languages all named mentions referring to the same real-world entity. Formally, the Multilingual Named Entity Recognition task is subdivided into three sub-tasks:

- **Named Entity Mention Detection and Classification:** Recognizing all named mentions of entities of five types: persons (PER), organizations (ORG), locations (LOC), products (PRO), and events (EVT).

- **Name Normalization:** Mapping each named mention of an entity to its corresponding *base form*. By "base form" we generally mean the lemma ("dictionary form") of the inflected word-form. In some cases normalization should go beyond inflection and transform a derived word into a base word's lemma, e.g., in case of personal possessives (see below). Multi-word names should be normalized to the *canonical multi-word expression*—rather than a sequence

[2] http:\\poleval.pl

of lemmas of the words making up the multi-word expression.

- **Entity Linking.** Assigning a unique identifier (ID) to each detected named mention of an entity, in such a way that mentions referring to the same real-world entity should be assigned the same ID—referred to as the cross-lingual ID.

The task does not require positional information of the name entity mentions. Thus, for all occurrences of the same form of a NE mention (e.g., an inflected variant, an acronym or abbreviation) within a given document, no more than one annotation should be produced.[3] Furthermore, distinguishing typographical case is not necessary since the evaluation is case-insensitive. If the text includes lowercase, uppercase or mixed-case variants of the same entity, the system should produce only one annotation for all of these mentions. For instance, for "*ISIS*" and "*isis*" (provided that they refer to the same NE type), only one annotation should be produced. The recognition of common-noun or pronominal references to named entities does not constitute part of the task.

### 3.1 Named Entity Classes

The task defines the following five NE classes.

**Person names (PER):** Names of real (or fictional) persons). Person names should not include titles, honorifics, and functions/positions. For example, in the text fragment "*. . . President Vladimir Putin. . .* ", only "*Vladimir Putin*" is recognized as a person name. Both initials and pseudonyms are also considered named mentions of persons. Similarly, toponym-based named references to groups of people (that do not have a formal organization unifying them) should also be recognized, e.g., "*Germans*." In this context, mentions of a single member belonging to such groups, e.g., "*German*," should be assigned the same cross-lingual ID as plural mentions, i.e., "*Germans*" and "*German*" when referring to the nation receive the same cross-lingual ID.

Named mentions of other groups of people that do have a formal organization unifying them should be tagged as PER, e.g., in the phrase "*Spart'ané vyhráli*" (Spartans won), "*Spart'ané* are to be tagged as PER.

[3]Unless the different occurrences have different entity types (different *readings*) assigned to them, which is rare.

Personal possessives derived from a person's name should be classified as a Person, and the base form of the corresponding name should be extracted. For instance, in "*Trumpov tweet*" (Croatian) one is expected to classify "*Trumpov*" as PER, with the base form "*Trump*."

**Locations (LOC):** All toponyms and geopolitical entities—cities, counties, provinces, countries, regions, bodies of water, land formations, etc.—including named mentions of *facilities*—e.g., stadiums, parks, museums, theaters, hotels, hospitals, transportation hubs, churches, streets, railroads, bridges, and similar facilities.

In case named mentions of facilities *also* refer to an organization, the LOC tag should be used. For example, from the text "*San Rafaelle Hospital hired new staff due to Covid-19 pandemic*" the mention "*San Rafaelle Hospital*" should be classified as LOC.

**Organizations (ORG):** All organizations, including companies, public institutions, political parties, international organizations, religious organizations, sport organizations, educational and research institutions, etc.

Organization designators and potential mentions of the seat of the organization are considered to be part of the organization name. For instance, from the text "*...Zakład Ubezpieczeń Społecznych w Bydgoszczy...*" (The Social Insurance Institution in Bydgoszcz), the full phrase "*Zakład Ubezpieczeń Społecznych w Bydgoszczy*" should be extracted.

**Products (PRO):** All names of *products and services*, such as electronics ("*Samsung Galaxy A41*"), cars ("*Honda Pilot*"), newspapers ("*Der Spiegel*"), web-services ("*Pintertest*"), medicines ("*Oxycodone*"), awards ("*Pulitzer Prize*"), books ("*Animal Farm*"), TV programmes ("*Wiadomości TVP*"), etc.

When a company name is used to refer to a *service*, e.g., "*na Instagramie*" (Polish for "on Instagram"), the mention of "*Instagramie*" is considered to refer to a service/product and should be tagged as PRO. However, when a company name refers to a service, expressing an opinion of the company, it should be tagged as ORG.

This category also includes legal documents and treaties, e.g., "*Układ z Schengen*" (Pol-


Figure 1: Screenshot of the Inforex Web interface, the tool used for data annotation.

Jacob Serrano (23) z americké Floridy se stal vůbec prvním Američanem, který byl očkován experimentální vakcínou proti koronaviru, ta vznikla za spolupráce vědců z Oxfordské univerzity a farmaceutické společnosti AstraZeneca. Podle WHO jde zatím o nejslibnější očkovací látku. Serrano se neváhal zapojit se do boje s koronavirem, který způsobuje nemoc covid-19, nákaza ho totiž připravila o 7 příbuzných, uvedl list Daily Mail.

114

| | | | |
|---|---|---|---|
| Američanem | Američan | PER | GPE-USA |
| AstraZeneca | AstraZeneca | ORG | ORG-AstraZeneca |
| Daily Mail | Daily Mail | PRO | PRO-Daily-Mail |
| Floridy | Florida | LOC | GPE-Florida |
| Jacob Serrano | Jacob Serrano | PER | PER-Jacob-Serrano |
| Oxfordské univerzity | Oxfordská univerzita | ORG | ORG-University-of-Oxford |
| Serrano | Serrano | PER | PER-Jacob-Serrano |
| WHO | WHO | ORG | ORG-World-Health-Org |
| covid-19 | covid-19 | EVT | EVT-Covid-19 |
| koronavirem | koronavirus | EVT | EVT-Covid-19 |
| koronaviru | koronavirus | EVT | EVT-Covid-19 |

Figure 2: Example input and output formats.

ish: "Schengen Agreement") and initiatives, e.g., "*Horizon 2020*".

**Events (EVT):** This category covers named mentions of events, including conferences, e.g. "24. Konference Žárovného Zinkování" (Czech: "Hot Galvanizing Conference"), concerts, festivals, holidays, e.g., "*Święta Bożego Narodzenia*" (Polish: "Christmas"), wars, battles, disasters, e.g., "*Katastrofa Smoleńska*" (Polish: "the Smoleńsk air disaster"), outbreaks of infectious diseases ("*Spanish Flu*"). Future, speculative, and fictive events—e.g., "'Polexit"—are considered event mentions too.

### 3.2 Complex and Ambiguous Entities

In case of complex named entities, consisting of nested named entities, only the *top-most* entity should be recognized. For example, from the text "*Università Commerciale Luigi Bocconi*" one should not extract "*Luigi Bocconi*", but only the top-level entity.

In case one word-form (e.g., "*Georgia*") is used to refer to more than one different real-world entities in different contexts in the same document (e.g., a person and a location), two annotations should be returned, associated with different cross-lingual IDs.

In case of coordinated phrases, like "*European and German Parliament*," two names should be extracted (as ORG). The lemmas would be "*European*" and "*German Parliament*", and the IDs should refer to "*European Parliament*" and "*German Parliament*" respectively.

In rare cases, plural forms might have two annotations—e.g., in the phrase "*a border between Irelands*"—"*Irelands*" should be extracted twice with identical lemmas but different IDs.

### 3.3 System Input and Response

**Input Document Format:** Documents in the collection are represented in the following format. The first five lines contain the following metadata (in the respective order): `<DOCUMENT-ID>`,

<LANGUAGE>, <CREATION-DATE>, <URL>, <TITLE>, <TEXT>. The text to be processed begins from the sixth line and runs till the end of file. The <URL> field stores the origin from which the text document was retrieved. The values of <CREATION-DATE> and <TITLE> were not provided for all documents, due to unavailability of such data or due to errors in parsing during data collection.

**System Response.** For each input file, the system should return one output file as follows. The first line should contain only the <DOCUMENT-ID>, which corresponds to the input. Each subsequent line contains one annotation, as tab-separated fields:

```
<MENTION> TAB <BASE> TAB <CAT> TAB <ID>
```

The <MENTION> field should be the NE as it appears in text. The <BASE> field should be the base form of the entity. The <CAT> field stores the category of the entity (ORG, PER, LOC, PRO, or EVT) and <ID> is the cross-lingual identifier. The cross-lingual identifiers may consist of an arbitrary sequence of alphanumeric characters. An example document in Czech and the corresponding response is shown in Figure 2.

The detailed descriptions of the tasks are available on the web page of the Shared Task.[4]

## 4 Data

For Russian, Polish, Czech and Bulgarian, the training and test data sets from the 2019 Shared Task were used as training data for 2021. For the new languages—Ukrainian and Slovene—new training sets were annotated. The test data in all six languages covered two major current topics: the COVID-19 pandemic and the 2020 USA Presidential elections (USA 2020 ELECTIONS).

The 2019 training data consist of four sets of documents extracted from the Web, each related to a given *focus* entity. We tried to choose entities related to events in 2018 and 2019 covered in mainstream news in many languages. ASIA BIBI, which relates to a Pakistani woman involved in a blasphemy case, BREXIT, RYANAIR, which faced a massive strike, and NORD STREAM, a controversial Russian-European project.

Each dataset was created as follows. For the focus entity, we posed a search query to Google and/or publicly available crawled data repositories, in each of the target languages. The query returned documents in the target language. We removed duplicates, downloaded the HTML—mainly news articles—and converted them into plain text. Since the result of HTML parsing may include not only the main text of a Web page, but also spurious text, some additional manual cleaning was applied whenever necessary. The resulting set of "cleaned" documents were used to manually select documents for each language and topic, for the final datasets.

Documents were annotated using the Inforex[5] web-based system for annotation of text corpora (Marcińczuk et al., 2017). Inforex allows parallel access and resource sharing by multiple annotators. It let us share a common list of entities, and perform entity-linking semi-automatically: for a given entity, an annotator sees a list of entities of the same type inserted by all annotators and can select an entity ID from the list. A snapshot of the Inforex interface is in Figure 1.

In addition, Inforex keeps track of all lemmas and IDs inserted for each surface form, and inserts them automatically, so in many cases the annotator only confirms the proposed values, which speeds up the annotation process a great deal. All annotations were made by native speakers. After annotation, we performed automatic and manual consistency checks, to reduce annotation errors, especially in entity linking.

Training and test data statistics are presented in Table 1 and 2 respectively.

The testing datasets—COVID-19 and USA 2020 ELECTIONS—were released to the participants who were given circa 2 days to return up to 5 system responses. The participants did not know the topics in advance, and did not receive the annotations. The main drive behind this decision was to push participants to build a general solution for Slavic NER, rather than to optimize their models toward a particular set of names.

## 5 Evaluation Methodology

The NER task (exact case-insensitive matching) and Name Normalization (or "lemmatization") were evaluated in terms of precision, recall, and F1-measure. For NER, two types of evaluations were carried out:

[4] http://bsnlp.cs.helsinki.fi/System_response_guidelines-1.2.pdf

[5] github.com/CLARIN-PL/Inforex

| | Brexit | | | | | | Asia Bibi | | | | | | Nord Stream | | | | | | Ryanair | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PL | CS | RU | BG | SL | UK | PL | CS | RU | BG | SL | UK | PL | CS | RU | BG | SL | UK | PL | CS | RU | BG | SL | UK |
| Documents | 500 | 284 | 153 | 600 | 52 | 50 | 88 | 89 | 118 | 101 | 4 | 6 | 151 | 161 | 150 | 130 | 74 | 40 | 146 | 163 | 150 | 87 | 52 | 63 |
| PER | 2 650 | 1 108 | 1 308 | 2 515 | 532 | 242 | 683 | 570 | 643 | 583 | 36 | 39 | 538 | 570 | 392 | 335 | 548 | 78 | 136 | 161 | 72 | 147 | 107 | 33 |
| LOC | 3 524 | 1 279 | 666 | 2 407 | 403 | 336 | 403 | 366 | 567 | 388 | 24 | 57 | 1 430 | 1 689 | 1 320 | 910 | 1 362 | 339 | 821 | 871 | 902 | 344 | 384 | 455 |
| ORG | 3 080 | 1 039 | 828 | 2 455 | 301 | 166 | 286 | 214 | 419 | 245 | 10 | 30 | 837 | 477 | 792 | 540 | 460 | 449 | 529 | 707 | 500 | 238 | 408 | 193 |
| EVT | 1 072 | 471 | 261 | 776 | 165 | 62 | 14 | 3 | 1 | 8 | 0 | 0 | 15 | 9 | 5 | 6 | 50 | 14 | 7 | 12 | 0 | 4 | 8 | 0 |
| PRO | 668 | 232 | 137 | 490 | 31 | 17 | 55 | 42 | 49 | 63 | 2 | 1 | 405 | 364 | 510 | 331 | 243 | 8 | 114 | 66 | 82 | 79 | 101 | 20 |
| Total | 10 994 | 4 129 | 3 200 | 8 643 | 1 445 | 823 | 1 441 | 1 195 | 1 679 | 1 287 | 72 | 127 | 3 225 | 3 116 | 3 020 | 2 122 | 2 664 | 948 | 1 607 | 1 817 | 1 556 | 812 | 1008 | 701 |
| *Distinct* | | | | | | | | | | | | | | | | | | | | | | | | |
| Surface forms | 2 820 | 1 111 | 783 | 1 200 | 596 | 234 | 508 | 303 | 406 | 412 | 51 | 87 | 845 | 770 | 892 | 504 | 902 | 336 | 514 | 475 | 400 | 323 | 673 | 187 |
| Lemmas | 2 133 | 840 | 568 | 1 091 | 411 | 177 | 412 | 248 | 317 | 360 | 41 | 77 | 634 | 550 | 583 | 448 | 600 | 244 | 419 | 400 | 332 | 315 | 520 | 137 |
| Entity IDs | 1 506 | 583 | 268 | 772 | 288 | 127 | 273 | 160 | 178 | 230 | 31 | 64 | 441 | 392 | 321 | 305 | 465 | 177 | 322 | 306 | 251 | 245 | 428 | 108 |

Table 1: Overview of the training datasets.

| | Covid-19 | | | | | | USA 2020 Elections | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PL | CS | RU | BG | SL | UK | PL | CS | RU | BG | SL | UK |
| Documents | 103 | 155 | 83 | 151 | 178 | 85 | 66 | 85 | 163 | 151 | 143 | 83 |
| PER | 419 | 478 | 559 | 351 | 834 | 215 | 566 | 447 | 3203 | 1539 | 2589 | 672 |
| LOC | 369 | 474 | 701 | 759 | 1228 | 364 | 827 | 277 | 3457 | 1093 | 1268 | 541 |
| ORG | 402 | 318 | 628 | 589 | 965 | 455 | 243 | 99 | 2486 | 557 | 578 | 384 |
| EVT | 240 | 393 | 435 | 465 | 612 | 269 | 86 | 63 | 396 | 170 | 118 | 257 |
| PRO | 137 | 155 | 400 | 168 | 274 | 143 | 87 | 56 | 846 | 240 | 254 | 124 |
| Total | 1567 | 1818 | 2723 | 2332 | 3913 | 1446 | 1810 | 942 | 10398 | 3599 | 4807 | 1978 |
| *Distinct* | | | | | | | | | | | | |
| Surface forms | 688 | 941 | 1436 | 1092 | 2190 | 622 | 484 | 377 | 3440 | 1117 | 1605 | 537 |
| Lemmas | 557 | 745 | 1133 | 1016 | 1774 | 509 | 356 | 279 | 2593 | 1019 | 1129 | 390 |
| Entity IDs | 404 | 562 | 796 | 764 | 1400 | 369 | 278 | 200 | 1669 | 668 | 833 | 270 |

Table 2: Overview of the test datasets.

- **Relaxed:** An entity mentioned in a given document is considered to be extracted correctly if the system response includes *at least one* annotation of a named mention of this entity (regardless of whether the extracted mention is in base form);

- **Strict:** The system response should include exactly one annotation *for each* unique form of a named mention of an entity in a given document, i.e., identifying all variants of an entity is required.

In relaxed evaluation we additionally distinguish between *exact* and *partial matching*: in the latter case, an entity mentioned in a given document is considered to be extracted correctly if the system response includes at least one partial match of a named mention of this entity.

We evaluate systems at several levels of granularity: we measure performance for (a) all NE types and all languages, (b) each given NE type and all languages, (c) all NE types for each language, and (d) each given NE type per language.

In the name normalization task, we take into account only correctly recognized entity mentions and only those that were normalized (on both the annotation and system's sides). Formally, let $N_{correct}$ denote the number of all correctly recognized entity mentions for which the system returned a correct base form. Let $N_{key}$ denote the number of all normalized entity mentions in the gold-standard answer key and $N_{response}$ denote the number of all normalized entity mentions in the system's response. We define precision and recall for the name normalization task as:

$$Recall = \frac{N_{corrrect}}{N_{key}} \qquad Precision = \frac{N_{corrrect}}{N_{response}}$$

In evaluating document-level, single-language and cross-lingual entity linking we adopted the Link-Based Entity-Aware metric (LEA) (Moosavi and Strube, 2016), which considers how important the entity is and how well it is resolved. LEA is defined as follows. Let $K = \{k_1, k_2, \ldots, k_{|K|}\}$ denote the set of key entities and $R = \{r_1, r_2, \ldots, r_{|R|}\}$ the set of response entities, i.e., $k_i \in K$ ($r_i \in R$) stand for set of mentions of the same entity in the key entity set (response entity set). LEA recall and precision are then defined as follows:

$$Recall_{LEA} = \frac{\sum_{k_i \in K} \left( imp(k_i) \cdot res(k_i) \right)}{\sum_{k_z \in K} imp(k_z)}$$

$$Precision_{LEA} = \frac{\sum_{r_i \in R} \left( imp(r_i) \cdot res(r_i) \right)}{\sum_{r_z \in R} imp(r_z)}$$

where $imp$ and $res$ denote the measure of importance and the resolution score for an entity, respectively. In our setting, we define $imp(e) = \log_2 |e|$ for an entity $e$ (in $K$ or $R$), $|e|$ is the number of mentions of $e$—i.e., the more mentions an entity has the more important it is. To avoid biasing the importance of the more frequent entities log is used. The resolution score of key entity $k_i$ is computed as the fraction of correctly resolved co-reference links of $k_i$:

$$res(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}$$

where $link(e) = (|e| \times (|e| - 1))/2$ is the number of unique co-reference links in $e$. For each $k_i$, LEA checks all response entities to check whether they are partial matches for $k_i$. Analogously, the resolution score of response entity $r_i$ is computed as the fraction of co-reference links in $r_i$ that are extracted correctly:

$$res(r_i) = \sum_{k_j \in K} \frac{link(r_i \cap k_j)}{link(r_i)}$$

LEA brings several benefits. For example, LEA considers resolved co-reference relations instead of resolved mentions and has more discriminative power than other metrics for co-reference resolution (Moosavi and Strube, 2016).

The evaluation was carried out in "case-insensitive" mode: all named mentions in system response and test corpora were lower-cased.

## 6 Participant Systems

Six teams submitted descriptions of their systems as BSNLP Workshop papers. We briefly review these systems here; for complete descriptions, please see the corresponding papers. Two additional teams submitted their results with short descriptions of their systems, which appear in this section.

The **UL FRI** system, (Prelevikj and Zitnik, 2021), generated results for several settings, models and languages, although the team's main motivation is to develop effective NER tools for Slovenian. The system uses contemporary BERT and RoBERTa multilingual pre-trained models, which include Slovene among other languages. The system was further trained on the SlavNER dataset for the NER task and used the Dedupe method for the Entity Matching task. The best performing models were pre-trained on Slovene. The results also indicate that two-step prediction of NE could be beneficial. The team made their code publicly available.

The **Priberam Labs** system, (Ferreira et al., 2021), focuses on the NER task. It uses three components: a multilingual contextual embedding model, a character-level embedding model, and a bi-affine classifier model. The paper reports results for different multilingual contextual embedding models, which included Multilingual BERT, XLM-RoBERTa, or the Slavic BERT. For different languages the best-performing models where different, but having the same language within the large pre-trained model usually improved the results—e.g., Slavic BERT, which used additional resources for Bulgarian, Russian and Polish, also performed best for these languages. The system uses heuristics to predict and resolve spans of NEs, and in this way it is able to tag overlapping entities. The code for the system is made available.

The **TLD** system, (Vīksna and Skadina, 2021), uses a staged approach. The first stage is identification of NEs in context, which is treated as a sequence labeling problem and is performed by a multilingual BERT model from Google, modified by the team. Entity linking is the second stage, which uses a list of LaBSE embeddings; matched entries need to pass a pre-defined threshold of cosine similarity with existing entries; otherwise they are added as new values to the list. The third stage is normalisation of identified entities, which is performed using models provided with Stanza.

The **L3i** system, (Cabrera-Diego et al., 2021), combines BERT models with the "Frustratingly Easy" domain adaptation algorithm. It also uses other techniques to improve system's NER performance, such as marking and enrichment of uppercase tokens, prediction of NE boundaries with a multitask approach, prediction of masked tokens, fine-tuning the language model to the domain of the document.

The **TraSpaS** system, (Suppa and Jariabka, 2021), tests the assumption that the universal open-source NLP toolkits (such as SpaCy, Stanza or Trankit) could achieve competitive performance on the Multilingual NER task, using large pre-trained Transformer-based language models available from HuggingfaceTransformers, which have not been available in previous editions of the Shared Task. The team tests the generalizability of the models to new low-resourced domains, and to languages such as Slovene and Ukrainian.

The **UWr-VL** system, (Rychlikowski et al., 2021), utilizes large collections of unstructured and structured documents for unsupervised training of embedding of lexical units and for recognizing and linking multiple real-world NEs. In particular, the team makes use of CommonCrawl news articles, Wikipedia, and its structured counterpart Wikidata as knowledge sources, to address the problem of data scarcity, building neural gazetteer via collecting different embeddings from these knowledge sources. The system further uses standard neural approaches to the NER task, with a RNN classifier, in order to determine for every input word the probability of labelling it with various beginning and end NE tags.

Two more systems generated the results for the shared task—**CTC-NER** from the Cognitive Technologies Center team, and **PAISC_wxd**:

**CTC-NER** is a baseline prototype of a NER component of an entity recognition system currently under development at the Cognitive Technologies Center. The system has a hybrid architecture combining rule-based and ML techniques; the ML-component is loosely related to (Antonova and Soloviev, 2013). The languages currently processed include Russian, English and Ukrainian.

**PAISC_wxd** uses the XLM-Roberta model, followed by BiLSTM-CRF on top. In addition, the system uses data enhancement based on machine translation.

## 7 Evaluation Results

Figure 3 shows the performance of the systems averaged across all languages and both test corpora. For each team that provided a solution for all six languages (7 teams except **CTC-NER**), we present the best scores ($F1$, $Precision$, and $Recall$) obtained by the team in three evaluation modes.[6]

As the plots show, the best performing model, Priberam, yields F-measure 85.7% according to the *relaxed partial* evaluation, and 79.3% according to the strict evaluation. The Priberam submission scores highest in precision — 89,4% relaxed partial, and 85.1% strict — but much lower in recall — 82.2% relaxed partial, and 74.3% strict.

Among the teams that submitted results for *cross-lingual entity linking*, only two achieved results comparable with the benchmarks achieved on the Second Challenge, and this year's results surpass those benchmarks by a substantial margin. The best results for each team, averaged across two corpora, are shown in Table 3. These results

[6]Complete results available on the Workshop's Web page: bsnlp.cs.helsinki.fi/final-rank-2021.pdf

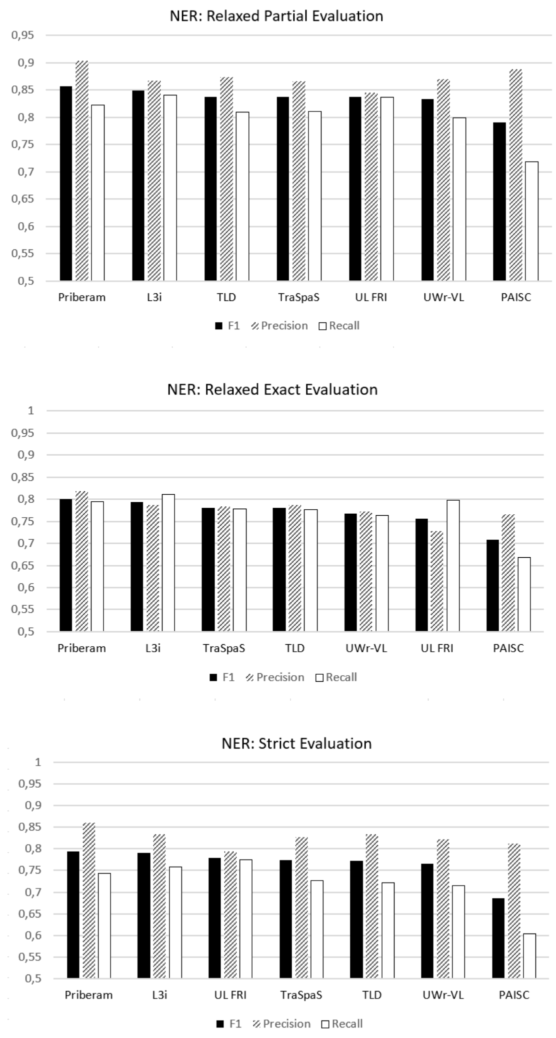


Figure 3: Best average performance scores obtained by the teams on the two test data

show that this task is much more difficult than entity extraction. The best performing model, TLD, achieves F-measure 50.4%.

Note that in our setting the performance on entity linking depends on the performance on name recognition and normalization: each system had to link entities that it had extracted from documents upstream, rather than link a set of correct entities.

Tables 4 and 5 present the F1-measures separated by language, for all tasks for the COVID-19 and USA 2020 ELECTIONS data sets These tables show only the top-performing model for each team. For recognition, we show only the *relaxed* evaluation, since the results obtained on the three evaluation schemes are correlated, as can be seen from Figure 3.

The tables indicate some variation in scores obtained on the test corpora This variation could be

| Covid-19 | | USA 2020 Elections | |
|---|---|---|---|
| System | F1 | System | F1 |
| TLD | 47.5 | TLD | 52.0 |
| UWr-VL | 32.8 | UWr-VL | 27.9 |
| Priberam | 5.8 | Priberam | 8.0 |
| L3i | 4.4 | TraSpaS | 7.9 |
| PAISC | 2.8 | L3i | 7.3 |
| TraSpaS | 2.7 | PAISC | 6.2 |
| UL FRI | 1.9 | CTC-NER | 2.9 |
| CTC-NER | 1.2 | UL FRI | 0.4 |

Table 3: Cross-lingual entity linking.

due to a number of factors, including actual differences in the test data, as well as differences in annotation across languages. This variation should and will be investigated in greater depth.

In Table 6 we present the results of the evaluation by entity type. As seen in the table, performance was higher overall for Loc and Per, and substantially lower for Org and Pro, which corresponds with our findings from the previous editions of the shared task, where Org and Misc were the most problematic categories (Piskorski et al., 2017). The Pro category also exhibits higher variance across languages and corpora than other categories, which might point to possible annotation artefacts. The results for the Evt category are less informative, since the task heavily depends on detecting the repeated central events of the corpora.

## 8 Conclusion

This paper reports on the $3^{rd}$ Multilingual Named Entity Challenge focusing on recognizing mentions of NEs in Web documents in six Slavic languages, normalization of the NEs, and cross-lingual entity linking. The Challenge has attracted substantial interest, following the prior Challenges in 2017 and 2019, with 10 teams registering for the competition and eight teams submitting results from working systems, with multiple variants. Most systems use state-of-the-art neural network models. Overall, the results of the best-performing systems are quite strong for extraction and normalization, while cross-lingual linking is the most challenging of the tasks.

We show summary results for the main aspects of the challenge and the best-performing model for each team. For detailed, in-depth evaluations of all participating systems and their performance please consult the Shared Task's Web page and the papers by the respective teams.

To stimulate further research into NLP for Slavic languages, including cross-lingual entity linking, our training and test datasets, the detailed annotations, and scripts used for evaluations are made available to the public on the Shared Task's Web page.[7] The annotation interface is released by the Inforex team, to support further annotation of additional data for future tests.

This challenge covered six Slavic languages. For future editions of the Challenge, we plan to expand the data sets, covering a wider range of entity types, and supporting cross-lingual entity linking. We plan to expand the training and test data to include *non-Slavic* languages. We will also undertake further refinement of the underlying annotation guidelines—a highly complex task in a real-world setting. More complex phenomena also need to be addressed, e.g., coordinated NEs, contracted versions of multiple NEs, etc.

We believe that the reported results and the annotated datasets will help stimulate further research on robust, end-to-end analysis of real-world texts in Slavic languages.

## Acknowledgments

Work on Bulgarian was in part supported by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies for the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO1-377/18.12.2020.

Work on Czech was in part supported by ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no. CZ.02.1.01/0.0/0.0/17 048/0007267), and by Grant No. SGS-2019-018 "Processing of heterogeneous data and its specialized applications."

Work on Inforex and on Polish was supported in part by investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

We thank the students of Pushkin State Russian Language Institute for their assistance with annotation of Russian data.

This work has been partially supported by the European Union Horizon 2020 research and in-

[7] bsnlp.cs.helsinki.fi/shared_task.html

| COVID-19 | | *Language* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phase* | *Metric* | bg | | cs | | pl | | ru | | sl | | uk | |
| Recognition | Relaxed Partial | Priberam | 83.2 | UWr-VL | 86.7 | Priberam | 87.8 | L3i | 76.0 | UWr-VL | 87.6 | UWr-VL | 84.8 |
| | | L3i | 82.8 | Priberam | 86.3 | UWr-VL | 86.9 | Priberam | 75.1 | Priberam | 87.5 | L3i | 80.6 |
| | | TLD | 82.2 | TLD | 84.1 | TLD | 86.4 | PAISC | 74.4 | L3i | 85.6 | TLD | 80.1 |
| | | UL FRI | 81.6 | L3i | 83.9 | L3i | 85.0 | TLD | 72.9 | TLD | 84.2 | Priberam | 79.9 |
| | | UWr-VL | 81.2 | TraSpaS | 82.0 | UL FRI | 83.4 | UL FRI | 71.9 | TraSpaS | 83.9 | PAISC | 78.3 |
| | | TraSpaS | 80.9 | UL FRI | 80.4 | TraSpaS | 82.5 | TraSpaS | 70.2 | PAISC | 80.1 | UL FRI | 78.3 |
| | | PAISC | 79.7 | PAISC | 77.6 | PAISC | 81.0 | CTC-NER | 69.3 | UL FRI | 79.1 | TraSpaS | 78.1 |
| | | | | | | | | UWr-VL | 67.1 | | | CTC-NER | 65.0 |
| Normalization | | UWr-VL | 33.3 | TraSpaS | 47.0 | UWr-VL | 57.4 | CTC-NER | 40.4 | UWr-VL | 53.0 | TraSpaS | 53.7 |
| | | UL FRI | 21.4 | TLD | 45.2 | UL FRI | 47.2 | UL FRI | 39.9 | UL FRI | 40.5 | UWr-VL | 51.5 |
| | | TLD | 13.8 | UWr-VL | 44.8 | TraSpaS | 46.2 | TraSpaS | 38.6 | TraSpaS | 34.3 | UL FRI | 50.7 |
| | | TraSpaS | 10.0 | UL FRI | 44.4 | TLD | 45.3 | TLD | 36.2 | TLD | 32.3 | TLD | 46.3 |
| | | Priberam | 0.0 | Priberam | 0.0 | Priberam | 0.0 | UWr-VL | 27.2 | Priberam | 0.0 | CTC-NER | 39.2 |
| | | L3i | 0.0 | L3i | 0.0 | L3i | 0.0 | Priberam | 0.0 | L3i | 0.0 | Priberam | 0.0 |
| | | PAISC | 0.0 | PAISC | 0.0 | PAISC | 0.0 | L3i | 0.0 | PAISC | 0.0 | L3i | 0.0 |
| | | | | | | | | PAISC | 0.0 | | | PAISC | 0.0 |
| Entity linking | Document level | UWr-VL | 37.6 | TLD | 47.0 | UWr-VL | 61.2 | TLD | 42.5 | UWr-VL | 52.0 | TLD | 48.9 |
| | | TLD | 24.6 | UWr-VL | 46.0 | TLD | 44.7 | UWr-VL | 30.5 | TLD | 45.2 | UWr-VL | 45.3 |
| | | L3i | 13.3 | UL FRI | 29.8 | UL FRI | 26.4 | UL FRI | 20.4 | UL FRI | 29.6 | UL FRI | 24.7 |
| | | Priberam | 12.4 | Priberam | 23.9 | PAISC | 20.4 | Priberam | 15.5 | Priberam | 16.8 | Priberam | 23.7 |
| | | TraSpaS | 11.5 | L3i | 22.5 | L3i | 20.3 | PAISC | 13.8 | L3i | 15.6 | L3i | 22.3 |
| | | PAISC | 11.4 | TraSpaS | 22.1 | Priberam | 20.0 | L3i | 13.3 | TraSpaS | 14.9 | TraSpaS | 22.0 |
| | | UL FRI | 6.1 | PAISC | 21.2 | TraSpaS | 18.4 | TraSpaS | 12.2 | PAISC | 13.8 | PAISC | 17.8 |
| | | | | | | | | CTC-NER | 3.5 | | | CTC-NER | 2.3 |
| | Single language | UWr-VL | 67.9 | TLD | 66.5 | UWr-VL | 73.0 | TLD | 47.4 | UWr-VL | 66.4 | TLD | 61.7 |
| | | TLD | 57.1 | UWr-VL | 66.1 | TLD | 67.8 | UWr-VL | 38.9 | TLD | 59.2 | UWr-VL | 61.5 |
| | | PAISC | 16.4 | UL FRI | 40.2 | UL FRI | 38.8 | UL FRI | 20.1 | UL FRI | 32.7 | UL FRI | 36.8 |
| | | L3i | 10.9 | PAISC | 15.9 | PAISC | 13.7 | Priberam | 6.2 | TraSpaS | 10.0 | Priberam | 15.9 |
| | | Priberam | 8.7 | L3i | 11.2 | Priberam | 9.3 | L3i | 4.2 | Priberam | 7.2 | L3i | 7.7 |
| | | UL FRI | 7.6 | TraSpaS | 11.2 | TraSpaS | 8.2 | PAISC | 3.5 | L3i | 4.2 | PAISC | 7.5 |
| | | TraSpaS | 3.6 | Priberam | 8.0 | L3i | 7.9 | TraSpaS | 2.0 | PAISC | 1.8 | TraSpaS | 6.3 |
| | | | | | | | | CTC-NER | 1.8 | | | CTC-NER | 2.6 |

Table 4: F1-measure results for the COVID-19 corpus.

novation programme under grants 770299 (NewsEye).

Work on Slovene was financed through the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 825153, Project EMBEDDIA: Cross-Lingual Embeddings for Less-Represented Languages in European News Media, as well as Slovenian Research Agency's project: Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581).

## References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1724–1727, Reykjavík, Iceland.

AY Antonova and AN Soloviev. 2013. Conditional random field models for the processing of Russian. In *Computational Linguistics and Intellectual Technologies: Papers From the Annual Conference "Dialogue"(Bekasovo, 29 May–2 June 2013)*, volume 1, pages 27–44.

Krešimir Baksa, Dino Golović, Goran Glavaš, and Jan Šnajder. 2017. Tagging named entities in Croatian tweets. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 4(1):20–41.

Luis Adrián Cabrera-Diego, Jose G. Moreno, and Antoine Doucet. 2021. Using a frustratingly easy domain and tagset adaptation for creating slavic named entity recognition systems. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Nancy Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) program—tasks, data, and evaluation. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal.

Xiangyu Duan, Rafael E. Banchs, Min Zhang, Haizhou Li, and A. Kumaran. 2016. Report of NEWS 2016 machine transliteration shared task. In *Proceedings of The Sixth Named Entities Workshop*, pages 58–72, Berlin, Germany.

Pedro Ferreira, Ruben Cardoso, and Afonso Mendes. 2021. Priberam labs at the 3rd shared task on

| USA 2020 ELECTIONS | | *Language* | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Phase* | *Metric* | bg | | cs | | pl | | ru | | sl | | uk | |
| Recognition | Relaxed Partial | L3i | 89.8 | UWr-VL | 91.3 | Priberam | 92.3 | L3i | 83.7 | Priberam | 91.5 | TLD | 84.6 |
| | | Priberam | 88.7 | Priberam | 90.7 | L3i | 92.0 | Priberam | 83.4 | L3i | 91.5 | Priberam | 84.6 |
| | | TraSpaS | 88.1 | L3i | 90.2 | TLD | 90.8 | TraSpaS | 81.5 | UWr-VL | 90.4 | L3i | 84.5 |
| | | UWr-VL | 87.3 | TLD | 88.5 | UWr-VL | 89.8 | TLD | 80.9 | TLD | 89.8 | TraSpaS | 83.3 |
| | | TLD | 87.3 | UL FRI | 88.4 | TraSpaS | 89.2 | UL FRI | 80.5 | TraSpaS | 89.4 | UWr-VL | 83.3 |
| | | UL FRI | 86.9 | TraSpaS | 87.8 | UL FRI | 89.1 | UWr-VL | 77.2 | UL FRI | 88.6 | UL FRI | 83.2 |
| | | PAISC | 83.6 | PAISC | 82.6 | PAISC | 66.4 | PAISC | 77.1 | PAISC | 86.0 | PAISC | 77.0 |
| | | | | | | | | CTC-NER | 75.4 | | | CTC-NER | 71.1 |
| Normalization | | UWr-VL | 51.3 | UWr-VL | 51.9 | UWr-VL | 62.1 | TraSpaS | 50.7 | UWr-VL | 62.4 | UL FRI | 56.9 |
| | | UL FRI | 21.9 | TraSpaS | 42.0 | TLD | 51.0 | UL FRI | 48.8 | UL FRI | 43.9 | TraSpaS | 56.8 |
| | | TLD | 19.1 | TLD | 40.1 | UL FRI | 50.1 | TLD | 46.5 | TraSpaS | 34.2 | TLD | 55.3 |
| | | TraSpaS | 17.9 | UL FRI | 39.7 | TraSpaS | 42.4 | CTC-NER | 44.8 | TLD | 31.9 | CTC-NER | 36.9 |
| | | Priberam | 0.0 | Priberam | 0.0 | Priberam | 0.0 | UWr-VL | 25.6 | Priberam | 0.0 | UWr-VL | 26.5 |
| | | L3i | 0.0 | L3i | 0.0 | L3i | 0.0 | Priberam | 0.0 | L3i | 0.0 | Priberam | 0.0 |
| | | PAISC | 0.0 | PAISC | 0.0 | PAISC | 0.0 | L3i | 0.0 | PAISC | 0.0 | L3i | 0.0 |
| | | | | | | | | PAISC | 0.0 | | | PAISC | 0.0 |
| Entity linking | Document level | UWr-VL | 63.7 | UWr-VL | 64.3 | UWr-VL | 67.1 | TLD | 44.8 | UWr-VL | 67.3 | UWr-VL | 58.9 |
| | | TLD | 58.7 | TLD | 55.3 | TLD | 62.3 | UWr-VL | 35.8 | TLD | 59.3 | TLD | 52.2 |
| | | Priberam | 12.5 | UL FRI | 37.5 | UL FRI | 44.9 | UL FRI | 32.2 | UL FRI | 43.3 | UL FRI | 28.8 |
| | | L3i | 12.1 | L3i | 30.5 | Priberam | 18.2 | Priberam | 12.3 | L3i | 18.3 | Priberam | 25.4 |
| | | TraSpaS | 11.7 | Priberam | 29.5 | L3i | 18.0 | L3i | 12.3 | Priberam | 17.9 | L3i | 23.9 |
| | | PAISC | 11.4 | TraSpaS | 28.6 | TraSpaS | 17.4 | PAISC | 9.9 | TraSpaS | 17.1 | TraSpaS | 23.5 |
| | | UL FRI | 4.5 | PAISC | 21.6 | PAISC | 13.4 | TraSpaS | 9.8 | PAISC | 15.8 | PAISC | 16.8 |
| | | | | | | | | CTC-NER | 2.8 | | | CTC-NER | 1.5 |
| | Single language | UWr-VL | 68.5 | TLD | 69.0 | TLD | 74.9 | TLD | 50.1 | TLD | 68.7 | TLD | 62.2 |
| | | TLD | 67.1 | UWr-VL | 66.0 | UWr-VL | 69.9 | UWr-VL | 39.3 | UWr-VL | 66.5 | UWr-VL | 52.9 |
| | | PAISC | 12.8 | UL FRI | 50.0 | UL FRI | 37.7 | UL FRI | 13.6 | UL FRI | 21.3 | UL FRI | 23.0 |
| | | Priberam | 10.1 | L3i | 18.1 | Priberam | 14.8 | Priberam | 5.6 | Priberam | 8.4 | TraSpaS | 21.4 |
| | | TraSpaS | 8.6 | Priberam | 17.7 | L3i | 14.5 | L3i | 5.5 | L3i | 8.3 | L3i | 20.5 |
| | | L3i | 8.6 | TraSpaS | 17.7 | TraSpaS | 13.4 | TraSpaS | 5.1 | TraSpaS | 8.2 | Priberam | 20.2 |
| | | UL FRI | 8.3 | PAISC | 14.1 | PAISC | 10.7 | PAISC | 4.4 | PAISC | 7.2 | PAISC | 12.9 |
| | | | | | | | | CTC-NER | 3.6 | | | CTC-NER | 9.4 |

Table 5: Evaluation results (F1-measure) for the USA 2020 ELECTION corpus.

| | COVID-19 | | | | | | USA 2020 ELECTIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bg | cs | pl | ru | sl | uk | bg | cs | pl | ru | sl | uk |
| **Per** | 98.0 | 98.1 | 98.3 | 83.1 | 98.2 | 96.6 | 93.6 | 97.4 | 94.2 | 93.1 | 96.3 | 98.7 |
| **Loc** | 95.8 | 96.4 | 96.7 | 95.1 | 95.7 | 97.3 | 97.5 | 96.9 | 97.6 | 93.1 | 98.2 | 93.8 |
| **Org** | 86.5 | 89.4 | 91.3 | 82.9 | 88.8 | 87.6 | 86.6 | 89.6 | 86.3 | 76.6 | 76.7 | 81.7 |
| **Pro** | 55.1 | 76.2 | 75.6 | 47.6 | 63.4 | 49.4 | 80.7 | 87.4 | 90.2 | 66.9 | 77.7 | 69.9 |
| **Evt** | 52.6 | 40.1 | 57.8 | 52.6 | 63.5 | 75.9 | 29.6 | 26.1 | 40.5 | 55.7 | 38.0 | 16.1 |

Table 6: Recognition F1-measure (relaxed partial) by entity type—best-performing systems for each language.


slavner. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Diversity of scenarios in information extraction. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 entity discovery and linking tasks. In *Proceedings of Text Analysis Conference (TAC2014)*, pages 1333–1339.

Heng Ji, Joel Nothman, and Ben Hachey. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proceedings of Text Analysis Conference (TAC2015)*.

Mladen Karan, Goran Glavaš, Frane Šarić, Jan Šnajder, Jure Mijić, Artur Šilić, and Bojana Dalbelo Bašić. 2013. CroNER: Recognizing named entities in Croatian using conditional random fields. *Informatica*, 37(2):165.

Michal Konkol and Miloslav Konopík. 2013. CRF-based Czech named entity recognizer and consolidation of Czech NER research. In *Text, Speech and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 153–160. Springer Berlin Heidelberg.

A Kumaran, Mitesh M. Khapra, and Haizhou Li. 2010. Report of NEWS 2010 transliteration mining shared task. In *Proceedings of the 2010 Named Entities Workshop*, pages 21–28, Uppsala, Sweden.

Nikola Ljubešić, Marija Stupar, Tereza Jurić, and Željko Agić. 2013. Combining available datasets for building named entity recognition models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):35–57.

Michał Marcińczuk, Marcin Oleksy, and Jan Kocoń. 2017. Inforex - a collaborative system for text corpora annotation and analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2-8, 2017*, pages 473–482. INCOMA Ltd.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 632–642, Berlin, Germany.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2018. *Proceedings of the PolEval 2018 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2020. *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 63–74, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

Jakub Piskorski, Karol Wieloch, and Marcin Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299.

Marko Prelevikj and Slavko Zitnik. 2021. Bsnlp 2021 shared task: Multilingual named entity recognition and matching using bert and dedupe for slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Adam Przepiórkowski. 2007. Slavonic information extraction and partial parsing. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paweł Rychlikowski, Adrian Lancucki, Adam Kaczmarek, Bartłomiej Najdecki, Adam Wawrzyński, and Wojciech Janowski. 2021. Named entity recognition and linking augmented with large-scale structured data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Agata Savary and Jakub Piskorski. 2011. Language Resources for Named Entity Annotation in the National Corpus of Polish. *Control and Cybernetics*, 40(2):361–391.

Magda Ševčíková, Zdeněk Žabokrtskỳ, and Oldřich Kruza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195. Springer.

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81.

A. S. Starostin, V. V. Bocharov, S. V. Alexeeva, A. A. Bodrova, A. S. Chuchunkov, S. S. Dzhumaev, I. V. Efimenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, and S. Y. Toldova. 2016. FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue"*, pages 688–705.

Marek Suppa and Ondrej Jariabka. 2021. Benchmarking pre-trained language models for multilingual ner: Traspas at the bsnlp2021 shared task. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Erik Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rinalds Vīksna and Inguna Skadina. 2021. Multilingual slavic named entity recognition. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. European Association for Computational Linguistics.

Jakub Waszczuk, Katarzyna Głowińska, Agata Savary, and Adam Przepiórkowski. 2010. Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10)*, pages 531–539, Wisła, Poland. PTI.