Slav-NER: the 4$^{\text{th}}$ Cross-lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic languages

Roman Yangarber[1], Jakub Piskorski[2], Anna Dmitrieva[1],
Michał Marcińczuk[3], Pavel Přibáň[4], Piotr Rybak[2], Josef Steinberger[4]
[1]University of Helsinki, Finland `first.last@helsinki.fi`
[2]Polish Academy of Sciences, Warsaw, Poland `jpiskorski@gmail.com`
[3]Wrocław University of Science and Technology, Poland `marcinczuk@gmail.com`
[4]University of West Bohemia, Czech Republic `{pribanp,jstein}@kiv.zcu.cz`

Abstract

This paper describes Slav-NER: the 4$^{\text{th}}$ 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. This version of the Challenge covers three languages and five entity types. It is organized as part of the 9th Slavic Natural Language Processing Workshop, co-located with the EACL 2023 Conference.

Seven teams registered and three participated actively in the competition. Performance for the named entity recognition and normalization tasks reached 90% F_1 measure, much higher than reported in the first edition of the Challenge, but similar to the results reported in the latest edition. Performance for the entity linking task for individual language reached the range of 72-80% F_1 measure. 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 4th 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 9th Slav-NLP: Workshop on Natural Language Processing in Slavic languages, co-located with the EACL 2023 conference. The task covers three languages—Czech, Polish and Russian—and five types of NE: person, location, organization, product, and event. The data consists of documents collected from the Web involving certain "focal" events. 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 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. Conclusions and lessons learned are in Section 8.

2 Prior Work

The work described here builds on a series of Shared Tasks on Multilingual Named Entity Recognition, Normalization and cross-lingual Matching for Slavic Languages, (Piskorski et al., 2017, 2019, 2021), 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 con-

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

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 179–189
May 6, 2023 ©2023 Association for Computational Linguistics

text 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 alphabets/writing systems. A series of NE Transliteration Shared Tasks were organized as 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, in languages including English, Chinese, Tamil, Russian, and Arabic.

Research on NE focusing on Slavic languages includes 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), NE recognition in Slovene (Štajner et al., 2013; Ljubešić et al., 2013), a Czech corpus of 11K annotated NEs (Ševčíková et al., 2007), NER 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] (Ogrodniczuk and Łukasz Kobyliński, 2018, 2020) and LESZCZE[3] campaigns, recent shared tasks on NE Recognition in Russian (Starostin et al., 2016; Artemova et al., 2022), the latter utilizing the NEREL dataset (a Russian dataset for named entity recognition and relation extraction, described in Loukachevitch et al., 2021), and *SemEval 2022 Task 11: MultiCoNER Multilingual Complex Named Entity Recognition*[4] and *SemEval 2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition*,[5] which included Russian and Ukrainian respectively.

3 Task Description

The data for this edition of the shared task consists of a set of documents in three Slavic languages: Czech, Polish and Russian. To facilitate entity linking, the set of documents is chosen to involve one specific event. The documents were obtained from the Web, by posing keyword queries to search engines, 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 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

[2] http://poleval.pl

[3] https:/lepiszcze.ml/tasks/namedentityrecognition

[4] https://multiconer.github.io/multiconer_1

[5] https://multiconer.github.io

same real-world entity should be assigned the same ID—referred to as the cross-lingual ID.

These tasks do 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.[6] 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 "*UEFA*" and "*uefa*" (provided that they refer to the same NE type[7]), only one annotation should be produced. The recognition of common-noun or pronominal references to named entities is not included as 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 Volodymyr Zelenskiy. . .* ", only "*Volodymyr Zelenskiy*" 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 have no formal organization unifying them) should also be recognized, e.g., "*Ukrainians*." In this context, mentions of a single member belonging to such groups, e.g., "*Ukrainian*," should be assigned the same cross-lingual ID as plural mentions, i.e., "*Ukrainians*" and "*Ukrainian*" 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 "*Królewscy wygrali*" (The Royals won), "*Królewscy*" is to be tagged as PER.

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 "*Szpital Miejski im. Franciszka Raszei zatrudnił nowy personel ze względu na pandemie koronawirusa*" (The Franciszek Raszeia Hospital hired new staff due to the covid pandemics.) the mention "*Szpital Miejski im. Franciszka Raszei*" 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 "*...Narodowy Fundusz Zdrowia w Poznaniu...*" (National Health Fund in Poznań), the full phrase "*Narodowy Fundusz Zdrowia w Poznaniu*" should be extracted.

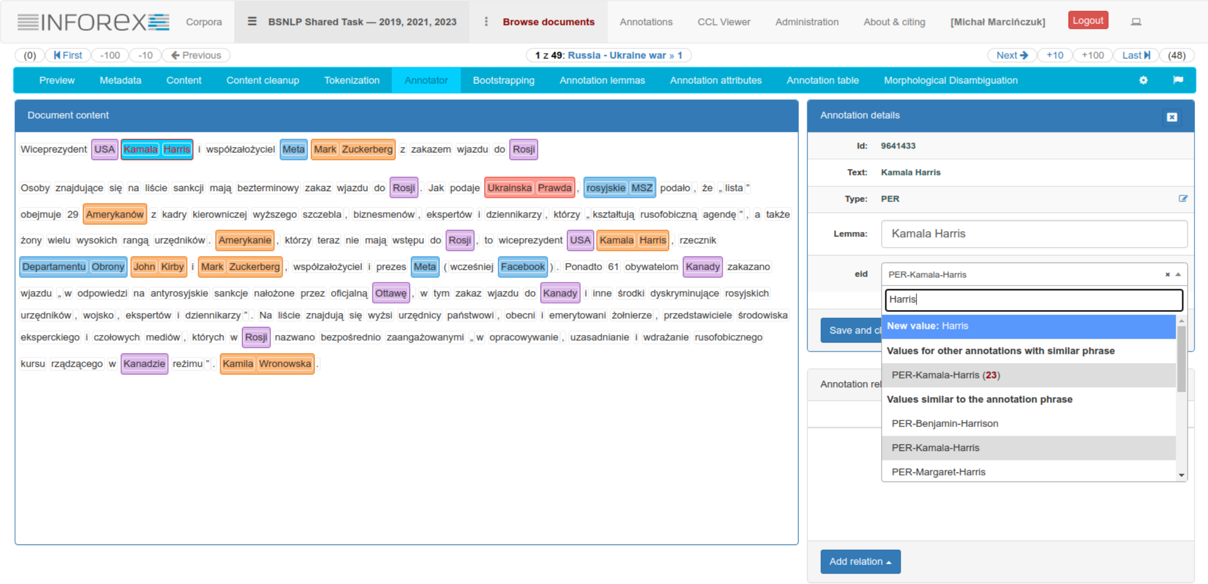
Products (PRO): All names of *products and services*, such as electronics ("*Samsung Galaxy A41*"), cars ("*Subaru Ascent*"), newspapers ("*Politico*"), web-services ("*The Telegraph*"), medicines ("*Oxycodone*"), awards ("*Nobel Prize*"), books ("*Hamlet*"), TV programmes ("*TVN News*"), etc.

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 Maastricht*" (Polish: "Maastricht Agreement") and initiatives, e.g., "*Horizon 2020*".

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

[7] Union of European Football Associations.



Wiceprezydent USA Kamala Harris i współzałożyciel Meta Mark Zuckerberg z zakazem wjazdu do Rosji

Osoby znajdujące się na liście sankcji mają bezterminowy zakaz wjazdu do Rosji . Jak podaje Ukrainska Prawda , rosyjskie MSZ podało , że „ lista ” obejmuje 29 Amerykanów z kadry kierowniczej wyższego szczebla , biznesmenów , ekspertów i dziennikarzy , którzy „ kształtują rusofobiczną agendę ” , a także żony wielu wysokich rangą urzędników . Amerykanie , którzy teraz nie mają wstępu do Rosji , to wiceprezydent USA Kamala Harris , rzecznik Departamentu Obrony John Kirby i Mark Zuckerberg , współzałożyciel i prezes Meta (wcześniej Facebook) . Ponadto 61 obywatelom Kanady zakazano wjazdu „ w odpowiedzi na antyrosyjskie sankcje nałożone przez oficjalną Ottawę , w tym zakaz wjazdu do Kanady i inne środki dyskryminujące rosyjskich urzędników , wojsko , ekspertów i dziennikarzy ” . Na liście znajdują się wyżsi urzędnicy państwowi , obecni i emerytowani żołnierze , przedstawiciele środowiska eksperckiego i czołowych mediów , których w Rosji nazwano bezpośrednio zaangażowanymi „ w opracowywanie , uzasadnianie i wdrażanie rusofobicznego kursu rządzącego w Kanadzie reżimu ” . Kamila Wronowska .

Id: 9641433
Text: Kamala Harris
Type: PER
Lemma: Kamala Harris
eid: PER-Kamala-Harris
Harris
New value: Harris
Values for other annotations with similar phrase
PER-Kamala-Harris (23)
Values similar to the annotation phrase
PER-Benjamin-Harrison
PER-Kamala-Harris
PER-Margaret-Harris

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

Input:

Za 120 dní 10 tisíc vojáků. Johnson nabídl v Kjyevě pomoc při výcviku armády Britský premiér Boris Johnson v pátek znovu přijel do ukrajinského Kyjeva, kde se sešel s prezidentem Volodymyrem Zelenským a představil mu konkrétní nabídku britské pomoci s výcvikem ukrajinských vojáků. Oba představitelé spolu také hovořili o dodávkách těžkých zbraní a protileteckých systémů, stejně jako o ekonomické podpoře Ukrajiny, která od konce února čelí ruské agresi, i o dalších možnostech zpřísnění protiruských sankcí.

Output:

Boris Johnson	Boris Johnson	PER	PER-Boris-Johnson
Johnson	Johnson	PER	PER-Boris-Johnson
Kyjeva	Kyjev	LOC	GPE-Kiev
Kyjevě	Kyjev	LOC	GPE-Kiev
Ukrajiny	Ukrajina	LOC	GPE-Ukraine
Volodymyrem Zelenským	Volodymyr Zelensky	PER	PER-Volodymyr-Zelensky
ruské agresi	ruská agrese	EVT	EVT-2022-Russian-Invasion-of-Ukraine

Figure 2: Example input and output formats.

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 lotnicza w Gibraltarze*” (Polish: “1943 Gibraltar Liberator AL523 crash”), outbreaks of infectious diseases (“*Spanish Flu*”). Future, speculative, and fictive events—e.g., “‘Czexit”—are considered event mentions.

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 “*Uniwersytet Adama Mickiewicza*” (Polish: “Adam Mickiewicz University”) one should not extract “*Adama Mickiewicza*”, but only the top-level entity.

In case one word-form (e.g., “*Washington*”) 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 “*Dutch and Belgian Parliament*,” two names should be extracted (as ORG). The lemmas would be “*Dutch*” and “*Belgian Parliament*”, and the IDs should refer to “*Dutch Parliament*” and “*Belgian Parlia-*

ment" 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.[8]

4 Data

In this edition of the Challenge the annotated datasets from previous editions were used as training data. In particular, the training and test datasets annotated in Bulgarian, Czech, Polish and Russian from 2019 Shared Task (Piskorski et al., 2019) and training and test datasets annotated in Bulgarian, Czech, Polish, Russian, Slovene and Ukrainian from 2021 Shared Task (Piskorski et al., 2021) were used. The prior datasets annotated in six languages covered various major topics, including, i.a., the COVID-19 pandemic, the 2020 USA Presidential elections (USA 2020 ELECTIONS), 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. The test data for the current edition of the challenge involves the RUSSIA-UKRAINE WAR.

Each of the datasets, including the latest test data, was created as follows. For the focus entity/event, 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 when 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[9] 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 *multiple phases* of automatic and manual consistency checks, to reduce annotation errors, especially in entity linking.

The training data statistics are shown in Table 1 and 2—for 2019 and 2021 datasets, respectively, while the test data statistics are shown in Table 3.

The participants received the test dataset—

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

[9] `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 dataset from the 2019 edition of the Slavic NER challenge.

	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 training dataset from the 2021 edition of the Slavic NER challenge.

	Russia-Ukraine war		
	PL	CS	RU
Documents	50	50	52
PER	276	229	236
LOC	599	345	454
ORG	252	159	355
EVT	62	49	15
PRO	85	43	78
Total	1274	825	1138
Distinct			
surface forms	723	498	725
Lemmas	563	384	594
Entity IDs	410	280	493

Table 3: Overview of the test dataset for the 2023 edition of the Slavic NER challenge.

Russia-Ukraine war—and were given circa 2 days to return up to 10 system responses. The topic was not announced in advance, and the annotations were not released. The rationale behind this decision was to motivate the participants to build a general solution for Slavic NER, rather than to optimize their models toward particular scenarios or sets 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 F_1 measure. For NER, two types of evaluations were carried out:

- **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 the 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 the systems at several levels of granularity: we measure the performance (a) for all NE types and all languages, (b) for each given NE type and all languages, (c) for all NE types for each language, and (d) for 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 the response 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 (LEA) metric (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 a set of mentions of the same entity in the key entity set—the response entity set. LEA recall and precision are then defined as follows:

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

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

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 $\log_2$ 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

Out of the seven registered teams, we received results from three. Further, two of these teams provided papers describing the details of their systems, presented in the 2023 Slavic NLP Workshop. We briefly review these systems here; for complete descriptions, please see the corresponding papers.

The **Tilde** system, (Rinalds Vīksna and Rozis, 2023), utilizes the multilingual XLM-R model to perform all subtasks. They enhance their training set by incorporating diverse NER datasets, in addition to the Slavic NER Challenge training set. The authors fine-tune five different variants of the XLM-R Large (Conneau et al., 2020) model that differ in the approach for the entity-linking subtask. For each variant, they use slightly different training datasets. In addition, one of the variants is an ensemble of five XLM-R Base models, one for each of the five NER entity labels. The base model was initially pre-trained on 2.6 GB of recent Czech, Polish and Russian news articles to integrate into the model new entities and events, which have emerged since the original model was trained. This process enables the model to embed the latest information and keep up-to-date with the evolving language usage.

The **AMU** system (Pałka and Nowakowski, 2023) combines a set of transformer-based models for named entity recognition, categorization, and lemmatization. They evaluated several monolingual (HerBERT, Czert, and RuBERT) and multilingual (Slavic-BERT and XLM-RoBERTa) BERT-like models for entity recognition and categorization. For entity lemmatization, sequence-to-sequence (seq2seq) models were applied, plT5 and mT5. The pre-trained models were fine-tuned on the dataset provided within the shared task and additional external resources, including datasets annotated with named entities: Collection3, MultiNERD, Polyglot-NER, WikiNEuRal; dictionaries of lemmatized named entities and multi-word expressions: SEJF, SEJFEK, PolEval 2019 Task 2. The additional resources for lemmatization were only for Polish. Thus, the authors used OPUS-MT to translate the resources to other languages to overcome the language limitation.

The third team—CTC, Cognitive Technologies Center—submitted results, but did not provide a description paper; their approach was similar to the one employed by this team in the 2021 Edition of the Shared Task (Piskorski et al., 2021).

7 Evaluation Results

Table 4 presents the F_1-measures separated by language, for all tasks for the test data—the "Russia-Ukraine war" dataset. The table shows only the one top-performing model for each team. The CTC team submitted results only for the Russian language. The best-performing team overall is the one that submitted the Tilde system based on the multilingual Transformer-based XLM-R model. The results of the AMU system are almost on par, trailing by only a small margin in most of the evaluated metrics, with the exception of the normalization task. The CTC system lags behind other systems by a margin of 4% F_1-measure in the recognition subtask.

Only the Tilde team submitted results for *cross-lingual entity linking*, achieving 66.9% F_1 score. This is a great improvement compared to the Third Challenge, where the best results were around 50% of F_1 score. To date, the task of cross-lingual linking remains much more challenging than the task of entity extraction.

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

In Table 5 we present the results of the evaluation by entity type. As seen in the table, performance was higher overall for LOC, PER and PRO in the case of Czech. Substantially lower results were achieved for ORG and EVT in all languages and PRO in Polish and Russian, which corresponds with our findings from the previous editions of the shared task, where ORG, PRO and EVT were the most challenging categories (Piskorski et al., 2017, 2021). The results for the EVT category are less informative since the task heavily depends on detecting the repeated central events of the corpora.

		Language					
Phase	*Metric*	cs		pl		ru	
Recognition	*Relaxed Partial*	Tilde	91.6	Tilde	89.9	Tilde	89.8
		AMU	91.5	AMU	88.9	AMU	88.8
						CTC	84.4
	Relaxed Exact	Tilde	89.0	Tilde	86.0	Tilde	85.1
		AMU	88.3	AMU	84.1	AMU	85.0
						CTC	81.0
	Strict	Tilde	89.9	Tilde	87.0	Tilde	86.8
		AMU	89.7	AMU	85.4	AMU	86.2
						CTC	73.4
Normalization		AMU	76.9	AMU	82.4	AMU	81.5
		Tilde	54.3	Tilde	53.9	Tilde	72.6
						CTC	66.0
Entity Linking	*Document level*	Tilde	80.2	Tilde	76.4	Tilde	71.7
		AMU	25.8	AMU	19.7	AMU	19.4
						CTC	4.8
	Single language	Tilde	77.6	Tilde	72.9	Tilde	61.0
		AMU	7.5	AMU	8.8	AMU	5.8
						CTC	2.9

Table 4: F_1-measure results for the test dataset.

	Language		
Entity Class	cs	pl	ru
Per	99.6	97.9	98.0
Loc	94.7	94.6	96.5
Org	88.8	83.3	87.2
Pro	93.3	89.4	71.2
Evt	42.0	49.9	28.6

Table 5: Recognition F_1-measure (relaxed partial) by entity type—best-performing systems for each language.

8 Conclusion

This paper reports on the 4^{th} Multilingual Named Entity Challenge focusing on recognizing mentions of NEs in Web documents in three Slavic languages, normalization of the NEs, and cross-lingual entity linking.

Seven teams registered and three of them actively participated in the Challenge and submitted system results with multiple variants. Most systems use state-of-the-art transformer-based models. Overall, the results of the best-performing systems are quite strong for extraction and normalization of names, while entity linking—and in particular, cross-lingual entity linking—remains a very challenging task.

We present the summary results for the main aspects of the challenge and the best-performing model from each team.

To foster further research into NLP for Slavic languages, including cross-lingual entity linking, our training and test datasets, the detailed annotations, and scripts used for the evaluations are made available to the research community on the Shared Task's Web page.[10]

Acknowledgments

Work on Czech has been partly supported by grant No. SGS-2022-016 Advanced methods of data processing and analysis and computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

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 Anisia Katinskaia for her kind assistance in preparation of the test corpora.

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.

Ekaterina Artemova, Maxim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Vladimir Ivanov, and Elena Tutubalina. 2022. Runne-2022 shared task: Recognizing nested named entities. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "DIALOGUE"*, pages 33–41.

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.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

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.

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

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. NEREL: A Russian dataset with nested named entities, relations and events. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 876–885, Held Online. INCOMA Ltd.

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.

Gabriela Pałka and Artur Nowakowski. 2023. Exploring the use of foundation models for named entity recognition and lemmatization tasks in slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing*. European Association for Computational Linguistics.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine. Association for Computational Linguistics.

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.

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.

Daiga Deksne Rinalds Vīksna, Inguna Skadiņa and Roberts Rozis. 2023. Large language models for multilingual slavic named entity linking. In *Proceedings of the 9th Workshop on 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.

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.

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.