Universal NER v2: Towards a Massively Multilingual Named Entity Recognition Benchmark

Terra Blevins[1], Stephen Mayhew[2], Marek Šuppa[3], Hila Gonen[4], Shachar Mirkin[5], Vasile Pais[6], Kaja Dobrovoljc[7], Voula Giouli[8], Jun Kevin[9], Eugene Jang[1], Eungseo Kim[10], Jeongyeon Seo[11], Xenophon Gialis[12], Yuval Pinter[13]

[1]Northeastern University, USA [2]Duolingo, USA [3]Comenius University Bratislava, Slovakia [4]University of British Columbia, Canada [5]Alpinference, France [6]Research Institute for Artificial Intelligence, Romanian Academy, Romania [7]University of Ljubljana, Slovenia [8]Aristotle University of Thessaloniki / ILSP, Athena Research Center, Greece [9]Universitas Pelita Harapan, Indonesia [10]Seoul National University, South Korea [11]Independent Researcher, South Korea [12]Democritus University of Thrace, Greece [13]Ben-Gurion University of the Negev, Israel

t.blevins@northeastern.edu, stephen@duolingo.com, marek@suppa.sk, hilagnn@gmail.com, shacharmirkin@gmail.com vasile@racai.ro, kaja.dobrovoljc@ff.uni-lj.si, pgiouli@del.auth.gr junkevin88@gmail.com, ej16056@gmail.com juniormoo@snu.ac.kr, yena.seo@kaist.ac.kr, xenogial@pme.duth.gr, yuvalpinter@gmail.com

Abstract

We present Universal NER (UNER) v2, a significant extension of the initial version released in 2024. UNER is a collaborative dataset for multilingual named-entity annotations, built to support research on NER methods in a cross-linguistic setting. UNER v2 adds 11 new datasets in 10 typologically varied language varieties to the resource, including multiple parallel evaluation benchmarks aligned with each other and other datasets in UNER v1, while maintaining the same annotation guidelines and high standards for inter-annotator agreement. We report detailed statistics for the dataset and benchmark UNER v2 using both encoder-based model architectures and LLMs.

Keywords: Named Entity Recognition, Multilingual, Benchmark, Cross-lingual Transfer, Large Language Models

1. Introduction

While multilingual language models promise to bring the benefits of LLMs to speakers of many languages, gold-standard evaluation benchmarks in most languages to interrogate these assumptions remain scarce. The Universal NER project, now entering its fourth year, is dedicated to building gold-standard multilingual Named Entity Recognition (NER) benchmark datasets. Inspired by existing massively multilingual efforts for other core NLP tasks (e.g., Universal Dependencies; de Marneffe et al., 2021), the project uses a general tagset and thorough annotation guidelines to collect standardized, cross-lingual annotations of named entity spans. The first installment (UNER v1) was released in 2024 (Mayhew et al., 2024), and the project has continued and expanded since then, with various organizers, annotators, and collaborators in an active community.

We present a substantial update to the Universal Named Entity Recognition (UNER) project, with new, gold-standard annotations on eleven datasets in ten languages (nine of which are new to the UNER collection), comprising 59,000 entities over 1.5 million tokens. As shown in Figure 1, the addition of these datasets brings the composition of the overall Universal NER repository to 30 datasets across 22 languages, with a total of 3 million annotated tokens. We release this aggregated dataset as UNER v2.

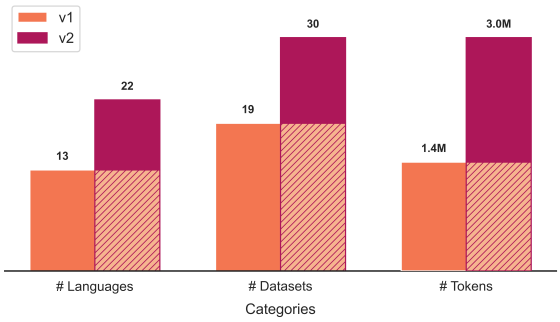

Figure 1: Comparison of the dataset statistics for UNER v1 and UNER v2.

In this work, we first provide a summary of the Universal NER project thus far (Section 2) and then describe the creation of UNER v2 datasets, along with relevant dataset statistics and analyses (Section 3). We then present baseline experiments on

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6609–6618
11–16 May 2026.

Language		Dataset	
EL	Greek	gdt	Greek Dependency Treebank
HE	Hebrew	htb	Hebrew Dependency Treebank
NNO	Norwegian Nynorsk	ndt	Norwegian Dependency Treebank
NOB	Norwegian Bokmål	ndt	Norwegian Dependency Treebank
SL	Slovenian	ssj	Slovenian SSJ Treebank
SV	Swedish	lines	Swedish LinES Treebank
CS	Czech	pud	Parallel Universal Dependencies
ID	Indonesian	pud	Parallel Universal Dependencies
JA	Japanese	pud	Parallel Universal Dependencies
KO	Korean	pud	Parallel Universal Dependencies
RO	Romanian	legalnero	Romanian Named Entities in the Legal Domain

Table 1: Languages and associated original dataset names used in UNER v2.

the new datasets introduced in UNER v2, benchmarking them against both encoder-based and generative language models (Section 4). Our experimental results show that state-of-the-art generative models achieve at most 0.50 average F1, with performance dropping sharply on typologically distant and lower-resourced languages. UNER continues to be an essential benchmark for developing more robust, truly multilingual NER systems.

2. Universal NER

The development of multilingual Named Entity Recognition benchmarks has accelerated in recent years. The Universal NER project (UNER v1 Mayhew et al., 2024) established a community-driven gold standard resource in 13 languages, following the philosophy of projects such as Universal Dependencies (UD; de Marneffe et al., 2021) or the PARSEME initiative for multiword expressions (Savary et al., 2017). Complementary efforts have also emerged in domain-specific contexts, such as LegalNERo (Păiș et al., 2024), which focuses on the Romanian legal domain, and NorNE (Jørgensen et al., 2020), which contains named entity annotations for Norwegian.

The hallmark of the Universal NER project is a simple and transferable tagset consisting of 3 tags – Person (PER), Location (LOC), and Organization (ORG) – as well as thorough annotation guidelines. Annotation is performed using shared software based on TALEN (Mayhew and Roth, 2018), and gathered on GitHub. To encourage high-quality annotations, we require (with exceptions) that at least 2 annotators provide annotations on at least 5% of the data, so that we can calculate agreement statistics. In many cases, we have found that even when there is low overlap between annotators, agreement statistics can uncover annotation guideline misunderstandings or incompetent annotators, and identify potential quality issues.

Impact of UNER v1 Since its introduction in 2024, the first version of this dataset has had a substantial impact on the multilingual NLP community, garnering citations from over 40 research works. Its significance is further evidenced by its integration into major multilingual resources: the Aya Dataset, a prominent open-access collection for multilingual instruction tuning (Singh et al., 2024), and skLEP, a comprehensive Slovak language understanding benchmark (Suppa et al., 2025).

Other works have leveraged the first version of the dataset to investigate targeted questions—such as the effect of NER on translation (Singh et al., 2025) and the efficacy of synthetic data for NER tasks (Kamath and Vajjala, 2025)—as well as broader fundamental questions concerning cross-lingual transfer (Liu and Niehues, 2025; Chen et al., 2023b) and multilingual knowledge distillation (Wibowo et al., 2024). The dataset continues to enable explorations across a wide spectrum of multilingual research directions (Africa et al., 2025; Feher et al., 2025; Kesen et al., 2025; Straková and Straka, 2025; Berger et al., 2024).

3. UNER v2 Creation and Dataset

Universal NER (UNER) v2 is an extension of UNER that adds 11 new datasets in 10 new language varieties (with Norwegian represented by two written standards, Nynorsk and Bokmål), namely Modern Greek, Hebrew, Norwegian Nynorsk, Norwegian Bokmål, Slovenian, Swedish, Czech, Indonesian, Japanese, Korean, and Romanian (Table 2; the full data composition of UNER v2 is given in Figure 1), for a total of 30 datasets spanning 22 languages now included in the UNER project. This updated version of UNER also incorporates minor annotation fixes for some existing datasets originally released in v1 (namely, English's EWT and PUD). Here, we present the annotation process for UNER v2 (Section 3.1) and the dataset statistics for the

		Sentences				Entities				Tokens			
Lang.	Dataset	Train	Dev	Test	All	Train	Dev	Test	All	Train	Dev	Test	All
EL	gdt	1,662	403	456	2,521	1,551	501	436	2,488	42,326	10,443	10,672	63,441
HE	htb	5,241	484	491	6,216	6,013	434	439	6,886	137,717	11,412	12,282	161,411
NNO	ndt	14,174	1,890	1,511	17,575	10,348	1,111	897	12,356	245,330	31,250	24,773	301,353
NOB	ndt	15,696	2,409	1,939	20,044	10,062	1,438	1,259	12,759	243,886	36,369	29,966	310,221
SL	ssj	10,903	1,250	1,282	13,435	7,293	907	724	8,924	215,155	26,500	25,442	267,097
SV	lines	3,176	1,032	1,035	5,243	1,221	433	451	2,105	55,451	18,515	16,994	90,960
CS	pud	–	–	1,000	1,000	–	–	1,008	1,008	–	–	18,610	18,610
ID	pud	–	–	1,000	1,000	–	–	1,110	1,110	–	–	19,446	19,446
JA	pud	–	–	1,000	1,000	–	–	1,150	1,150	–	–	28,788	28,788
KO	pud	–	–	1,000	1,000	–	–	1,172	1,172	–	–	16,584	16,584
RO	legalnero	–	–	8,284	8,284	–	–	8,996	8,996	–	–	265,335	265,335

Table 2: Dataset statistics for the new benchmarks included in UNER v2.

new datasets released in v2 (Section 3.2). Additionally, we conduct a cross-lingual analysis of the pud labels to evaluate inter-language consistency in named entity usage across parallel datasets (Section 3.3).

3.1. Data Annotation

The annotation process for UNER v2 generally followed the procedure laid out in the first UNER data collection effort. We used the same annotation guidelines, with minor wording changes to clarify the class distinctions: documents are annotated for location (LOC), organization (ORG), and person (PER) entities.[1] Annotators for each dataset were recruited as volunteers from the NLP community, and the annotation effort was primarily coordinated through a collaborative Discord server.

Following UNER v1, annotations were collected using the TALEN platform (Mayhew and Roth, 2018), a web-based tool developed for span-level sequence labeling. Additionally, all datasets annotated from scratch for UNER v2 include overlapping annotations from multiple annotators for at least 5% of dataset documents to calculate inter-annotator agreement (Table 4). Each dataset is released in a custom .conllu format (termed .iob2); the overall format follows .conllu (Nivre et al., 2020) but contains the following column types for word-level information: word id, word form, UNER label, XNER label,[2] and annotator id. Since text is annotated at the word-level in this format, NER spans are annotated using the same IOB2 annotation schema as the datasets in v1.

The newly added Greek (EL) dataset was annotated on top of the Greek gdt (Prokopidis and Papageorgiou, 2017) in accordance with the UNER guidelines. Language-specific examples from prior – yet compatible – annotation efforts in the domains of finance (Boutsis et al., 2000) and in multi-domain journalistic texts (Giouli et al., 2006) were actively consulted during annotation to guide challenging cases and support consistent decisions.

Dataset Transfer UNER v2 also incorporates three existing datasets converted to align with the UNER annotation standard. Norwegian ndt (containing two written standards NNO and NOB) and Slovenian (SL) ssj are originally annotated in existing Universal Dependencies treebanks (Øvrelid and Hohle, 2016; Dobrovoljc et al., 2017). Norwegian ndt is annotated with named entities as an independent effort by Jørgensen et al. (2020), and then automatically converted to the UNER format with a fixed label mapping. The Slovenian ssj dataset, by contrast, contains partial named entity annotations based on the JANES-NER scheme (Arhar Holdt et al., 2024; Zupan et al., 2017). The conversion to UNER completed these annotations by adding 1,383 new entities and then automatically converting and manually correcting all labels to conform to the UNER guidelines.

However, Romanian (RO) legalnero is not part of UD and is instead annotated on a dataset of Romanian legal documents (Păiș et al., 2024). The original version provides gold annotations for organizations, locations, persons, time expressions, and legal resources mentioned in legal documents. The raw text files were extracted from the Romanian part (Tufiș et al., 2020) of the MARCELL corpus (Váradi et al., 2020). The annotations are automatically converted to the UNER format by selecting only the entity types supported by the UNER scheme. The corpus contains 265,335 tokens in 8,284 sentences over 370 text files.

Annotation Corrections In addition to adding new datasets to Universal NER, UNER v2 also comprises revisions of the English EWT and PUD datasets. Specifically, we relabel a handful of enti-

[1] While entities that do not fall into these categories are labeled as OTH (Other) during annotation, these additional entities are not included in the final dataset release.

[2] XNER is only present for datasets that have been transferred to the UNER format, and contains the NER labels from the original datasets.

Data Source	Lang.	Dataset	Domains	Entity Dist. (%) LOC	ORG	PER
Prokopidis and Papageorgiou (2017)	EL	gdt	news, spoken, wiki	38.0%	41.5%	20.5%
Tsarfaty (2013); McDonald et al. (2013)	HE	htb	news	24.0%	39.7%	36.3%
Øvrelid and Hohle (2016)	NNO	ndt	blog, news, nonfiction	28.4%	30.1%	41.5%
	NOB	ndt	blog, news, nonfiction	27.3%	31.6%	41.1%
Dobrovoljc et al. (2017)	SL	ssj	fiction, news, nonfiction	34.1%	22.1%	43.8%
Ahrenberg (2015)	SV	lines	fiction, nonfiction, spoken	18.5%	7.9%	73.6%
Zeman et al. (2018)	CS	pud	news, wiki	38.3%	20.8%	40.9%
	ID	pud	news, wiki	47.0%	16.1%	36.9%
	JA	pud	news, wiki	47.0%	16.9%	36.1%
	KO	pud	news, wiki	40.0%	24.7%	35.3%
Păiș et al. (2024)	RO	legalnero	legal	21.7%	68.4%	9.9%

Table 3: Domains and distribution of entity types for datasets in UNER v2. Domains are categorized for the underlying UD datasets at https://universaldependencies.org/ or author descriptions if not included in UD.

ties (≤ 3) in these datasets that were found to be incorrect in the v1 version. The total number of annotated entities remains the same.

3.2. Dataset Statistics

Table 2 illustrates the dataset statistics for each of the new datasets released in UNER v2. This new version of UNER adds six full NER datasets, as well as four new parallel NER evaluation sets that are aligned with six additional languages already annotated in UNER v1. The new datasets are also typologically diverse, spanning five macrolanguage families (Indo-European, Afroasiatic, Austronesian, Koreanic, and Japonic); this brings the full coverage of the UNER datasets to 22 languages across six language families.[3] The new languages include fusional (Czech, Greek, Norwegian, Slovenian, Swedish), agglutinative (Japanese, Korean), analytic (Indonesian), and introflexive (Hebrew) morphological types, written in four scripts (Latin, Greek, Hebrew, and CJK). This diversity matters for NER: morphological complexity affects where entity boundaries fall (e.g., agglutinative suffixes in Korean), and script differences impact how multilingual models tokenize text.

In addition to linguistic diversity, the datasets in UNER v2 are also stylistically diverse (Table 3). The datasets broadly cover fiction and nonfiction text, as well as spoken-language transcripts. Additionally, legalnero adds the legal domain to UNER in Romanian, while SV lines expands UNER's Swedish coverage to include literary texts. We correspondingly observe distribution shifts in entity use across domains; for example, the legal documents in legalnero contain a much higher proportion of ORG entities than the other datasets, while having very few ($< 10\%$) PER entities.

[3]UNER v1 includes the Sino-Tibetan family with two Chinese datasets.

Finally, we present the inter-annotator agreement (IAA) for new datasets in UNER v2 in Table 4. We observe similar agreement trends to those seen in the original UNER annotation process: specifically, agreement between the annotators on PER entities is usually much higher than on the LOC and ORG entities, due to the prevalence of ambiguous entities that depend on (often vague or missing) context. For instance, metonymic organizations are common, leading to the use of the same form to describe an organization (*The White House* announced a new policy today) and a location (The meeting was held at *the White House*).

3.3. Analyzing Named Entities across Parallel Multilingual Data

A key contribution of Universal NER is the creation of parallel gold-standard evaluation benchmarks for Named Entity Recognition, annotated on top of Parallel UD (PUD; Zeman et al., 2018); this release of UNER adds four more typologically diverse languages to this resource: Czech (fusional SVO), Indonesian (analytic SVO), and Japanese and Korean (agglutinative SOV). In this section, we examine the cross-lingual consistency of named entity usage across all ten languages in the UNER dataset.

Figure 2 displays the results of our cross-dataset analysis. Similar to the initial analysis of the UNER v1 parallel datasets, the Indo-European (IE) languages (including the addition of Czech), strongly agree in terms of entity distribution and use. In contrast, by adding examples from other families, we observe that the behavior of languages outside the IE family is more complex. Indonesian and Korean behave similarly to Chinese, with less overlap with IE languages (and moderate agreement with each other and with Chinese); in contrast, Japanese exhibits alignment patterns similar to those of the IE languages, albeit with many additional LOC entities.

		Train				Dev				Test			
Lang.	Dataset	LOC	ORG	PER	% Docs	LOC	ORG	PER	% Docs	LOC	ORG	PER	% Docs
EL	gdt	0.750	0.556	0.962	25%	0.843	0.607	0.883	25%	0.401	0.408	0.485	26.9%
HE	htb	0.777	0.785	0.839	29.7%	0.741	0.664	0.968	100%	0.746	0.831	0.914	100%
SV	lines	**	**	**	**	0.927	0.526	0.974	100%	0.571	0.667	0.962	25%
CS	pud	–	–	–	–	–	–	–	–	0.656	0.840	0.775	5.3%
ID	pud	–	–	–	–	–	–	–	–	0.629	0.590	0.930	8.8%
JA	pud	–	–	–	–	–	–	–	–	0.651	0.769	0.839	6.3%
KO	pud	–	–	–	–	–	–	–	–	0.613	0.443	0.945	11.8%

Table 4: Inter-annotator agreement (IAA) for the new benchmarks included in UNER v2. – indicates that the data split does not exist, while ** indicates that IAA was not collected. IAA was also not calculated for datasets transferred into UNER v2: NNO and NOB ndt, SL ssj, and RO legalnero.

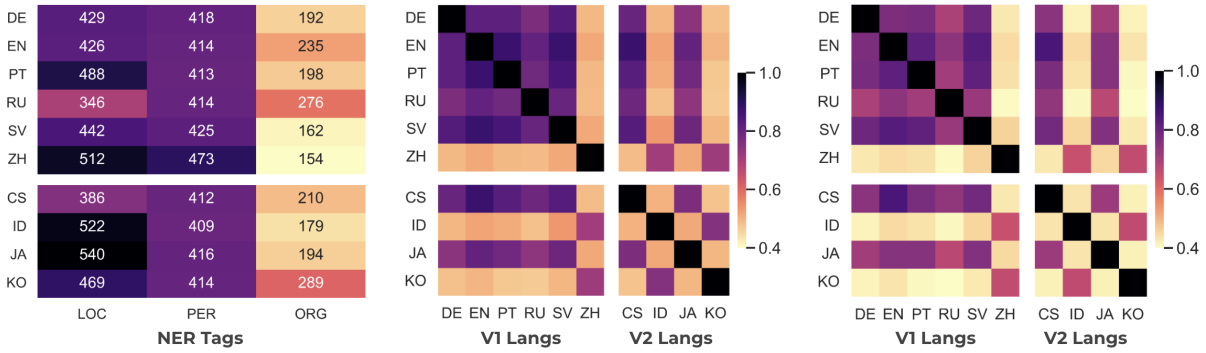

Figure 2: A cross-lingual comparison of UNER annotations on top of parallel text (PUD). We consider the overlap of datasets from UNER v1 and v2. **Left**: The overall tag distribution in each PUD dataset. **Center**: Sentence-level agreement between languages for entity count. **Right**: Sentence-level agreement on entity count within each class (LOC, ORG, PER) between languages.

4. Experiments

We perform two sets of experiments: (a) using a finetuned cross-lingual encoder, XLM-R (Conneau et al., 2020), to be comparable with our prior work, Section 4.1; (b) by directly prompting three large language models (LLMs), as detailed in Section 4.2.

4.1. Traditional baselines

Experiment Setup This section establishes baselines on the new datasets in UNER v2 and provides in-language and cross-lingual results with XLM-R_{Large}. We finetuned XLM-R_{Large} (560M parameters) (Conneau et al., 2020) on the UNER datasets in which train and dev sets are available,[4] using a single NVIDIA GeForce RTX 3090 GPU. We used a learning rate of 3e-5 and batch size of 8, except for Korean (kor_pud), where we used a batch size of 4. All code was adapted from the Huggingface transformers package (Wolf et al., 2020).

[4]Greek (el_gdt), Hebrew (he_htb), Norwegian Nynorsk (nno_ndt), Norwegian Bokmål (nob_ndt), Slovenian (sl_ssj), Swedish (sv_lines), English (en_ewt)

Results and Discussion Figure 3 reports the micro F1 scores on all test sets when XLM-R_{Large} is finetuned on different languages. The in-language performance shown on the diagonal is almost always the highest among all test sets, with a few exceptions for closely related languages. We observe that in most cases, cross-lingual transfer performs well between European languages, achieving over 0.60 F1. However, transfer results in lower performance on non-European languages such as Japanese (ja_pud) and Korean (ko_pud), which aligns with observations from previous work that cross-lingual transfer to typologically distant languages remains challenging (Chen et al., 2023a; Wu et al., 2020).

Overall, the tag-level performance breakdown reveals that F1 on ORG is consistently the lowest, and LOC is often the second lowest. This is consistent with the ambiguity between ORG and LOC entities (e.g., metonymic uses discussed in Section 3.2), whereas person names are usually less ambiguous, resulting in a higher F1 on PER for most datasets.

Figure 3: Experimental results for in-language and cross-lingual UNER performance with XLM-R_{large}.

4.2. LLM Results

LLMs have been shown to be beneficial for text annotation tasks, particularly for Named Entity Recognition (Tan et al., 2024; Wang et al., 2023). Recent studies have explored their potential as cost-effective annotators and data generators (Dagdelen et al., 2024; Bogdanov et al., 2024), though their performance typically falls short of fine-tuned models (Hu et al., 2024).

We ran a set of experiments with LLMs, providing the model with the annotation guidelines and asking it to follow them when annotating each sentence in the dataset. We experimented with three models as our LLM annotator: *Gemini 2.5 Flash Lite* (*Gemini* below), *Claude Sonnet 4* (*Claude*), and *GPT-5 Mini* (*GPT* below).[5] These were chosen to represent the balance between state-of-the-art performance and inference cost at the time of writing. We deliberately did not engage in prompt engineering, as we aimed to observe what LLMs can achieve when presented with the exact same guidelines as human annotators received. In that, we can consider the results shown here as a baseline for LLM-based annotations on this dataset.

Figure 4 presents the inter-annotator agreement between each model and the human annotator who annotated the most documents in a given dataset. For brevity, we include a sample of the results, specifically removing datasets that are too small to be considered meaningful for this measurement. The left-most column of the table shows the human IAA for comparison.

Among the three models, Claude achieves the highest average F1 (0.497), followed by GPT and Gemini, though all remain substantially below human inter-annotator agreement (0.741 average F1). Performance varies across languages, from strong agreement on Greek `el_gdt` to low agreement on Czech `cs_pud`.[6] As with human IAA scores, PER is the easiest entity type for LLMs to annotate, while LOC and ORG are harder.

A closer look at annotation counts reveals systematic over- and under-annotation patterns. On the PUD parallel datasets, GPT and Gemini consistently produce 30–90% more entity annotations than humans (e.g., on English PUD, humans annotate 682 entities while GPT produces 1,262), suggesting widespread annotation of non-**named** entities such as generic references to organizations. This pattern is language-specific: on Hebrew and Korean, all three models instead *under*-annotate, producing 30–60% fewer entities than humans. We also observed that models overlooked nuances in the guidelines, such as the requirement to annotate geopolitical entities as organizations rather than locations, and the distinction between nationalities and locations.

Although LLM annotation quality falls behind that of humans on this dataset, it can likely be improved with prompt engineering or an agentic flow (equivalent to a human annotator discussing a specific annotation). We note that LLMs have potential not only as annotators, but also as a means to surface errors in human annotations and to identify flaws in the annotation guidelines. We intend to explore these directions in the future.

[5]Exact model versions: `gemini-2.5-flash-lite`, `anthropic.claude-sonnet-4-20250514-v1`, `gpt-5-mini-2025-08-07`.

[6]We note that the agreement depends not only on the language and the dataset but also on the human annotators themselves.

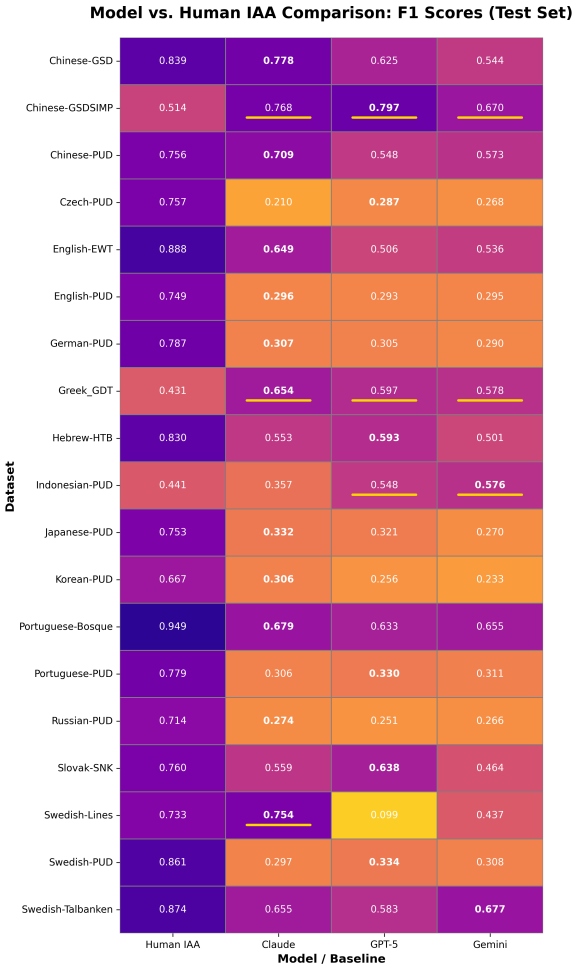

Figure 4: F1 score comparison across 19 multilingual NER datasets (test sets) for three LLMs against human Inter-Annotator Agreement (IAA) baseline. Bold values indicate the best-performing model for each dataset; underlined values indicate scores exceeding human agreement.

5. Prior Work

In parallel to the UNER efforts, new multilingual datasets are being created to support broader instruction tuning and evaluation. The Aya dataset (Singh et al., 2024) provides large-scale multilingual resources for instruction tuning, while Zhang and Xiao (2024) propose a classification framework to better organize and understand the diversity of NER datasets. Beyond traditional NER, benchmarking efforts have expanded to more complex evaluation settings, such as situational awareness of large language models (Tang et al., 2024).

On the modeling side, advances in multilingual representation learning have also influenced NER. Recent work has investigated middle-layer alignment for cross-lingual transfer in fine-tuned large language models (Liu and Niehues, 2025), meta-pretraining strategies for zero-shot NER in low-resource Philippine languages (Africa et al., 2025), and dynamic tokenization approaches for retrofitting large language models (Feher et al., 2025). Novel paradigms such as multilingual pretraining for pixel language models (Kesen et al., 2025) also highlight emerging directions beyond conventional text-only approaches.

6. Conclusion

We presented Universal NER v2, a new and substantially expanded version of the ongoing Universal NER project. We are excited to see the steady pace at which the resource is growing, and hope that it can reach the magnitude of other massive endeavors in multilingual linguistic annotation such as Universal Dependencies and UniMorph. As we release v2, significant work is being done towards incorporating more languages in the resource, including all levels from annotation to validation. Through its coverage of less-resourced or even underrepresented and typologically diverse languages, the dataset contributes to a more inclusive and universal approach to language technology. We invite more collaborators to contribute to future versions, whether in existing or new languages.

UNER v1 has already made an impact on the NER community, facilitating multilingual and cross-lingual evaluation of modeling techniques in a controlled, and even parallel setting. UNER v2 enhances and furthers this core ability. We eagerly anticipate the further advances in multilingual understanding that this will facilitate.

Data and Code Availability

The UNER v2 dataset, annotation guidelines, and benchmarking code are publicly available. The project website,[7] including annotation guidelines,[8] provides an overview of the resource. All annotations are released under the CC-BY-SA-4.0 license and hosted on GitHub.[9] The dataset is also available on Hugging Face.[10]

Acknowledgments

Multiple datasets new to UNER 2.0 were developed thanks to collaboration via the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). Slovenian dataset

[7] https://www.universalner.org/
[8] https://www.universalner.org/guidelines/
[9] https://github.com/UniversalNER
[10] https://huggingface.co/datasets/universalner/universal_ner

creation was partially supported by grants ARIS-GC-002 and HORIZON-WIDERA-2023-TALENTS-01-01-101186647. MŠ was partially funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I02-03-V01-00029.

7. References

David Demitri Africa, Suchir Salhan, Yuval Weiss, Paula Buttery, and Richard Diehl Martinez. 2025. Meta-pretraining for zero-shot cross-lingual named entity recognition in low-resource philippine languages. *arXiv preprint arXiv:2509.02160*.

Lars Ahrenberg. 2015. Converting an english-swedish parallel treebank to universal dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 10–19.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek, Tina Munda, Nejc Robida, Luka Terčon, and Slavko Zitnik. 2024. SUK 1.0: A new training corpus for linguistic annotation of modern standard Slovene. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15428–15435, Torino, Italia. ELRA and ICCL.

Uri Berger, Tal Baumel, and Gabriel Stanovsky. 2024. In-context learning on a budget: A case study in named entity recognition. *arXiv e-prints*, pages arXiv–2406.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. NuNER: Entity recognition encoder pre-training via LLM-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11829–11841.

Sotiris Boutsis, Iason Demiros, Voula Giouli, Maria Liakata, Harris Papageorgiou, and Stelios Piperidis. 2000. A system for recognition of named entities in greek. In *Proceedings of the Second International Conference on Natural Language Processing*, pages 424–436.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023a. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Yang Chen, Vedaant Shah, and Alan Ritter. 2023b. Translation and fusion improves zero-shot cross-lingual information extraction. *arXiv preprint arXiv:2305.13582*.

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.

John Dagdelen, Alex Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Darius Feher, Ivan Vulić, and Benjamin Minixhofer. 2025. Retrofitting large language models with dynamic tokenization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29866–29883, Vienna, Austria. Association for Computational Linguistics.

Voula Giouli, Alexis Konstandinidis, Elina Desypri, and Harris Papageorgiou. 2006. Multi-domain multi-lingual named entity recognition: Revisiting & grounding the resources issue. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(8):1812–1820.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik

Velldal. 2020. Norne: Annotating named entities for norwegian. In *LREC 2020, Twelfth International Conference On Language Resources And Evaluation*. European Language Resources Association (ELRA).

Gaurav Kamath and Sowmya Vajjala. 2025. Does synthetic data help named entity recognition for low-resource languages? *arXiv preprint arXiv:2505.16814*.

Ilker Kesen, Jonas F Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. *arXiv preprint arXiv:2505.21265*.

D. Liu and J. Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Stephen Mayhew and Dan Roth. 2018. TALEN: Tool for annotation of low-resource ENtities. In *Proceedings of ACL 2018, System Demonstrations*, pages 80–86, Melbourne, Australia. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Lilja Øvrelid and Petter Hohle. 2016. Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1579–1585, Portorož, Slovenia. European Language Resources Association (ELRA).

Prokopis Prokopidis and Harris Papageorgiou. 2017. Universal dependencies for greek. In *Proceedings of the nodalida 2017 workshop on universal dependencies (udw 2017)*, pages 102–106.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2024. Legalnero: A linked corpus for named entity recognition in the romanian legal domain. *Semantic Web*, 15:831–844.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, pages 31–47.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Sumit Singh, Pankaj Goyal, and Uma Tiwary. 2025. silp_nlp at semeval-2025 task 2: An effect of entity awareness in machine translation using llm. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2389–2394.

Jana Straková and Milan Straka. 2025. Nametag 3: A tool and a service for multilingual/multitagset ner. *arXiv preprint arXiv:2506.05949*.

Marek Suppa, Andrej Ridzik, Daniel Hládek, Tomáš Javůrek, Viktória Ondrejová, Kristína Sásiková,

Martin Tamajka, and Marian Simko. 2025. skLEP: A Slovak general language understanding benchmark. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26716–26743, Vienna, Austria. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

G. Tang, Z. Chu, W. Zheng, M. Liu, and B. Qin. 2024. Towards benchmarking situational awareness of large language models: Comprehensive benchmark, evaluation and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7904–7928.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–584, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. 2020. Collection and annotation of the romanian legal corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2766–2770, Marseille, France. European Language Resources Association.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. The marcell legislative corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3754–3761, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Haryo Akbarianto Wibowo, Thamar Solorio, and Alham Fikri Aji. 2024. The privileged students: On the value of initialization in multilingual knowledge distillation. *arXiv preprint arXiv:2406.16524*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Y. Zhang and G. Xiao. 2024. Named entity recognition datasets: a classification framework. *International Journal of Computational Intelligence Systems*, 17(1):71.

Katja Zupan, Nikola Ljubešić, and Tomaž Erjavec. 2017. Annotation guidelines for slovenian named entities Janes-NER. Technical report, Centre for Language Resources and Technologies (CJVT), University of Ljubljana.