Studia Metrica et Poetica 12.2, 2025, 7–18

Named Entity Recognition and Linking in PoeTree Corpora

Petr Plecháč, Artjoms Šeļa, Silvie Cinková, Mirella De Sisto, Lara Nugues, Neža Kočnik, Robert Kolár, Thomas Haider*

Abstract. Named entity recognition (NER) and named entity linking (NEL) remain underexplored in poetic texts. This study provides the first large-scale evaluation of contemporary NER and NEL systems on poetry across seven European languages – Czech, German, English, French, Italian, Russian, and Slovenian – using corpora from the PoeTree project. We benchmark three NER systems (flair, NameTag 2, spaCy) and three GPT models (GPT-3.5, GPT-4, GPT-4 Turbo) against manually annotated gold standards. While results fall short of in-domain benchmarks, they significantly outperform earlier findings. Manual correction further raises final annotation quality to estimated F_1 scores between 0.77 and 0.93 across languages. We additionally evaluate two NEL systems – spaCy fishing and mGenre – showing that mGenre consistently outperforms spaCy fishing, achieving in-KB F_1-scores of 0.70–0.81. By analysing geographic distances between predicted and gold-standard links, we demonstrate that a substantial portion of "incorrect" predictions are near-miss ambiguities rather than substantive errors. The resulting manually verified geolocation annotations have been integrated into PoeTree and made available through an interactive map interface.

Keywords: poetry, named entities, computational poetics, natural language processing

* Authors' addresses: Petr Plecháč, Institute of Czech Literature of the Czech Academy of Sciences, Na Florenci 1420/3, 110 00 Prague, Czechia, email: plechac@ucl.cas.cz; Artjoms Šeļa, Institute of Czech Literature of the Czech Academy of Sciences, Na Florenci 1420/3, 110 00 Prague, Czechia, email: sela@ucl.cas.cz; Silvie Cinková, Institute of Formal and Applied Linguistics, Charles University, Malostranské náměstí 2/25, 118 00, Prague, Czechia, email: cinkova@ufal.mff.cuni.cz; Mirella De Sisto, Department of Computational Cognitive Science, Tilburg University, Warandelaan 2, 5037 AB Tilburg, Netherlands, email: M.DeSisto@tilburguniversity.edu; Lara Nugues, University of Basel, Maiengasse 51, CH-4056 Basel, Switzerland, email: lara.nugues@unibas.ch; Neža Kočnik, University of Maribor, Slomškov trg 15, 2000 Maribor, Slovenia, email: neza.kocnik2@um.si; Robert Kolár, Institute of Czech Literature of the Czech Academy of Sciences, Na Florenci 1420/3, 110 00 Prague, Czechia, email: kolar@ucl.cas.cz; Thomas Haider, University of Passau, Innstr. 41, 94032 Passau, Germany, email: thomas.haider@uni-passau.de.

https://doi.org/10.12697/smp.2025.12.2.01

0. Introduction

Named entity recognition (NER) and named entity linking (NEL) rank among the most popular fields in contemporary natural language processing (NLP). The task is to automatically identify mentions of real world entities – such as people, locations, or organizations – within a text and to disambiguate them by linking each mention to a specific entry in a knowledge base. For instance, in the sentence "Before the pilgrims landed in Plymouth, we were here" the system must not only recognize "Plymouth" as a reference to a real location (NER) but also to determine that it refers to Plymouth, Massachusetts, rather than Plymouth, UK (NEL).

Despite the crucial role the concept of "topoi" plays in poetics, these methods have rarely been used with poetic texts. Cross-domain application of NLP tools usually results in performance drop, and to date, no large NER/NEL models have been specifically trained on poetry. Moreover, there is an important pitfall that further complicates the task: capitalization – a usual marker of proper names – occurs in poetry in unusual contexts, be it the convention of capitalizing line-initial letters or irregular capitalization inside the lines ("Hail, holy Light, offspring of Heaven firstborn / Or of the Eternal coeternal beam"). In 2019 Foley found out that F_1-score of spaCy multilingual model for NER does not exceed even 0.05 when applied to English poetry, performing only slightly better than random (Foley 2019: 46).

Since then, however, NER/NEL systems have made significant progress, with the recent advent of LLMs further advancing the field. With regards to it, we set out to annotate a collection of poetry corpora in seven different languages for named entities, namely Czech (*cs*), German (*de*), English (*en*), French (*fr*), Italian (*it*), Russian (*ru*) and Slovenian (*sl*); all coming from the PoeTree project (Plecháč et al. 2024). In *it*, which unlike other corpora spans over more than six centuries, this was limited to authors born after 1700. In *de* this was done prior to enlarging the corpus with data from the *Deutsches Lyrik Korpus*, therefore only data from the *Metricalizer* corpus were included. Here we first evaluate the performance of several NER systems, describe how toponyms were labelled and linked in our data, and finally provide the evaluation of two NEL systems.

1. Named Entity Recognition

To evaluate the performance of NER systems, we created a random sample of 100 poems for each of the seven target languages and manually annotated named entities of the types LOC (location) and PER (person). Other types such as ORG (companies, institutions) were disregarded as they were deemed irrelevant to the domain. This served as a gold standard for testing the performance of three NER systems: flair (Akbik et al. 2019), NameTag 2 (Straková et al. 2019), and spaCy (Honnibal et al. 2020).

None of these, however, currently provide models for all seven languages. flair offered models for *de*, *en*, and *fr*, while NameTag 2 supported *cs*, *de*, and *en*. For spaCy we utilized the following pipelines: en_core_web_trf, sl_core_trf, and {de,fr,it,ru}_core_news_lg. Since no spaCy NER pipeline was available for Czech, we relied on multilingual xx_ent_wiki_sm.

In addition to these NER systems, we evaluated the performance of few-shot prompting using three GPT models: GPT-3.5 Turbo, GPT-4 and and GPT-4 Turbo (GPT-4o was not yet available during testing). Each poem was processed using a single prompt with the following structure:

> You are an expert in Natural Language Processing. Your task is to identify Named Entities in a given text. The possible Named Entity types are exclusively: ("LOC", "PER"). EXAMPLE: Text: "You may take me, Robert, if you will! There was a time when all thy sons were proud to speak thy name, England, when Europe echoed back aloud thy fearless fame. And thence through Berlin, Dresden, and the like, until he reached the castellated Rhine. This is William Wallace."{{"LOC": ["England", "Europe", "Berlin", "Dresden", "Rhine"], "PER": ["Robert", "William Wallace"],}} – TASK: [text of the poem to be annotated][1]

We evaluate the performance of NER systems and GPT models using *precision* (number of true positives among all spans of text annotated by the system) and *recall* (number of true positives among all spans of text annotated in the gold standard) with four different sets of conditions. More complex tagsets were simplified if necessary. For instance, spaCy's GPE (countries, cities, states) and

[1] We thank the reviewer of our article for noting that the choice of "an expert in Natural Language Processing" in our prompt may have been suboptimal. Indeed, expertise in NLP does not necessarily involve skills in literary annotation or the interpretation of poetic language; alternative roles, such as "a literary scholar" or "an expert in digital humanities annotation", might have been more appropriate.

LOC (non-GPE locations, such as mountain ranges and bodies of water) were merged under a unified LOC label in our evaluation.

- *TAG.strict*: span annotated as named entity is considered a true positive if it *lines up exactly* with a span in the gold standard and *matches its label* (LOC/PER). This is the most common approach in NER evaluation.

- *TAG.relaxed*: span annotated as a named entity is considered a true positive if it *overlaps* with a span in the gold standard and *matches its label*. This accommodates cases where e.g. "Kingdom of the Netherlands" appears in gold standard annotated as LOC, while the system labels only the Netherlands as LOC.

- *SPAN.strict*: span annotated as a named entity is considered a true positive if it *lines up exactly* with a span in the gold standard *irrespective of the assigned label*. This accounts for cases where the mention was recognized but mislabeled.

- *SPAN.relaxed*: span annotated as a named entity is considered a true positive if it *overlaps* with a span in the gold standard *irrespective of the assigned label*. This accounts for cases where at least part of the mention was correctly recognized but mislabeled.

Figure 1 gives the results of our evaluation, along with F_1-scores (harmonic mean of precision and recall) reported for individual systems when tested on standard NER datasets (see Appendix). While none of the NER systems achieves the performance of in-domain applications, the results are far from the catastrophic outcomes reported in Foley (2019). Overall, the best performance is found in *en* (likely due to English-centric bias), where spaCy leads in both precision and recall at all levels of conditions. In contrast, the poorest performance is seen in *de* (likely due to a convention of capitalizing all nouns), where even the most lenient metric (SPAN.relaxed.precision) reaches only 0.6 (GPT-4 Turbo), while SPAN.relaxed.recall does not exceed 0.8 (GPT-4). As for other languages, in *cs* NameTag 2 leads by far in recall and only slightly lags behind GPT-4 Turbo in precision. Similarly, in *sl* spaCy reaches the best recall with only a marginal loss to GPT-4 Turbo in precision. For the remaining languages, NER models fall short of GPT-4 and GPT-4 Turbo. Notably, GPT-4 generally outperforms GPT-4 Turbo in recall but is consistently outperformed by GPT-4 Turbo in precision. As expected (with the exception of *sl*), GPT-3.5 fails to match the performance of its successors.

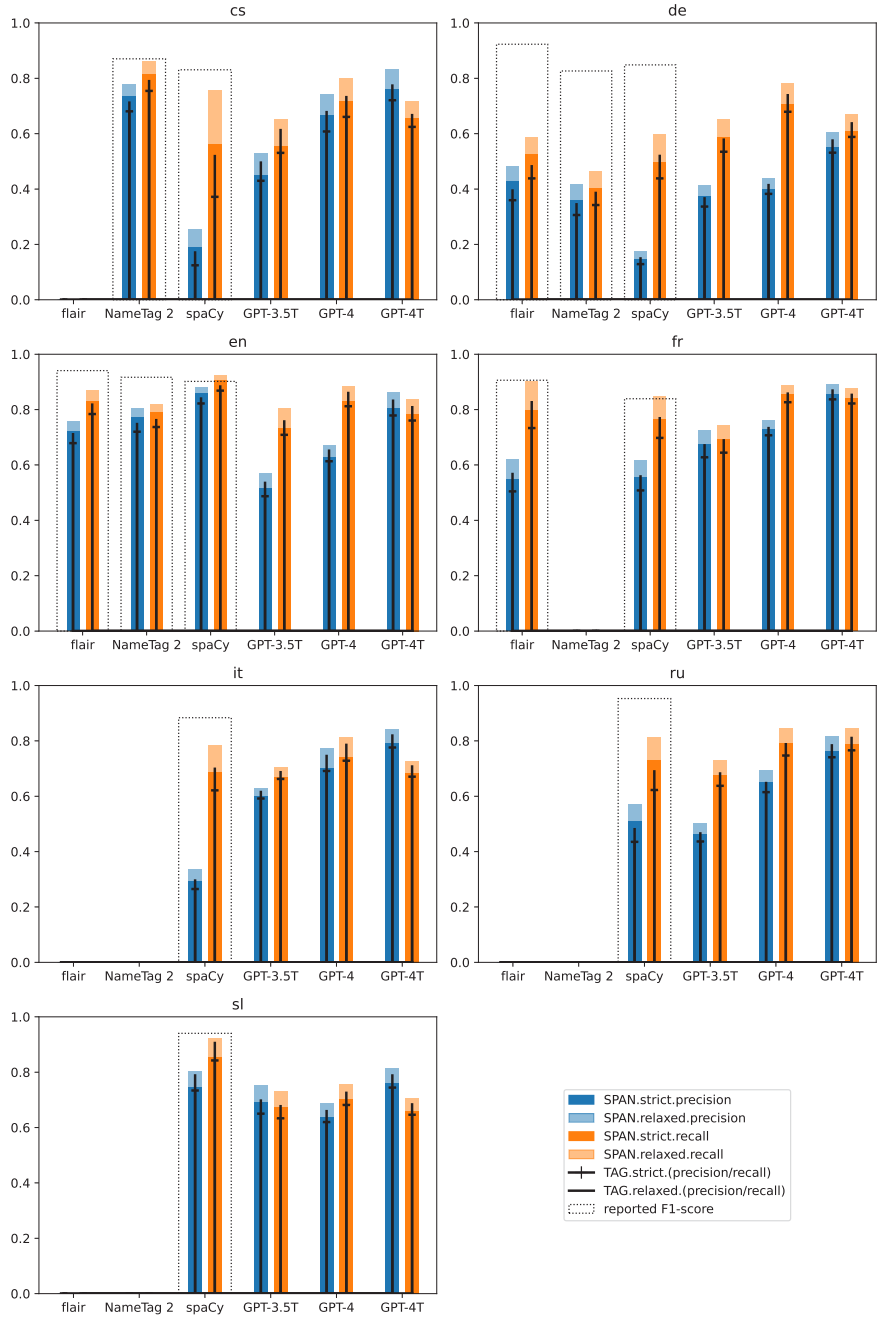

Figure 1: Evaluation of three NER systems and three GPT models.

In summary, the best performing models in *en* and *fr* are able to capture the mentions precisely in their extent and to assign them correct labels with TAG.strict.F_1-scores of approximately 0.85. In *cs*, *it*, *ru* and *sl*, the scores are slightly lower averaging around 0.75, while *de* lags significantly with a score of only 0.55. When relaxing the evaluation criterion to consider mentions that are *at least partially* captured (TAG.relaxed), F_1-scores increase modestly across languages, with gains ranging from 0.02 and 0.07. These are by no means great results, but definitely something one can build upon manually. We thus selected the best-performing system for each language to annotate the entire corpora: NameTag 2 for *cs*, spaCy for *en* and *sl*, and GPT-4 Turbo for the remaining languages. (Although the performance of GPT-4 and GPT-4 Turbo was close in some cases, we consistently opted for the latter as its API was approximately three times cheaper at the time of processing.)

2. Manual Interventions

Each corpus was assigned to an annotator tasked with reviewing all recognized named entities and performing the following steps: (1) discard false positives, (2) correct labels where necessary, and (3) link mentions to their corresponding Wikidata items. For now, this task is limited to geolocations (LOC) only, due to the volume of data.

Despite the annotators' meticulous work, occasional errors are unavoidable in a dataset of this size. We conservatively estimate the error rate of false positive removal to be as high as 5% (Marsh and Perzanowski 1998), although the actual figure is likely much lower. (Since the gold standards were produced by the same annotators, external validation is not feasible.) Under these conditions, we estimate that manual intervention raises the annotation precision to at least 0.95 across corpora. Because annotators have access to the full textual context, they can also correctly link partially recognized mentions to their proper Wikidata entries. We can therefore approximate the resulting improvement in TAG.strict.F_1 as:

$$F_1' = \frac{2 \cdot 0.95 \cdot \text{TAG.relaxed.recall}}{0.95 + \text{TAG.relaxed.recall}}$$

Table 1 presents these expected values, along with the corpus sizes and the number of named entities recognized within each corpus. The improvements are significant across languages, with three surpassing F_1' of 0.9 and only two falling below 0.85.

Table 1. F_1-scores expected after manual interventions (F_1'')

	# of poems	NER system	# of mentions		# of types		TAG. strict. F1	F_1''
			LOC	PER	LOC	PER		
cs	80,229	NameTag 2	43,661	186,682	10,322	31,840	0.72	0.87
de	53,133	GPT-4 Turbo	40,404	76,629	10,548	18,508	0.55	0.77
en	40,735	spaCy	55,358	138,199	12,439	29,687	0.85	0.92
fr	18,226	GPT-4 Turbo	24,646	41,715	5,046	9,504	0.83	0.90
it	4,698	GPT-4 Turbo	9,478	14,602	2,652	3,866	0.72	0.81
ru	45,563	GPT-4 Turbo	35,438	56,329	9,561	16,034	0.75	0.88
sl	5,587	spaCy	3,250	8,116	1,385	2,813	0.78	0.93

2. Named Entity Linking

Using manually evaluated data, we can measure how well automatic systems perform entity linking in comparison with human annotators. Here, we assess the performance of two NEL systems: the lightweight and user-friendly spaCy fishing (Terriel et al. 2022), and mGENRE, a robust system developed by Meta Research (De Cao et al. 2022), widely regarded as the current state-of-the-art in multilingual entity linking.

Common metric used for NEL evaluation is the in-KB (in-knowledge base) F_1-score. Both systems are provided with all mentions that have been manually verified and linked, along with their context and tasked to provide a link to their corresponding Wikidata item. For each mention, one of three outcomes is possible:

1. The returned link matches the one provided by the human annotator (correct).
2. The returned link differs from the one provided by the human annotator (incorrect).
3. No link is returned (null).

The in-KB F_1-score is then calculated as the harmonic mean of in-KB precision and in-KB recall, which are defined as follows:

$$\text{in-KB precision} = \frac{\text{\# of correct}}{\text{\# of correct} + \text{\# of incorrect}}$$

$$\text{in-KB recall} = \frac{\text{\# of correct}}{\text{\# of correct} + \text{\# of incorrect} + \text{\# of null}}$$

As shown in Table 2, spaCy fishing performs significantly worse than mGenre in *de*, *en*, and *ru*, and fails entirely in *fr* and *it*, primarily due to null responses (models for *cs* and *sl* were not available). mGenre on the other hand offers quite consistent performance across languages with in-KB F_1-scores ranging from 0.70 to 0.81 and no null responses.

Table 2. In-KB F_1-scores for two NEL systems

	cs	de	en	fr	it	ru	sl
SpaCy fishing	-	0.60	0.72	0.38	0.42	0.68	-
mGenre	0.75	0.70	0.78	0.81	0.77	0.78	0.73

There is, however, more granularity in the systems' output beyond simple correct/incorrect/null classification that may be worth inspecting. For example, in the following line:

> To respict the galliant Irish upon Shannon shore.
> (William Makepeace Thackeray: *The Battle of Limerick*)

mGenre does not link the mention to Shannon River (Wikidata identifier: Q192820) but rather to Shannon Town (Q611919), located just a few miles from Limerick. This is a minor error, far less severe than the one found in Wordsworth, where the system confuses Derwent River in the Lake District (Q506996) with its namesake in Tasmania (Q583565):

> In the green dales beside our Rotha's stream,
> Greta, or Derwent, or some nameless rill
> (William Wordsworth: *The Prelude*)

Moreover, there are instances where multiple links may be considered equally valid. For example:

> Shores of Lucerne! where many a winding bay
> Shone beauteous to the morn's returning ray
> (William Lisle Bowles: *Lucerne*)

In this case, the gold standard links the mention to Lake Lucerne (Q14381), while mGenre links it to the city of Lucerne (Q4191) situated on the lake's banks. These ambiguous instances frequently involve historical sites, regions, and events, for example, the Lebanese city of Tyre (Q82070) versus Tyre as a heritage site (Q3991738), Egypt (Q79) versus Ancient Egypt (Q11768), or Waterloo municipality (Q179034) versus the Battle of Waterloo (Q48314).

To take account of different types of errors we have taken all cases where the output of the NEL system was classified as incorrect. For these mentions, we retrieved the geographic coordinates of both the predicted and the gold standard items using the Wikidata REST API, discarding cases where coordinates were not available. For each of these incorrect predictions we then measured the distance between the gold standard location and the location predicted by the system. Figure 2 shows the proportion of incorrect predictions that are located up to 500 km from their target by means of the empirical distribution cumulative function. In all cases, the growth follows a quasi-logarithmic trend. From 15% to 40% of these predictions in individual corpora fall within a radius of no more than 5 kilometers from their targets. Vast majority of these would likely be considered valid responses upon closer inspection, yet these may be found even further: the two large steps (spaCy fishing/*it*/ ~ 70 km and mGenre/*cs*/ ~ 50 km) arise from the "mislabeling" of the first most frequent entity in *it* – Italy (Q38) as the Kingdom of Italy (Q172579) – and the second most frequent entity in *cs* – Czechia (Q213) as Bohemia (Q39193).

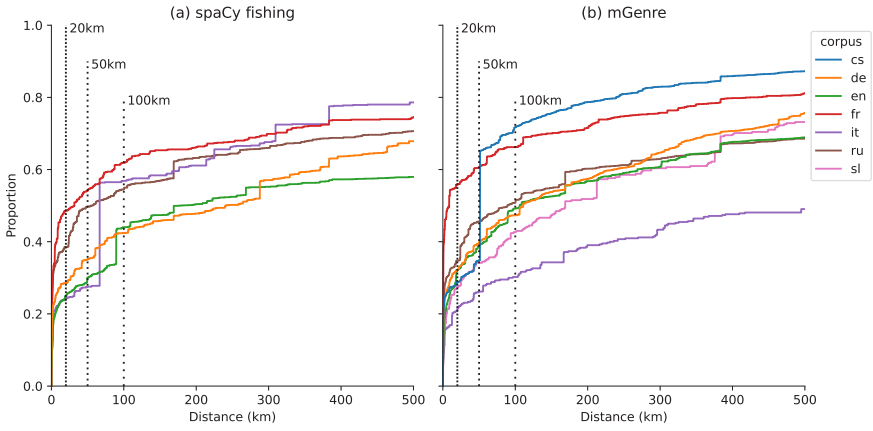

Figure 2: Incorrect responses of NEL systems with known coordinates. Empirical distribution cumulative functions of distances to their target locations

Figure 3 shows that in-KB F_1-scores based solely on linking mismatches may indeed underestimate actual performance. Allowing for a deviation of no more

than 1 kilometer increases scores of both systems by an average of 3%. Extending the tolerance to 20 kilometers results in an average increase of 7.5% for spaCy fishing and 5.5% for mGenre. Although the scores for spaCy fishing remain well below 0.8 across all languages, one can imagine research scenarios in computational poetics where mGenre's precision may be considered sufficient.

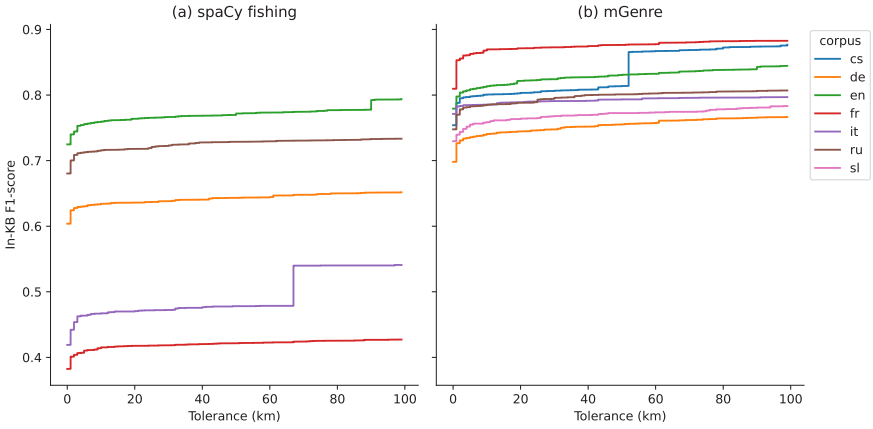

Figure 3: Growth of in-KB F_1-scores when responses located in proximity to their targets are considered correct

3. Conclusion and Future Work

We have provided a detailed evaluation of NER/NEL systems when applied to poetry in seven European languages. While NER performance still falls far short of its in in-domain applications, the results are substantially better than those reported in Foley (2019). mGenre shows promising results in linking geolocations, however, even with a tolerance for proximate matches, its in-KB F_1-scores never exceed 0.9 and surpass 0.8 only in a few languages.

Geolocations annotation has been included in the latest version of the PoeTree dataset and is also accessible via an interactive online map (https://versologie.cz/poetree/map). Besides other research tasks, we aim to use this data to train and release a poetry-specific NEL model in the near future, hoping to achieve a better performance with poetic texts.[2]

[2] Data & code available at https://github.com/versotym/ner_nel_poetree. This article was supported by the Czech Science Foundation (project ga23-07727S).

References

Akbik, Alan; Bergmann, Tanja; Blythe, Duncan; Rasul, Kashif; Schweter, Stefan; Vollgraf, Roland 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In: *NAACL 2019: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis: Association for Computational Linguistics, 54–59. https://doi.org/10.18653/v1/N19-4010

De Cao, Nicola; Wu, Ledell; Popat, Kashyap; Artetxe, Mikel; Goyal, Naman; Plekhanov, Mikhail; Zettlemoyer, Luke; Cancedda, Nicola; Riedel, Sebastian; Petroni, Fabio 2022. Multilingual Autoregressive Entity Linking. In: *Transactions of the Association for Computational Linguistics* 10, 274–290. https://doi.org/10.1162/tacl_a_00460

Foley, John 2019. *Poetry: Identification, Entity Recognition, and Retrieval*. PhD thesis. University of Massachusetts Amherst.

Honnibal, Matthew; Montani, Ines; Landeghem, Sofie Van; Boyd, Adriane 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Marsh, Elain; Perzanowski, Dennis 1998. MUC-7 Evaluation of IE Technology: Overview of Results. In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 – May 1, 1998*. https://aclanthology.org/M98-1002/

Plecháč, Petr; Cinková, Silvie; Kolár, Robert; Šeļa, Artjoms; De Sisto, Mirella; Nugues, Lara; Haider, Thomas; Kočnik, Neža 2024. PoeTree: Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish. In: *Research Data Journal for the Humanities and Social Sciences* 9, 1–17. https://doi.org/10.1163/24523666-bja10044

Straková, Jana; Straka, Milan; Hajič, Jan 2019. Neural Architectures for Nested NER through Linearization. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 5326–5331. https://doi.org/10.18653/v1/P19-1527

Terriel, Lucas; Berenstein, David; Romagnuolo, Giuseppe; Scheithauer, Hugo 2022. spaCy fishing 0.1.8. https://github.com/Lucaterre/spacyfishing

Appendix

F_1-scores reported for systems tested on standard NER datasets were sourced from:

- NameTag.{cs,de,en}: https://ufal.mff.cuni.cz/nametag/2/models
- flair.{de, en}: https://github.com/flairNLP/flair
- flair.fr: https://huggingface.co/flair/ner-french
- spaCy.cs: https://huggingface.co/spacy/xx_ent_wiki_sm
- spaCy.de: https://huggingface.co/spacy/de_core_news_lg
- spaCy.en: https://huggingface.co/spacy/en_core_web_trf
- spaCy.fr: https://huggingface.co/spacy/fr_core_news_lg
- spaCy.it: https://huggingface.co/spacy/it_core_news_lg
- spaCy.ru: https://huggingface.co/spacy/ru_core_news_lg
- spaCy.sl: https://huggingface.co/spacy/sl_core_news_trf