

Clinical Entity Recognition from Electronic Health Records and Linking to Biomedical Knowledge Bases in Low-resource Language Settings

Eno-Martin Lotman

Tallinn University of Technology and
North Estonia Medical Centre
Tallinn, Estonia

Svetla Boytcheva

Sofia, Bulgaria

David Hübner and Kris Collins

Averbis GmbH
Freiburg, Germany

Ivelina Nikolova-Koleva

Ontotext and
Bulgarian Academy of Sciences
Sofia, Bulgaria

Abstract

Clinical Natural Language Processing (NLP) in low-resource languages is severely constrained by strict privacy regulations and a scarcity of localized biomedical ontologies. In Estonian, for instance, the localized SNOMED CT translation covers less than 10% of the international target concept space. In this paper, we present a unified framework for clinical Named Entity Recognition (NER) and Entity Linking (EL) in low-resource settings, evaluated on real-world clinical narratives in Estonian, German, and Dutch. For NER, we deploy a hybrid architecture combining fine-tuned multi-task Transformers, a few-shot generative LLM (Qwen2.5-14B), and dictionary-based lookups. For EL, we implement a two-stage pipeline using deterministic gazetteers alongside SapBERT encoders cross-lingually augmented via SNOMED CT terms and multi-ontology mappings. On clinical datasets, the NER pipeline achieved F1-scores of 0.87 (Dutch), 0.84 (Estonian), and 0.67 (German). However, direct cross-lingual performance comparisons are confounded by substantial variations in corpus size. Compact SapBERT models deliver superior EL precision (0.76) compared to larger multilingual baselines. This work aims to inform future NLP projects by identifying both limitations and promising pathways for advancing low-resource language technologies in biomedical informatics and beyond.

1 Introduction

In natural language processing, a low-resource language is defined not by the size of its speaker base, but by the scarcity of the linguistic infrastructure required for computational work, including annotated corpora, digital lexicons, domain-specific datasets, and pre-trained models. Estonian is considered a low-resource language due to the limited available digital datasets and resources compared to

more widely spoken languages. Clinical documentation in low-resource languages faces unique challenges stemming from both technological and ethical constraints. Privacy regulations and strict data protection laws severely limit the availability and public sharing of authentic clinical language samples, making it more difficult to develop sophisticated NLP tools for these languages. Regulations typically mandate clinical documentation in local languages, reinforcing the need for robust local-language NLP capabilities for improved patient care and safety, but also for maintaining linguistic diversity and precise domain-specific vocabulary into the future. Named Entity Recognition (NER) and Entity Linking (EL)—or entity normalization—are essential for interpreting and utilizing unstructured electronic health records (EHR). By identifying medical terms and mapping them to standardized ontologies, these tasks ensure high data accuracy, semantic interoperability, and effective secondary reuse.

2 Related Work

Clinical NLP in Low-Resource Settings While clinical NLP has achieved remarkable performance in high-resource languages like English, adaptation to lower-resource or morphologically complex languages remains a critical bottleneck. Early clinical systems relied heavily on rule-based patterns and localized dictionaries, which fail to capture natural syntactic variations. The advent of transformer-based models domain-adapted for medicine, such as SapBERT (Liu et al., 2021), revolutionized concept extraction and entity linking (Gallego et al., 2024). However, because these models are predominantly pre-trained on English corpora (like MIMIC-IV or PubMed), direct zero-shot transfer to low-resource languages typically results in steep performance degradation. Consequently, recent

efforts have shifted toward cross-lingual transfer, machine translation of localized ontologies, and parameter-efficient fine-tuning on small, manually annotated clinical datasets (Vassileva et al., 2024).

Estonian General NLP and NER As a Finno-Ugric language characterized by rich morphology, extensive compounding, and free word order, Estonian poses distinct challenges for tokenization and syntactic parsing. The computational foundation for the language was significantly advanced by the introduction of `EstNLTK` (Orasmaa et al., 2016), a Python-based library tailored to handle Estonian-specific morphological complexities.

The Named Entity Recognition in Estonian began with the work of Tkachenko et al. (2013), who developed the first Estonian NER corpus (`EstNER`) using news articles annotated with classic semantic types (person, location, organization) using Conditional Random Fields. This corpus was later expanded and re-annotated to suit neural architectures. The transition to deep learning in Estonian NLP was accelerated by the development of `EstBERT` (Tanvir et al., 2021), a monolingual pre-trained transformer model that consistently outperformed multilingual models, such as `mBERT` (Devlin et al., 2018) and `XLM-RoBERTa` (Conneau et al., 2020), across various syntax and semantic tasks, including general-domain NER.

Clinical NLP and Concept Mapping in Estonian

Compared to general-domain NLP, clinical text processing in Estonian is still in its infancy, largely due to strict privacy regulations governing the Estonian National Health Information System (*Tervise Infosüsteem*). However, a few notable initiatives have paved the way for structured clinical extraction.

Researchers have increasingly focused on processing unstructured clinical narratives, such as epicrisis, using specialized machine learning workflows. For instance, Šuvalov et al. (2023) successfully built an information extraction pipeline for breast cancer clinical records in Estonian, utilizing active learning and human-in-the-loop workflows to train a localized BERT model for TNM clinical staging extraction. Perli (2022) demonstrated that the general-purpose `EstBERT` model cannot match the clinical domain accuracy achieved by a BERT model pre-trained from scratch on clinical notes. Furthermore, the challenge of mapping localized clinical texts to standardized terminologies has been explored through cross-lingual resource

alignment. Since the official Estonian translation of SNOMED CT covers only a fraction of the international version, researchers have focused on mapping local terminologies, ICD-10 classifications, and pharmaceutical registries to global standards like the OMOP Common Data Model (Uusküla et al., 2023) to enable semantic interoperability.

Entity linking Entity linking is the task of connecting named entities mentioned in text to their corresponding resources in a knowledge base or database (e.g., linking "Diabetes" in a sentence to its identifier in a knowledge base like UMLS). This involves assigning a unique database identifier to a text snippet that represents the entity. Several architectures have been proposed for clinical entity linking for low-resource languages. For example, `ClinLinker` (Gallego et al., 2024) combines a two-stage pipeline with initial candidate retrieval using a `SapBERT`-based bi-encoder and subsequent re-ranking with a cross-encoder, following a contrastive learning strategy tailored to medical concepts in Spanish. `ClinLinker` outperformed multilingual `SapBERT` by a substantial margin. The `SympTEMIST` (Vassileva et al., 2024) multilingual linking approach fine-tuned a `RoBERTa`-based token-level classifier with Bidirectional LSTM and CRF layers on an augmented train set. EL was performed via a hybrid approach with dictionaries, generating candidates from a knowledge base using the cross-lingual `SapBERT` and reranking the top candidates using GPT-3.5.

The state-of-the-art EL approaches usually employ two-stage architectures combining candidate generation with transformer-based ranking (Wang et al., 2020). BERT-based and domain-adapted models, such as `SapBERT` and `PubMedBERT` have become foundational, with domain-specific pre-training providing gains of 6.5 points over general BERT initialization (Zhang et al., 2022).

Our work builds on these efforts by evaluating a unified, hybrid NER and EL pipeline designed specifically to tackle the extreme data scarcity of specialized clinical concepts in Estonian.

2.1 Overcoming the Resource Restrictions

Mitigating data scarcity in low-resource clinical NLP requires targeted, privacy-preserving strategies. Modern approaches leverage local language models to generate expert-reviewed synthetic text, alongside hybrid de-identification pipelines to safely anonymize patient data before annotation.

Engaging clinicians and native speakers ensures local linguistic and domain accuracy, while federated learning enables multi-institutional model training directly on local datasets without exposing raw patient data.

2.2 Motivation

The project of which this work is a part employs an array of artificial intelligence tools—including data onboarding assistants, multilingual NLP models (developed in Estonian, Dutch, and German), and entity linking frameworks—combined with an advanced human-in-the-loop curation pipeline. Central to the project is the transformation of disparate patient health data into an integrated Personal Health Knowledge Graph aligned with international standards such as SNOMED CT and HL7 FHIR¹, thereby ensuring semantic interoperability and sustainable data reuse. By automating data quality assessment and annotation, and by publishing reusable benchmark datasets and tools, the project not only addresses technical challenges but also lays the groundwork for the wider adoption of safe, efficient, and standardized health data management in both high- and low-resource language contexts. Consequently, the NLP tools employed to structure health data into a knowledge graph play a major role in driving this transformation.

3 Multilingual Clinical NLP Pipeline

3.1 NER and Temporal Information Extraction

Named Entity Recognition is a fundamental task in Natural Language Processing (NLP) wherein certain spans of text are identified and classified into semantic categories, such as people, locations, or organizations. In the clinical domain, this entails the recognition and classification of disorders, procedures, medications, laboratory values and more.

Although NLP has profited from the deployment of large annotated training corpora and the advent of pre-trained Transformer models, clinical NLP presents unique challenges through specialized language, including ambiguous abbreviations, Latin terminology, institution-specific conventions and domain-specific ontologies. We developed a hybrid NER architecture to address these challenges. Rather than relying on a single approach, we combine three complementary components: fine-tuned multitask Transformer models for high-precision

extraction of well-defined entity types, a generative LLM (Qwen2.5-14B) deployed in a few-shot regime to extend coverage to entities lacking annotated training data, and a dictionary-based lookup for specific, well-enumerated entities linked to SNOMED CT codes.

3.1.1 Hybrid NER Architecture

The NER module is part of a broader clinical NLP pipeline developed for two primary use cases: Breast Cancer and Cardiovascular Disease. The architecture shown on Fig. 1 is designed to recognize spans belonging to SNOMED CT semantic categories, primary among them Body Structure, Disorder, Finding, Medicinal Product, Morphological Abnormality, Observable Entity, Procedure, Product, Qualifier Value and Substance. This system is designed to achieve both high precision on core entity types and adequate coverage across all semantic categories.

Fine-Tuned Transformer Models The foundation of the NER module is a set of multitask Transformer models fine-tuned on annotated clinical data. We utilize a pre-trained multilingual BERT model, fine-tuned for parallel sequence tagging of all semantic tags. Models are trained for 27 epochs (with early stopping) using a batch size of 32, a maximum sequence length of 128 tokens (with an overlap window of 14 tokens to accommodate sliding window inference across longer narratives), and a learning rate of 3×10^{-5} optimized against validation loss.

Generative LLM with Few-Shot Prompting To extract sparse or out-of-vocabulary entities, we employ few-shot prompting with Qwen2.5-14B. The model identifies five core semantic types (Disorder, Finding, Observable Entity, Procedure, Substance) and temporal expressions (in a different request). The 3-shot NER prompt is structured as follows:

System: You are an expert AI assistant for medical information extraction. Your task is to analyze a clinical document (in German, Dutch, or Estonian) and extract key medical concepts.

For each concept you identify, provide a JSON object with the following features:

- **coveredText:** The exact, verbatim text of the concept as it occurs in the document. It should be concise.
- **SemanticTag:** The concept's category, chosen from the five options below.

Semantic Tag Definitions (based on SNOMED CT)

- **disorder:** A disease, diagnosis, or pathological condition. E.g., *Hypertonie* (DE), *hartfalen* (NL, heart failure), *diabeet* (ET).
- **finding:** A sign, symptom, or a specific lab/test result, representing evidence or an observation about a condition. E.g., *Fieber* (DE, fever), *pijn op de borst* (NL, chest pain), *kõha* (ET, cough).

¹<http://hl7.org/fhir/R4/>

- **observable entity:** A characteristic, quality, or state that can be measured or assessed. This represents the parameter being measured. E.g., *Körpertemperatur* (DE, body temperature), *bloeddruk* (NL, blood pressure), *südame löögisagedus* (ET, heart rate).
- **procedure:** A diagnostic or therapeutic action, test, surgery, or the act of making a measurement. E.g., *EKG* (DE), *operatie* (NL, operation), *röntgenuuring* (ET, x-ray), *Blutdruckmessung* (DE, blood pressure measurement).
- **substance:** A drug, medication, or chemical. Only extract the name, no auxiliary information. E.g., *Ibuprofen* (DE), *paracetamol* (NL), *insuliin* (ET, insulin).

Extraction Rules

- **Be Precise:** The `coveredText` must be the exact text. For Diagnose: *akute Bronchitis*, extract *akute Bronchitis*.
- **Separate Listed Items:** If a text lists multiple items like 'Hypertonie und Diabetes', extract them as two separate concepts.
- **No Concepts:** If no valid concepts are found, return nothing.

[Few-shot examples are truncated to save space ...]

Dictionary-based Lookup For entities which have limited lexical variation and well-defined format, like measurements, we use a dictionary-based lookup. This method matches surface strings in the document against a curated dictionary and assigns the corresponding SNOMED CT code. Dictionary lookup is precise, interpretable, and computationally efficient.

Post-Processing: Filtering and Deduplication

Due to the nature of this hybrid approach, the pipeline may produce overlapping, duplicate or conflicting annotations. To minimize these unwanted results, we incorporate two harmonization steps before the final results are output.

- **Confidence-Based Filtering:** Each annotation coming from the Transformer model has a confidence score. To optimize for precision, annotations with confidence scores which don't pass our threshold are discarded. LLM predictions are likewise filtered based on confidence scores produced during Entity Linking.
- **Deduplication:** Overlapping and redundant annotations produced by two different components are resolved through rule-based logic.

3.1.2 NER Results

As summarized in Table 1, the NER pipeline achieved F1-scores of 0.84 on a large Estonian corpus, 0.87 for Dutch, and 0.67 for German. Direct cross-lingual comparisons are confounded by structural differences in the datasets (Tables 1 and 5). Notably, the Dutch corpus contains only 708 entities compared to 11,523 for Estonian, and the

Estonian test set has a near-total concept overlap with its training data (215 of 216 concepts).

A recognized limitation of our current study is the absence of an ablation analysis to isolate the performance gains of each distinct NER module. Unfortunately, strict institutional data protection protocols meant our access to the raw clinical narratives was revoked upon completion of the initial evaluation runs, preventing any subsequent component-wise testing.

3.2 Entity Linking

In alignment with the recent advances in EL, here we present the Estonian EL pipeline in detail and only refer to the German and Dutch ones. The goal is to resolve entities to 207,799 SNOMED CT concepts. Although the approach is concerned only with a SNOMED CT subset, the complexity of the EL task is high as it constitutes a classification problem across more than 207,000 candidate classes.. This scale requires a machine learning approach and a significant amount of training data.

3.2.1 Data

During training, we teach the model that the different labels of the same concept are closer to each other in the model space than the labels of other concepts. For this reason we provide mapping for each of these labels, to the concept ID and feed them to the model. E.g. in the Estonian version of SNOMED CT, the concept with identifier 44054006 has 11 labels:

- 5 labels in Estonian: "*2. tüüpi diabeet*", "*2. tüüpi diabetes mellitus*", "*2. tüüpi meliitdiabeet*", "*2. tüüpi suhkurtõbi*", "*Insuliinisõltumatu suhkurtõbi*" and
- 6 labels in English: "*Diabetes mellitus type 2 (disorder)*", "*Type 2 diabetes mellitus*", "*Diabetes mellitus type 2*", "*Diabetes mellitus type II*", "*T2DM - diabetes mellitus type 2*", "*Type II diabetes mellitus*".

At training time we create exhaustive list of combinations of size 2 from these labels and map them to the concept ID, 44054006 in this case, and supply these mappings as training samples to the model.

Although the project's target concepts span over 20 semantic types, the majority of entities in both SNOMED CT and our training dataset are covered by just four high-priority classes: Measurement (snomed:363787002), Drug (snomed:410942007),

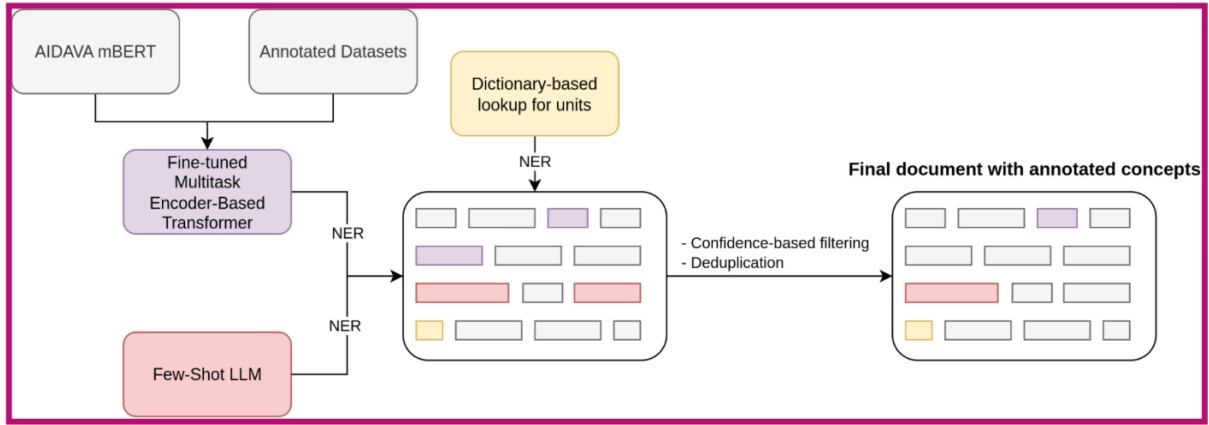


Figure 1: The hybrid architecture for Named Entity Recognition (NER). The system combines three parallel approaches: a fine-tuned Transformer model (mBERT trained on project-annotated data) for high-precision extraction, a few-shot LLM for broad coverage of less common entities, and a dictionary-based lookup for unambiguous units. The outputs are merged and then refined through confidence-based filtering and deduplication to produce the final annotated document.

Table 1: Overview of Named Entity Recognition (NER) performance evaluation across the three target languages at Medical University of Graz (MUG), University of Maastricht (UM), and North Estonia Medical Centre (NEMC).

Site (lang.)	F1	Precision	Recall	TPs	FPs	FNs	Entities
MUG (de)	0.67	0.64	0.69	624	353	274	898
UM (nl)	0.87	0.92	0.83	588	49	120	708
NEMC (et)	0.84	0.84	0.84	9665	1791	1858	11,523

ProblemCondition (snomed:404684003), and Procedure (snomed:71388002).

Terminological sources A natural source of mappings between concept identifiers (IDs) and their corresponding labels is the international edition of SNOMED CT, given that it serves as the target KB for the EL task and offers multiple lexicalizations for many concepts. Among the three target languages in the project, the Dutch translation of SNOMED CT is the most comprehensive. In contrast, the German and Estonian releases contain a significantly limited number of translated concepts, with the Estonian version being the most constrained. Alongside Estonian descriptors, English labels are integrated into the training data. This inclusion enables the processing of narratives that may retain English terminology.

In addition to *<label-to-concept.id>* mappings from SNOMED, more resources are used to extend the coverage of the model. These are included in the training set through inferred mappings between SNOMED codes and UMLS codes, SNOMED and ICD classifications (ICD-9, ICD-10 and ICD-O), SNOMED and ATC. By applying these mappings, the training set for Estonian was

augmented with concepts and labels from ICD-10, EMA drugs register; Estonian Ravimiregister; and MedDRA. Fig. 2 shows an example how the label “Lahtine haav” originating from MedDRA has been added to the Estonian training set.

The contribution of each of the resources to the final training set is shown on Table 2.

Table 2: Model 1 training data.

Source dataset	#concepts	#labels
SNOMED-EN	207,798	647,315
SNOMED-ET	8,980	12,974
ICD-10	5,211	5,283
EMA-DRUGS	161	372
Ravimiregister	534	2,781
MedDRA	12,074	17,479
Machine-translated	194,010	579,389
Manual annotation	264	7,506
Total	208,106	4,488,782

Estonian SNOMED CT scarcity It is worth mentioning that Estonian is less represented in SNOMED than the other project languages - the Estonian SNOMED contains only about 5% of the concepts in the Dutch version and less than 50% of the concepts in the German and Austrian SNOMED together. As shown in Table 3, despite being the

From the inferred mappings:
MEDDRA ID 10058834 == UMLS CUI C0332798 == SNOMED ID 125643001
→
“*Lahtine haav*” (MEDDRA ID 10058834) is a label of the SNOMED concept with ID **125643001** which is not available in the localized ET version of SNOMED therefore it contributes to the training set.

Figure 2: MEDDRA-UMLS-SNOMED mappings

primary source of localized labels, the Estonian edition of SNOMED is highly limited. Indeed, the number of translated concepts represents only 9% of the target concepts. Furthermore, nearly 10% of these share labels with other concepts, introducing significant ambiguity and making model training more challenging.

Table 3: Dataset Statistics: Concept and Label Ambiguity Analysis

Concept Statistics	
Total number of concepts	17,749
Clean concepts (no ambiguous labels)	16,188
Concepts with ambiguous labels	1,561
Label Statistics	
Number of unambiguous labels	16,966
Number of ambiguous labels	783
Ambiguity Distribution	
Max. concepts per ambiguous label	4
Labels mapping to exactly 2 concepts	753
Labels mapping to exactly 3 concepts	29
Labels mapping to exactly 4 concepts	1

Machine-translated Labels To overcome this problem, all concepts in the international version of SNOMED (in English), which were among the project-related concepts were translated automatically to Estonian, using EasyNMT engine² and the opus-mt model³ for English-Estonian. The translation was done locally to preserve privacy signed with SNOMED CT. These labels are marked as Machine Translated on Table 2 and Table 4 visualizing the training data in the EL experiments. As a result more than 194 000 concepts and 579 000 labels were added to the training set of Model 1 and 117 concepts and 426 labels to Model 2, respectively. It is important to mention that these translated labels were manually checked (a subset) and were not included in the indexed KB, which is used for the resolution task. They were only for training.

Manually Annotated Data Manually annotated data is of great importance to creating a good qual-

ity training dataset, as it represents the concepts’ usage in reality with their specific lexicalizations originating from project-relevant documents. For each of the three languages, a manually annotated corpus was provided. The entities in the documents are manually tagged with a semantic type and the most relevant SNOMED concept ID. Table 5 shows the statistics on the manually annotated data. Each annotation is a mention/label of a given concept. Since a concept may have more than one label and a label may occur more than once in the dataset, the number of annotations exceeds the number of labels and concepts. The Estonian dataset has the largest number of documents; however, it covers the lowest number of concepts among the three language datasets, and it’s the only dataset whose training set covers more than 99% of the concepts in the test set. This is a prerequisite for good results on the test set and a low degree of generalization to unseen data/concepts.

The quality and relevance of the training data has a direct impact on the model’s ability to make correct predictions. However, due to the sensitivity of the data in the project domain - healthcare, only limited access to such data was available at development time. It was granted in compliance with the regulations and ethical policies established in the project and in the respective countries.

The contribution of each dataset to the training data for Model 1 in terms of labels and concepts in Estonian is shown in Table 2, and the resources attributed to Model 2 are shown in Table 4.

Table 4: Model 2 training data

Source dataset	#concepts	#labels
SNOMED-EN	183	780
SNOMED-ET	58	84
ICD-10	13	13
MedDRA-ET	30	68
Machine-translated ET	117	426
Manual annotation	259	7,425
MA-context	264	7,506
Total	264	16,302

²<https://easynmt.net/demo/>

³<https://github.com/Helsinki-NLP/OPUS-MT-train>

Table 5: Distribution of the annotations in the manually annotated data.

Site (lang.)	Split	# annotations	# labels	# labels in common with train	# concepts	# concepts in common with train
DE	train	4,087	1,210	—	575	—
	test	415	201	132	142	132
	validation	483	245	152	169	145
ET	train	76,382	7,147	—	264	—
	test	9,688	2,004	1,504	216	215
	validation	9,639	2,157	1,599	220	219
NL	train	5,127	2,106	—	912	—
	test	740	487	206	308	183
	validation	745	447	231	278	198

3.2.2 Approach

The input for the EL module consists of an entity string and its corresponding semantic type. The objective is to return the most appropriate SNOMED concept ID for that entity. Two types of annotators are employed for this task: gazetteers and machine learning models. Gazetteers are tailored exclusively for unambiguous entities; consequently, their confidence score is very high, and they are executed up-front. If a gazetteer fails to find a match, the entity string is passed to a transformer model to generate an embedding. This embedding is subsequently used to search for the nearest concepts within an indexed knowledge base.

Gazetteer The gazetteer comprises all labels from the complete training dataset (Table 2) that are associated with a single concept identifier. Ambiguous labels are strictly excluded from the gazetteer. Consequently, the accuracy of this annotator is exceptionally high—approaching 1.

Models Following on the related approaches, the machine learning model we chose for solving this problem is a transformer model with applied Self-Aligned Pretraining (SapBERT). This is a fine-tuning method that leverages synonymy sets extracted from UMLS to improve BERT’s ability to act as a biomedical entity encoder. It has proven good performance and SotA results on the EL task.

Experiments The final training setup was made after a series of experiments with subsets of training data, split by semantic type, data source and language. Two models were fine-tuned for Estonian: big (model 1) and small (model 2). Both are trained in 3 epochs with 128 batch size, maximum length of 25 and optimized against multi-similarity loss.

Model 1 (big): SapBERT-UMLS-2020AB-all-lang-from-XLMR fine-tuned with the whole training data listed on Table 2.

Model 2 (small): SapBERT-UMLS-2020AB-all-lang-from-XLMR tuned only on the subset of concepts available in the manually annotated data for the corresponding language. This resulted in a smaller dataset, shown on Table 4.

3.2.3 EL Results

The EL module performance is tested on 2 datasets: the test (unseen) and validation (used for development) splits of the Manually annotated dataset. Table 6 evaluates the contribution of gazetteers to the models and compares performance between the large and small model variants. The baseline model—a multilingual SapBERT—is trained on SNOMED CT labels across all three languages, supplemented by Estonian translation equivalents and German labels extracted from hospital patient records. In contrast, the subsequent models are additionally trained on manually annotated data.

The evaluation on the test and validation splits shows similar results; they are both randomly selected from the manually annotated documents and contain similar terminology. The results are comparable on both datasets, except for Dutch, where the results on the validation dataset were higher by more than 10 points. The small models for Estonian and German outperformed the big models (Table 6 Row 2 and 3) – 0.758 vs. 0.27 for Estonian and 0.874 vs. 0.79 for German. For Dutch, however, the big model performed slightly better than the small one (0.59 vs. 0.55 in terms of Precision). The performance of Dutch is surprising, it is worth seeing the number of concepts overlapping in the training and test set in this respect. For Dutch there is the lowest overlap.

Gazetteer impact: it was difficult to estimate the

Table 6: Model Performance Evaluation Across Data Splits and Languages

Data split	Model	Semantic types	ET		DE		NL	
			model + gaze	model only	model + gaze	model only	model + gaze	model only
test	baseline	all	—	0.119	—	0.157	—	0.177
	big	all	0.270	0.270	0.790	0.790	0.590	0.590
	small	all	0.758	0.758	0.874	0.874	0.552	0.543
validation	small	all	0.762	0.762	0.858	0.858	0.6875	0.6807

exact gazetteer contribution to the model performance, because the concepts in the test are almost fully covered by the training set. Dutch is an exception, as the set of concepts in the Netherlands test dataset is covered by the train set only in about 2/3 and probably that’s why the gazetteer shows impact – it succeeds in recognizing some of the remaining 1/3 of the concepts which were not seen by the model. To further prove the gazetteer impact, additional experiments with the Dutch models were carried out on the least represented concepts dataset and the most represented concepts dataset to test the models’ performance on more challenging datasets, which are outside the small model’s training data. On these datasets, the experiments were run with the small and the big model, with and without gazetteers and show that where models are troubled with unseen or less/over-represented concepts, the gazetteers raise the precision with 0.13 to 0.35 points for the big model and between 0.35 and 0.96 points for the small one.

4 Discussion and Conclusion

Here we describe the project’s NER and EL pipelines for the low-resource language Estonian in comparative evaluation with Dutch and German from clinical documentation, which yielded several important findings. (i) Although a low-resource language, the models show good performance for Estonian. (ii) Annotations were performed by common annotation guidelines and task complexity was similar across the three languages. (iii) This study offers valuable insights into the practical implications of working with low-resource languages in clinical NLP, presenting notable strengths and limitations. Among its strengths, the project’s cross-linguistic design, consistent methodology, and deployment across real-world clinical environments allowed for a robust comparison between Estonian, Dutch, and German — languages differing significantly in their NLP resource availability. (iv) Utilizing harmonized annotation protocols and evaluation

criteria minimized bias and enhanced the comparability of results, while the integration of human-in-the-loop curation captured both quantitative and qualitative measures of system performance. (v) Multilingual models trained on the three languages reached comparative results with models fine-tuned on a specific language.

However, there are inherent weaknesses. The project was limited by the scale of data available, potentially affecting the generalizability of findings to broader clinical language use. Additionally, some annotation and evaluation steps relied on domain experts, which, though necessary, introduced subjectivity and potential inter-annotator variability. The marked disparity in digital clinical ontologies and language tools—a defining trait of low-resource status—likely affected downstream processes such as knowledge graph construction and IPS generation, reinforcing the critical bottleneck imposed by limited language resources.

A methodological limitation of our evaluation is that while data access was available, NER and EL were evaluated independently on gold annotations, meaning downstream error propagation from imperfect NER predictions to the EL module remains unmeasured. Additionally, while experiments represent single runs, variance was minimized by employing deterministic transformer-encoder models and setting the LLM generation temperature to zero.

An important next step involves evaluating the developed NLP modules on external datasets to assess their ability to generalize beyond the project corpus.

Acknowledgments

This work is partially supported by European Union’s Horizon Europe research and innovation programme project AIDAVA [Grant Agreement No.101057062]⁴.

⁴<https://www.aidava.eu>

References

- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Multilingual BERT (mBERT). <https://github.com/google-research/bert/blob/master/multilingual.md>.
- F. Gallego, G. López-García, L. Gasco-Sánchez, M. Krallinger, and F. J. Veredas. 2024. [Clinlinker: Medical entity linking of clinical concept mentions in spanish](#). *Lecture Notes in Computer Science*, pages 266–280.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. [Self-alignment pretraining for biomedical entity representations](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.
- Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. [EstNLTK - NLP toolkit for Estonian](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2460–2466, Portorož, Slovenia. European Language Resources Association (ELRA).
- Meelis Perli. 2022. Representation learning on free text medical data. Master’s thesis, University of Tartu, Tartu, Estonia.
- Hendrik Šuvalov, Sven Laur, Tanel Laisaar, and Alan Altraja. 2023. [Information extraction from medical texts with BERT using human-in-the-loop labeling](#). In *Caring is Sharing – Exploiting the Value in Data for Health and Innovation*, volume 305 of *Studies in Health Technology and Informatics*, pages 831–835. IOS Press.
- Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19.
- Alexander Tkachenko, Timo Petmanson, and Sven Laur. 2013. Named entity recognition in Estonian. In *Proceedings of the 4th Workshop on Balto-Slavic Natural Language Processing*, pages 78–83.
- Anneli Uusküla, Heti Pisarev, Kaja-Triin Laisaar, Sven Laur, Kärt Võilo, Helen Task, Reelika Tamm, and Tarmo Meister. 2023. [Transforming Estonian health data to the Observational Medical Outcomes Partnership \(OMOP\) Common Data Model: lessons learned](#). *Database*, 2023:baad080.
- S. Vassileva, G. Graždanski, I. Koychev, and S. Boytcheva. 2024. [Transformer-based approach for symptom recognition and multilingual linking](#). *Database*, 2024.
- Yizhe Wang, Beatrice Daille, and Nabil Hathout. 2020. [A study of semantic projection from single word terms to multi-word terms in the environment domain](#). In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 50–54, Marseille, France. European Language Resources Association.
- S. Zhang, H. Cheng, S. Vashishth, C. Wong, J. Xiao, X. Liu, T. Naumann, J. Gao, and H. Poon. 2022. [Knowledge-rich self-supervision for biomedical entity linking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*.