Exploring automatic terminology extraction from historical medical data

Leonardo Zilio
CENTAL, UCLouvain, Belgium
leonardo.zilio@uclouvain.be

Maria José Bocorny Finatto
PPGLetras, UFRGS, Brazil
mariafinatto@gmail.com

Abstract

This paper analyzes the performance of several terminology extraction methods when confronted with historical specialized texts that do not conform with modern orthographical norms. We tested two extraction methods based on linguistic patterns, four prompt-based generative artificial intelligence (GenAI) models, and one BERT-like model. Some of these models went through fine-tuning for terminology extraction, and one of these is specialized in the extraction of medical terms from documents written in Portuguese. For the GenAI models, we tested four different prompting strategies. As test set, we used chapter fifteen of the second part of the book *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health], originally written in French by G. Mauran at the end of the 18th century, and translated and adapted to Portuguese in 1794. The chapter was manually annotated with terminology, and the evaluation was conducted separately as an f-measure automatic evaluation, as well as a manual precision-based evaluation. This second evaluation method was applied to observe if the automatic extraction methods could complement the original token-based annotation. Results show that using automatic extraction methods to complement the manual annotation can improve coverage, even if individual models do not achieve high extraction quality. By combining two or more models though, a recall of more than 90% could be achieved in the test data.

1 Introduction

Automatic terminology extraction (ATE) is an important Natural Language Processing (NLP) task that serves as basis for several downstream linguistic and computational tasks, such as the lexicometrical analysis and consistent translation of specialized texts. ATE can also further advance research in Digital Humanities, as it contributes to the description and understanding of historical practices in different technical and scientific domains. As it happens with many NLP tasks, most computer tools are not developed to work with historical documents (cf. Quaresma and Finatto, 2020; Vieira et al., 2021; Cameron et al., 2022; Zilio et al., 2022, 2024a). As such, tools that can achieve good performance in modern data might fall short when confronted with historical writing norms. At the same time there is growing interest for the extraction of information from historical medical documents, as can be seen, among other evidences, in the recent appearance of the book *Discursos Médicos no Século XVIII* [Medical Discourses in the 18th Century] (Finatto, 2025).

In this context, this paper sets forth the task of testing a series of off-the-shelf tools to evaluate their performance in ATE using medical data written in 18th-century Portuguese. We evaluated seven tools, ranging from statistical and linguistic ATE tools to large language models (LLMs), including models trained for ATE using medical data. Far from being an attempt at a comprehensive study, this broad range of tools allowed us to start exploring the landscape of ATE for historical documents. Our test data is a single chapter of the medical handbook *Aviso a' Gente do Mar sobre a sua Saude* [Advice to Sea People about their Health], published in 1794, in which terms were manually annotated by a trained linguist.

In our exploration of ATE methods, we show results from two types of evaluation done by two linguists that are trained in the analysis of historical data: one evaluation was based on a manually annotated test data, which generated a list of target terms to be extracted, and a second evaluation was conducted based on the precision of the extraction. This second evaluation was intended to identify elements that were not considered in the original annotation, but that could help in describing important elements of the historical medical context. A third and final evaluation type arose from the com-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 282–292

bination of these other two into a hybrid method.

The main contribution of this paper is the proposed methodologies for ATE evaluation, which include pre-annotation of a test set combined with an independent manual evaluation of the extraction. This method, while not without its faults, allowed us to greatly increase the coverage of the final extraction, and to have a larger test set.

2 Automatic Terminology Extraction

In the *Handbook of Terminology*, Heylen and De Hertog (2015, p. 203) indicate that "an expression's terminological status is often a matter of degree and open to individual variation", so ATE can help with a more objective approach to the selection of term candidates. However, continuing their argument, the authors also mention that "terms are *semantically* defined, as referring to a domain specific *concept*, and the full automatic modelling of semantics is still out of reach for computers". The handbook was written in 2015, before the development of current transformer models in NLP, and it covers ATE methods based mostly on statistics, such as collocation measures, and recurrent linguistic patterns.

Based on the results of TermEval 2020 (Terryn et al., 2020), Heylen and De Hertog (2015) seem to be right about the computer not getting a grasp of the semantics of for ATE. In the competition, a new dataset, ACTER, covering three languages (Dutch, English, and French), was released with terminological annotation, and four teams submitted their ATE tools. The winner team (for English and French) presented two BERT-based models trained on n-gram classification, which achieved f1-scores of 0.467 for English and 0.481 for French. These are very low scores to be able to reliably represent the terminology of a text. In addition, these two models were strictly language-specific.

More recently, with the further development of BERT-like transformers and the release of ALBERTINA-PT (Rodrigues et al., 2023), a new token-based classification model was developed specifically for recognizing medical terms: MediAlbertina (Nunes et al., 2024). This model was trained on named-entity recognition (NER) of medical information, and it had superior performance in comparison with other existing medical NER models, such as BioBERTpt (Schneider et al., 2020), achieving an f-score of 0.832 on the test data. This model was included in our test settings, representing the class of BERT-like, token-based classification models[1]. Because MediAlbertina was trained and fine-tuned on Portuguese data, we expected this model to be able to generalize over the slightly different historical spelling of our data.

With the current access to large language models (LLMs), Senger et al. (2025) developed a methodology to fine-tune LLMs for ATE. This methodology, Distant Supervision for Term Extraction (DiSTER), proved efficient in several datasets, but it still did not not surpass the winner of TermEval in the ACTER dataset. Because LLMs are multilingual by nature, and they are trained on huge amounts of data, they can be used for other languages, even if they were trained for extracting terms only in English. The training of the DiSTER model was also not focused on medical terms, but we expect the LLM to be able to generalize its training to the medical domain.

In this paper, we test whether the semantic definition of terms proposed by Heylen and De Hertog (2015) is "still out of reach for computers", while we acknowledge that there is still a lot of "individual variation" in what terminologists consider a term, which led us to use two complementary evaluation methods. We further complicate matters by using ATE as an umbrella term that also covers named-entity recognition in the specialized medical domain, which we will discuss in following section.

3 Of Terms and Named Entities

In the linguistic definition of terms provided by Cabré (2010, p. 357), terms are seen as "lexical units of language that activate a specialized value when used in certain pragmatic and discursive contexts. The special value results in a precise meaning recognized and stabilized within expert communities in each field". So, as much as terminology is a crucial part of our data, by itself, it would not provide us with sufficient information about the socio-historical context in which historical medical documents were produced.

This means that the information that we would like to extract from the historical sources goes beyond that of the specialized lexical units, and crosses into the territory of named entities and less-specialized lexical units. This moves our research

[1]For reference, we also tested BioBERTpt in our dataset, but we observed that the results were not satisfactory at all, possibly due to the historical spelling present in the data.

in the direction of a hybrid approach to terminology, and into the realm of a textual historical terminology, where not only, for instance, medications, diseases, and treatments are important, but also demographical and geographical data, as well as references to people that were either working on the field or being treated. Having said that, we use the word "term" throughout this paper to refer to our target units, even if we acknowledge the hybridity of our scope.

Because of this different approach to working with terminology in historical data, we could expect that some models, especially those trained on modern medical or specialized data, would fail in extracting demographical or geographical information. This is something that we take into account when prompting LLMs, as we provide them with some examples that lie outside mainstream terminology, and we observe whether they are able to adapt to this new type of information.

4 Corpus

The corpus selected for this study was originally collected in the scope of the project "Corpus Histórico da Linguagem da Medicina em Português (Séculos XVIII-XIX): Terminologia Diacrônica e Humanidades Digitais" [Historical corpus of medical language in Portuguese: Diachronic terminology and digital humanities][2] and it comprises seven chapters of the book *Aviso a' Gente do Mar sobre a sua Saude*. Its content was described in more details in the work of Zilio et al. (2024b).

The writing style in the book does not follow very strict rules for the use of terminology. One of the reasons for that could be that, by the time of writing of the document, many specialized texts were still being written in Latin, so the use of national languages for disseminating scientific knowledge and the scientific genre were still being shaped. This is also reflected in the way that terminology is used in the corpus, where a more stable, specialized vernacular lexicon was still under development for several domains. As such, instead of having precise terms and definitions, the text is written with greater fluidity, and less care for strict textual patterns. For instance, the same term referring to a "greenish stomach content" is described as "materias biliosas , e verdoengas", "materias tirante[s] a verde" e "materias biliosas , e tendentes a verde", as can be seen in these contexts, which were extracted from the test data (the highligths are ours):

	Train data	Test data	Total
Tokens	18482	2774	21256
Types	2850	923	3180

Table 1: Dataset size in types and tokens – observed with AntConc (Anthony, 2004).

> "[...] os doentes tem nauseas , vomitaõ mesmo algumas vezes espontaneamente **materias biliosas , e verdoengas** ; sua lingua se faz negra, e aspera."
> "[...] quando ha nauseas e vomitos de **materias tirante a verde**, he precizo fzer sangrias de dez para doze onças [...]"
> "[...] sobrevem-lhe desejos de vomitar ; e algumas vezes mesmo vomítos de **materias biliosas , e tendentes a verde**; todos estes symptomas chegam ao seu mais alto gráo em menos de vinte e quatro horas [...]"

Here it is important to also mention that we are dealing with a translated text. The original text was written in French by G. Mauran in 1786[3]. In 1794 it was translated and adapted by the High Surgeon of the Royal Portuguese Armada, Bernardo José de Carvalho, who took upon himself the responsibility of converting Mauran's text into a useful medical handbook for Portuguese sailors and ship surgeons. So, for instance, the three variant terminological expressions mentioned above were not variants in the original French, where Mauran consistently used "porracées" for "greenish", and even repeated the term "matières bilieuses & porracées" twice, where the Portuguese translator varied the terminology.

In this paper, we focus only on the Portuguese data, and quantitative details of the sample can be observed in Table 1. The corpus is purposefully split into train and test data, because some of the models needed input from similar information that should not come from the test data (so as to not contaminate the experiments). We thus used the train data to automatically generate linguistic patterns for the extraction with TBXTools and to extract examples of terms that were added to some of the prompts we used with LLMs.

[2]For more information about the transcription and files for download: https://sites.google.com/view/projeto38597/aviso-a-gente-do-mar-1794.

[3]The original title of the handbook was *Avis aux gens de mer, sur leur santé*.

4.1 Data Annotation

The whole corpus was annotated with terminological information using Label Studio (Tkachenko et al., 2020-2025), a Python package that provides a local Web-based annotation interface. The annotation was carried out by one linguist, who is specialized on the subject matter, but without following any annotation guidelines, except for a list of categories that were used to classify the data.

After the annotation was done, the list of terms was revised for annotation errors by the same linguist. This corpus was also annotated with morphosyntactical tags using spaCy's (Honnibal et al., 2020) *pt_core_news_lg* model.

The categories that were used as reference were the following: diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms. One issue arising from this list and the posterior manual analysis based on precision is that this list does not include places. So, while the manual annotation does include actors, places were left out of the list. This was one of the main sources of contributions from the system extractions. Almost all systems extracted names of places, even if not explicitly prompted to do so, because names of places can be considered as part of the information about the population. Later, in the precision-only evaluation, the extracted names of places were validated, as it can also be argued that places are an important source of information from historical data, as shown in the work of Opitz et al. (2026).

5 Methodology

Having described the corpus and the corpus annotation process, this methodology section focus on the tools that were used for (semi-)automatically extracting terms. We also dedicate some space for the second, precision-only manual evaluation process.

5.1 Off-the-shelf tools

We tested a total of seven off-the-shelf tools, some specifically developed for ATE, such as the TBXTools, DiSTER and the MediAlbertina models, and others that were developed with different purposes in mind, but that have features that allow for their use in ATE, such as the Sketch Engine, which provides, among several other features, keyword and term extraction tools, and the MedGemma, Gemma and EuroLLM models, which were developed to complete a text input provided by a user, but that can be prompted to extract terminology from data. Some of the tools mentioned in this subsection were already partially introduced in Section 2, so we will not expand on them as much as on the others.

5.1.1 Pattern-based extraction models

TBXTools (Oliver and Vàzquez, 2015). This tool presents two options for ATE: a linguistic extraction, and a statistical extraction based on n-gram sizes and, potentially, association measures. Because of the small size of the corpus, the statistical extraction feature was not used. The tool provides the means to automatically extract linguistic patterns based on existing annotated data, so we used our train data to extract morphsyntactical patterns. These extracted patterns were then manually revised and cleaned, so this was a semi-automatic extraction. The cleaning was necessary, because the morphosyntactical annotation done by spaCy is not very precise on historical data, and several patterns were spurious. Even with the cleaned patterns, there were still some basic issues with the ATE, as the corpus contained some errors in the recognition of stopwords such as "he" [is], "naõ" [no / not], which were sometimes mistakenly annotated, for instance, as nouns. This happens because of the historical spelling of these words. Having this in mind, after we used the cleaned patterns to extract term candidates from the test data, we used a post-processing python script to ensure that no term candidate would begin or end with stopwords. More details about this process are presented further down in this subsection. The semi-automatic extraction with TBXTools resulted in 331 terms that ranged from bigrams to heptagrams.

Sketch Engine (SkE) (Kilgarriff et al., 2014). Sketch Engine provides two types of extraction: keywords and terms. As such, contrary to TBXTools, SkE can also extract unigram terms. Although keywords and terms have different theoretical implications (the former being salient tokens in a text, and the latter being the carriers of specialized concepts in a given domain), there is usually a good amount of overlap between the two in specialized domains, so we treated the results of the keyword extraction as ATE. Similarly to the TBXTools, for terms longer than unigrams, SkE uses linguistic patterns for ATE, and so it also relies on its own morphosyntactical annotation to extract word combinations that satisfy previously

conceived terminological patterns. As such, for the same reasons explained for TBXTools, we applied a cleaning script at the end, to ensure that no extracted term candidate was a stopword, or had a stopword at its beginning or end. Because SkE also included unigram extraction, it provided us with more term candidates: 873 in total (724 unigrams).

Stopwords. As mentioned for both SkE and TBXTools, we used a custom list of stopwords to remove terms that either were a stopword (in the case of SkE), or started or ended with a stopword. To create this custom list, we extracted the first 150 most common tokens from the train corpus, manually removed a few non-stopwords, and manually added a few singular or plural forms of grammatical words that were already present, such as "elle" [he], "ás" [to the], etc. The resulting list contained 124 words, which were merged with a larger list of contemporary Portuguese stopwords[4]. The final list contained 592 unique items. In addition, due to slight differences in historical orthography, we also replaced all instances of "ão" in the stopword list with "aõ", to increase its coverage.

5.1.2 Large language models (LLMs)

We tested two types of large language models: specialized models, which were fine-tuned for ATE or for working with medical data, and generic models, which have no extra fine-tuning.

DiSTER (Senger et al., 2025). The *DiSTER-Llama-3-8B-Instruct*[5] model is a fine-tuned version of Llama-3-8B model. This is a generative LLM that was specifically fine-tuned for ATE using a combination of datasets from several domains, including biomedicine. All datasets used in the model's fine-tuning were in English, but Llama's multilingual nature was able to provide results for Portuguese as well. We used the default parameters as suggested on the model's Huggingface page.

Gemma 3 (Gemma Team, 2025). The model *Gemma-3-4B-It*[6] is a generic LLM, trained to complete a prompt with the most probable tokens. Only instruction fine-tuning was applied to this model, so we relied solely on the prompting strategies to guide it to perform ATE. Regarding the default parameters presented on the model's Huggingface page, we made two changes: *temperature* was set to 0.0, and *max_new_tokens* was set to 1024.

MedGemma (Sellergren et al., 2025). *MedGemma-4B-It*[7] is a Gemma model that was fine-tuned over several medical datasets, but not necessarily for ATE. Similarly to what we did with Gemma 3, for MedGemma we also only changed two parameters: *temperature* was set to 0.0, and *max_new_tokens* was set to 1024.

EuroLLM (Martins et al., 2025). *EuroLLM-9B-Instruct*[8] is a generic LLM trained on 35 languages. Similarly to Gemma, MedGemma, and DiSTER, it is an instruction-tuned model, which help in our task of prompting it with instructions to extract terminology. Again, we did not perform many changes to the model's parameters in regard to the default settings presented on the Huggingface page, but we did set its *temperature* to 0.01.

Prompting LLMs for ATE. We used four prompting strategies with each of the four models, and thus got 16 different results for the LLM extraction. Due to the slightly different training of the models, the prompts were also slightly different in terms of structure, but the content of the actual prompts was the same. Each prompt also contained a paragraph of the test data, with an average size of 132.29 tokens. There were 24 paragraphs in the test data, so each model was prompted 24 times for each prompting strategy (*i.e.*, 96 times per model). All prompts were written in English, which might have helped in highlighting the paragraphs written in Portuguese as the target of the task.

The strategies used were the following, increasing in information at each step:

- **Zero shot**: this strategy is the one that provides the least information to the system. Apart from mentioning the medical domain and the ATE task, no other information was provided, fully relying on the model's capacity to generate lists of terms.

- **Categories**: in addition to providing information about the domain, in this prompt, a few categories of interest were presented to the model: diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms.

[4]The larger, contemporary Portuguese stopword list was downloaded from https://github.com/stopwords-iso/stopwords-iso.

[5]https://huggingface.co/ElenaSenger/DiSTER-Llama-3-8B-Instruct.

[6]https://huggingface.co/google/gemma-3-4b-it.

[7]https://huggingface.co/google/medgemma-4b-it.

[8]https://huggingface.co/utter-project/EuroLLM-9B-Instruct.

- **One shot**: in this prompt, in addition to the categories, we also provided the system with a single example-term, which was not present in the test data.
- **Few shots**: here we gave the model one example-term per category. None of the example-terms were present in the test data.

The prompts used for each system are presented in more detail in Appendix A.

MediAlbertina (Nunes et al., 2024). MediAlbertina was created by fine-tuning Albertina PT (Rodrigues et al., 2023) for medical named-entity recognition. This is the only model that was trained exclusively on Portuguese data, and later fine-tuned on medical data written in Portuguese. Contrary to the previous generative models, whose task is to complete a prompt, MediAlbertina's task is to classify tokens of a text in a BIO fashion, so it evaluates each token in a text as a potential candidate, and outputs a label, either as a **b**eginning or **i**nternal part of a named entity, or as being **o**utside of named entities (i.e. as not being part of a named entity). This classification is then retrieved as a list of terms (or, to be more precise, as a list of medical named entities).

5.2 Evaluation

Three types of evaluation were carried out: a manual precision-only evaluation, an automatic f-measure evaluation, and an automatic hybrid evaluation.

In the precision-only evaluation, the list of extracted term candidates was analyzed by a linguist, and the candidates were classified as terms, non-terms, or partial terms. The classification was supported by contexts of occurrence (*i.e.*, the paragraphs from where candidates were extracted), and the occurrence as a term in at least one context would suffice for the classification of the candidate as a valid term. In total, 3,208 candidates were evaluated this way. The initial evaluation was semi-automatically revised by a second linguist, focusing on cases of inter-model annotation disagreements.

The second evaluation method, an f-measure evaluation, was carried out automatically, based on the annotated test data, as detailed in Subsection 4.1. The test data contained 193 terms. When contrasted with the precision-only annotation, another 252 unique terms were added to this list, which was then used in our final evaluation method: a hybrid evaluation, combining the annotated terms with the precision-only terms that were evaluated as valid. This goes to show that a single annotator, who is specialized in the topic, was able to cover around 43.37% of the information. This percentage rises to 61.12% if we consider that 79 of the new terms were part of the 193 annotated terms (that is, out of the 252 added terms, 79 were actually part of a longer term already present in the annotated data). This happened because, during the annotation process, the longest term would be selected, and any internal, shorter terms would not be individually annotated. We also have to consider that there are terms in the data that were potentially left out by the human annotator and by all the systems as well.

This process allowed us to observe not only the quality of the models used for ATE, but also the coverage that can be expected from human annotation, and how it can be improved by adding automatic tools in the process.

6 Results and Discussion

Table 2 presents the results divided by type of model: pattern-based extraction models (PB model), generative artificial intelligence models (GenAI models) and the single token-based classification model (TBC model). For each of the GenAI models, we also included subscripted information about the type of prompt used for obtaining the indicated results: *zero* stands for zero-shot prompting, *ner* represents the category-based approach, *one* refers to the one-shot approach, and *few* indicates the few-shot approach.

The table also contains information about how many unique term candidates each model extracted, and about how many new validated terms each model contributed to the hybrid evaluation. In the precision-only evaluation, the terms that were considered as partially correct were considered as correct in the "lenient precision" column. The top results for each evaluation category are highlighted in bold. There is no information about significance, because all models are deterministic or very close to deterministic[9].

The Gemma-family of models provided the best f-measure scoring, when we consider the hybrid evaluation (*i.e.*, annotation + precision-only). The

[9]The GenAI models, even at very low or zero temperature, still present minor fluctuations in their outputs, which are explained in (He and Lab, 2025). We consider that these minor fluctuations would not suffice to warrant several prompting attempts, in order to establish a mean and standard deviation as proposed by (De Pourcq et al., 2025).

Models	Precision-only Evaluation			Annotation Evaluation			Hybrid Evaluation			
	Extracted Candidates	Strict Precision	Lenient Precision	Precision	Recall	f-measure	New Terms	Precision	Recall	f-measure
PB models:										
Sketch Engine	873	0.2726	0.4742	0.1661	**0.7513**	0.2720	102	0.3253	**0.6382**	0.4310
TBXTools	331	0.4139	0.6344	0.1722	0.2953	0.2176	83	0.4381	0.3258	0.3737
GenAI models:										
DiSTER$_{zero}$	156	**0.7244**	**0.7821**	0.5000	0.4041	**0.4470**	38	**0.7564**	0.2652	0.3927
DiSTER$_{ner}$	99	**0.7172**	0.7677	**0.5556**	0.2850	0.3767	18	**0.7475**	0.1663	0.2721
DiSTER$_{one}$	97	**0.7629**	**0.7835**	**0.5876**	0.2953	0.3931	18	**0.7732**	0.1685	0.2768
DiSTER$_{few}$	128	0.6094	0.7109	0.4688	0.3109	0.3738	22	0.6719	0.1933	0.3002
EuroLLM$_{zero}$	357	0.3950	0.5854	0.2241	0.4145	0.2909	67	0.4370	0.3506	0.3890
EuroLLM$_{ner}$	365	0.4164	0.6082	0.2055	0.3886	0.2688	82	0.4548	0.373	0.4099
EuroLLM$_{one}$	320	0.4531	0.6344	0.2500	0.4145	0.3119	72	0.5031	0.3618	0.4209
EuroLLM$_{few}$	354	0.3927	0.5791	0.2062	0.3782	0.2669	74	0.4407	0.3506	0.3905
Gemma$_{zero}$	348	0.4511	0.6178	0.2759	0.4974	0.3549	66	0.4856	0.3798	0.4262
Gemma$_{ner}$	287	0.6132	**0.7909**	0.3659	0.5440	0.4375	76	0.6516	0.4202	**0.5109**
Gemma$_{one}$	285	0.6070	0.7719	0.3895	0.5751	**0.4644**	69	0.6526	0.418	**0.5096**
Gemma$_{few}$	260	0.6462	**0.8000**	0.3962	0.5337	**0.4547**	70	0.6808	0.3978	**0.5021**
MedGemma$_{zero}$	618	0.2670	0.4337	0.1553	0.4974	0.2367	80	0.3026	0.4202	0.3518
MedGemma$_{ner}$	365	0.5370	0.6959	0.3288	**0.6218**	**0.4301**	84	0.5863	**0.4809**	**0.5284**
MedGemma$_{one}$	344	0.5581	0.7413	0.3372	**0.6010**	**0.4320**	83	0.6017	**0.4652**	**0.5247**
MedGemma$_{few}$	389	0.4730	0.6247	0.2725	0.5492	0.3643	84	0.5090	0.4449	0.4748
TBC model:										
MediAlbertina	18	**0.7778**	**0.9444**	**0.5000**	0.0466	0.0853	5	**0.7778**	0.0315	0.0605

Table 2: Results of the manual precision-only evaluation, of the automatic f-measure evaluation based on the manually annotated test set, and of the automatic hybrid evaluation.

non-adapted version had slightly lower scores, but, given some basic information, such as the categories of interest, and possibly a single example, these models were able to achieve a good balance of number of extracted candidates, precision and recall. If the aim of the task is, however, to extract fewer candidates that are more precise, then the DiSTER model toped the table, achieving up to 77.32% precision, but with an extraction of only 97 terms. Usually, however, in tasks like ATE, the focus is on the extraction of as many terms as possible, and, in terms of recall, no system was better than SkE. It did extract a lot of candidates, which warranted the lowest score in precision in the trade-off, but, with a score of 63.82%, it toped the list in recall.

Considering that no system by itself achieved a high score, we tested combinations of two and three models, focusing on improving precision, recall and f-measure with two models, and then purely recall and f-measure with three models. These combinations were tested directly on the hybrid test set, that is, the one that emerged from the combination of the two manual evaluations. Tables 3 and 4 show results for combinations of two and three models, respectively. For the lack of space, not all possible permutations are shown in these tables, which focus on the most relevant results. In Table 4 we did not highlight any results for precision, because the best results in the combination of three models was 74.82% over 139 unique terms, when combining DISTER$_{ner}$ + DISTER$_{one}$ + MediAlbertina, but this result was already inferior to DISTER$_{one}$ by itself (see Table 2), and f-measure of the combo was much lower than the other combinations, at 35.62%.

The combinations of different models was a very promising path to explore. It is expected that, by joining different models, some will be complementary and achieve better results together, especially in terms of recall and f-measure. Here we could see that the combination of the two pattern-based extraction models, SkE and TBXTools toped the table in recall, even if TBXTools did not excel in any front, it was the only model that complemented SkE's extraction in such a way as to increase recall by almost 20 percentage points. Precision was not very high, as both models together extracted 1121 unique terms, but the f-measure was much better than many models by themselves. And when these two models were joined by MedGemma or by EuroLLM, the recall jumped above 90 percentage points, while still keeping an f-measure at around 48 points.

Combination	Unique Terms	Precision+	Recall+	f-measure+
DiSTER$_{ner}$ + Gemma$_{few}$	298	0.6443	0.4315	0.5168
DiSTER$_{one}$ + Gemma$_{few}$	298	0.6510	0.4360	0.5222
DiSTER$_{zero}$ + Gemma$_{few}$	317	0.6562	0.4674	0.5459
DiSTER$_{zero}$ + MedGemma$_{ner}$	411	0.5888	0.5438	**0.5654**
EuroLLM$_{ner}$ + TBXTools	685	0.4409	0.6787	0.5345
Gemma$_{few}$ + MediAlbertina	270	**0.6815**	0.4135	0.5147
Gemma$_{ner}$ + TBXTools	579	0.5095	0.6629	**0.5762**
Gemma$_{one}$ + MediAlbertina	291	0.6564	0.4292	0.5190
MedGemma$_{few}$ + TBXTools	675	0.4519	0.6854	0.5446
MedGemma$_{ner}$ + TBXTools	641	0.4805	0.6921	**0.5672**
MedGemma$_{one}$ + TBXTools	617	0.4830	0.6697	**0.5612**
SkE + TBXTools	1121	0.3318	**0.8360**	0.4751

Table 3: Results for model combinations using the hybrid evaluation test set.

Combination	Unique Terms	Precision+	Recall+	f-measure+
DiSTER$_{zero}$ + Gemma$_{ner}$ + TBXTools	627	0.5088	0.7169	**0.5951**
DiSTER$_{zero}$ + Gemma$_{one}$ + TBXTools	623	0.5072	0.7101	**0.5918**
EuroLLM$_{ner}$ + SkE + TBXTools	1253	0.3256	**0.9169**	0.4806
EuroLLM$_{one}$ + SkE + TBXTools	1230	0.3285	**0.9079**	0.4824
MedGemma$_{ner}$ + SkE + TBXTools	1213	0.3331	**0.9079**	0.4873
MedGemma$_{few}$ + SkE + TBXTools	1242	0.3285	**0.9169**	0.4837
MedGemma$_{one}$ + SkE + TBXTools	1215	0.3342	**0.9124**	0.4892

Table 4: Results from the combination of three models using the hybrid evaluation test set.

6.1 Of Models and Hallucinations

It is known that GenAI models can produce spurious results, usually referred to as hallucinations. They can generate outputs that are not real, or that do not correspond to the given task. What we saw in our data was that, apart from TBXTools and MediAlbertina, which would only produce hallucinations in the form of false positives, all the models had their own ways of producing hallucinations, not only the the LLMs.

The Gemma-family models sometimes "corrected" words present in the document. For instance, "vomitos" would sometimes be modernized to "vômitos" [vomit] in the extraction, "emulssaõ" was modified to "emulssa", which is not a word in Portuguese, and "respiraõ" was modified to "respira", a different form of the verb "respirar" [to breathe]. These modifications were accepted in the lenient precision-only evaluation (as partial matches), but not in the f-measure calculations.

EuroLLM also produced alterations in the spelling of extracted data, similar to Gemma. In addition, it would sometimes get into a loop, where it would repeat a word up to the maximum number of generated tokens. That's why it featured, for instance, the pronoun "tudo" [everything] 223 times, and "primeira" [first$_{feminin}$] 230 times in its outputs.

The DiSTER model frequently extracted information directly from the prompt, instead of extracting term candidates only from the target paragraph, especially in cases where the target paragraphs did not have any terms to be extracted. As such, the output would frequently contain the categories of interest (preserved in English) or the example-terms indicated in the prompt (even if they were not present in the test data). The zero-shot model was the only one not to produce such outputs.

SkE does not generate new output by itself, but, because it relies on lemmatized data to extract terms, and because its lemmatizer is not trained on historical Portuguese data, some candidates were extracted with bad grammatical agreement, such as "alimento mais succosos" [juicier$_{plural}$ food] and "remedios administrado" [administered$_{singular}$ medications]. As it happened with Gemma, when

these could be considered terms, they were considered correct in the lenient precision-only evaluation, but not when calculating the f-measure.

Even though MedGemma and EuroLLM had some hallucinations in the same way as the other LLMs did, they also provided some interesting results that we were not expecting to see. They were able to leverage their generating powers to combine words that were separated in the text, but that belonged together as a term. For instance, in the context "o ventre esta tenso , e dorido" [the abdomen is tense, and hurting], EuroLLM was able to extract "ventre dorido", and, in the context "quando a sede he urgente" [when the thirst is pressing], MedGemma joined together "sede urgente". These cases were rare, with five occurrences for EuroLLM and three for MedGemma, but they were considered as correct extractions and were added to the hybrid f-measure evaluation. DiSTER also showed potential for doing this, but it happened only twice, in the same context, as it extracted "vinho de Alicante" and "vinho de Chipre" from the context "algumas colheres do vinho de Malga , de Chipre , de Alicante" [some spoons of Malaga, Cyprus, or Alicante wine].

7 Final Remarks

In general, the models' ATE performance was not bad. We cannot draw a direct parallel with TermEval data, as we are working with a completely different dataset and evaluation method, but the f-score of 0.5951 for the combination of DiSTER$_{zero}$ + Gemma$_{ner}$ + TBXTools can be taken as a promising result. However, it was a surprise to see that, even with all the developments in neural and GenAI models, the pattern-based models still performed better in recall, which is arguably the most important metric for a terminologist, and arguably also the most important metric for us in the analysis of historical information. TBXTools and SkE, if not very performant by themselves, provided a great complement to one another and to other GenAI models, as they were the most recurrent models in the best combinations.

The use of two evaluation methods, which then generated a third evaluation, seemed to be a good approach for this experiment. By combining a token-based annotation with a precision-only evaluation, we were able to highlight how much models can contribute to a human annotation, and to show that neither a single human nor a single model can achieve high f-measure in ATE, even if the human is more precise in their annotations.

The outputs of GenAI models also presented some cause for concern, as EuroLLM and both Gemma-based models produced alterations in the spelling extracted candidates. Alterations and modernizations of spelling can have a significant impact in the description of historical texts and in the compilation of terminologies that were still being consolidated at the time.

Overall, the results of the experiments reported in this paper give us more confidence moving forward to the analysis of the remaining chapters in the dataset. By focusing on automatically extracted data, combined with a more detailed human analysis of the contexts, we can extract valuable information from the historical medical data that can be used to describe past medical practices and to further advance the field of Digital Humanities.

Limitations

One of the main limitations of this paper was its scope. Due to the amount of data that needed to be annotated and evaluated, we could not work with a larger dataset, and had to settle for a single chapter of the medical handbook as test sample.

A second limitation was the inexistence of annotation guidelines or of more annotators. We could not evaluate, for instance, how human annotators would agree on the annotation of terminological information that goes beyond the more strict boundaries of terms.

Perhaps more a trade-off than an actual limitation was the slight difference, in comparison to the literature, in the way we evaluated the tools. We did not perform a token-based evaluation, but rather focused on a type-based evaluation, where lists of terms were compared, instead of lists of frequencies or precise token positions. This approach was different from what is usually found in the literature. For instance, in Terryn et al. (2020), each occurrence of a type was taken into account for calculating the f-score. In our evaluation, even a single automatic extraction of a term that could occur ten times in the test set would result in an f-score of 1 for that term, but it also meant that terms that are very frequent would be treated in the same way as rarer terms. In this way, a system that would detect multiple occurrences of a frequent term, while letting slide a rarer, single-occurrence term, would be penalized by 50%.

Acknowledgments

This research was supported in Belgium by the Wallonia-Brussels Federation's Special Research Fund (ILC FSR24) and in Brazil by the National Council for Scientific and Technological Development (CNPq), grant PQ 307088/2023-5, PIBIC-CNPq-UFRGS, FAPERGS - Edital 06/2025, and TILD-IAR-CNPq (grant 408490/2024-1).

References

Laurence Anthony. 2004. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. pages 7–13.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Helena Cameron, Fernanda Olival, Renata Vieira, and Joaquim Santos. 2022. Named entity annotation of an 18th-century transcribed corpus: problems and challenges. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 18–25. CEUR.

Lena De Pourcq, Marie Gregoire, and Leonardo Zilio. 2025. Exploring the power of generative artificial intelligence for automatic term extraction from small samples. In *Electronic lexicography in the 21st century (eLex 2025) Intelligent Lexicography. Proceedings of the eLex 2025 conference*, pages 116–138. Lexical Computing CZ s.r.o.

Maria José Bocorny [Org.] Finatto. 2025. *Discursos Médicos no Século XVIII: genealogia de saberes e conhecimentos através da linguagem*. Editora da ABRALIN.

Gemma Team. 2025. Gemma 3.

Horace He and Thinking Machines Lab. 2025. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Kris Heylen and Dirk De Hertog. 2015. Automatic term extraction. *Handbook of terminology*, 1(01).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine. *Lexicography*, 1(1):7–36.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Miguel Nunes, João Boné, João C Ferreira, Pedro Chaves, and Luis B Elvas. 2024. Medialbertina: an european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

Antoni Oliver and Mercè Vàzquez. 2015. Tbxtools: A free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the international conference recent advances in natural language processing*, pages 473–479.

Juri Opitz, Corina Raclé, Emanuela Boros, Andrianos Michail, Matteo Romanello, Maud Ehrmann, and Simon Clematide. 2026. Clef hipe-2026: Evaluating accurate and efficient person-place relation extraction from multilingual historical texts. *arXiv preprint arXiv:2602.17663*.

Paulo Quaresma and Maria José Bocorny Finatto. 2020. Information extraction from historical texts: a case study. In *Proceedings of the Workshop on Digital Humanities and Natural Language Processing, co-located with International Conference on the Computational Processing of Portuguese, DHandNLP@PROPOR, Evora, Portugal, March 2, 2020.*, pages 49–56. CEUR.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Elena Senger, Yuri Campbell, Rob Van Der Goot, and Barbara Plank. 2025. Crossing domains without labels: Distant supervision for term extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1366–1378, Suzhou (China). Association for Computational Linguistics.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. 2021. Enriching the 1758 portuguese parish memories (Alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

Leonardo Zilio, Maria Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the XVIII century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Fortaleza, Brazil, 21st March, 2022*, pages 1–10. CEUR.

Leonardo Zilio, Rafaela R Lazzari, and Maria José B Finatto. 2024a. Can rules still beat neural networks? The case of automatic normalisation for 18th-century Portuguese texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 83–92.

Leonardo Zilio, Rafaela Radünz Lazzari, and Maria Jose Bocorny Finatto. 2024b. NLP for historical Portuguese: Analysing 18th-century medical texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 76–85.

A Query Structures and Prompts

There were, in total, three different query structures and four prompting strategies. So, for instance, the query structure used for the DiSTER model was the following:

```
'{"id": "test_0", "conversations": [\n'
'    {"from": "human", "value": "Text: ' + processed_line + '"},\n'
'    {"from": "gpt", "value": "I\'ve read this text."},\n'
'    {"from": "human", "value": "' + prompt + '"},\n'
'    {"from": "gpt", "value": ""}\n'
']}\n'
```

Where "processed_line" was a variable containing the current paragraph of the test set, and "prompt" was a variable containing the following string, for the zero-shot strategy:

> Extract a single list of medical terms from a short text. The terms can range from unigrams to n-grams. The output should only contain the list of terms, in a format that can be directly read as a Python list.

This query structure followed the one available on the model card on Huggingface.

For EuroLLM and the Gemma-family models, the query was very similar, with just a slight change. Here is the query structure for EuroLLM as an example:

```
{
    "role": "system",
    "content": "You are an expert terminologist that works with \
        18th-century medical documents written in Portuguese.",
},
{
    "role": "user",
    "content": current_prompt
}
```

Again, the "current_prompt" here contained the full prompt, such as this, for zero shot:

> Extract a single list of medical terms from a short text. The terms can range from unigrams to n-grams. The output should only contain the list of terms, in a format that can be directly read as a Python list. Here is the text: '##########'

Where "##########" is just a placeholder that would be replaced with the actual paragraph from the test set.

As these examples show, the queries for DiSTER were slightly different due to the different query structure, but the content of the instruction prompt was the same.

For approaches with information about categories, the following string was added after "n-grams":

> , and might include diseases, diagnostics, symptoms, treatments, medications, ingredients, body parts, actors, information about the population, and general medical terms.

This was further complemented with:

> Example of term (not included in the sample): peripneumonia.

for the one-shot strategy, and

> Examples of terms (not included in the sample): peripneumonia, pulsaçaõ, amollecido, panada nutriente, mel, intestinos, Professores, robusto, tratamento.

for the few-shot strategy.