How Important Is Tokenization in French Medical Masked Language Models?

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickaël Rouvier, Richard Dufour

How Important Is Tokenization in French Medical Masked Language Models?

Yanis Labrak[1,2], **Adrien Bazoge**[3]
Béatrice Daille[3], **Mickael Rouvier**[1], **Richard Dufour**[1,3]
[1] LIA, Avignon University [2] Zenidoc
[3] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
{first.last}@{univ-avignon.fr, univ-nantes.fr}

Abstract

Subword tokenization has become the prevailing standard in the field of natural language processing (NLP) over recent years, primarily due to the widespread utilization of pre-trained language models. This shift began with Byte-Pair Encoding (BPE) and was later followed by the adoption of SentencePiece and WordPiece. While subword tokenization consistently outperforms character and word-level tokenization, the precise factors contributing to its success remain unclear. Key aspects such as the optimal segmentation granularity for diverse tasks and languages, the influence of data sources on tokenizers, and the role of morphological information in Indo-European languages remain insufficiently explored. This is particularly pertinent for biomedical terminology, characterized by specific rules governing morpheme combinations. Despite the agglutinative nature of biomedical terminology, existing language models do not explicitly incorporate this knowledge, leading to inconsistent tokenization strategies for common terms. In this paper, we seek to delve into the complexities of subword tokenization in French biomedical domain across a variety of NLP tasks and pinpoint areas where further enhancements can be made. We analyze classical tokenization algorithms, including BPE and SentencePiece, and introduce an original tokenization strategy that integrates morpheme-enriched word segmentation into existing tokenization methods.

Keywords: Tokenization, Morphemes, Language Model, Biomedical, SentencePiece, BPE, RoBERTa, Transformers

1. Introduction

Word tokenization into subword units is a longstanding challenge in the field of natural language processing (NLP), initially conceived to address out-of-vocabulary words in language modeling (Larson, 2001; Bazzi and Glass, 2002; Szoke et al., 2008). In recent years, this strategy of splitting words into smaller units has gained prominence, primarily driven by the widespread adoption of pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020). This shift began with statistical tokenizers, particularly Byte-Pair Encoding (BPE) (Gage, 1994), as introduced in BERT (Sennrich et al., 2016). It was later extended by other data-driven variants, such as SentencePiece (SP) (Kudo and Richardson, 2018) and WordPiece (Devlin et al., 2019).

While empirical evidence consistently demonstrates that subword tokenization outperforms character and word-level tokenization (Wang et al., 2017; Wu et al., 2016), the precise reasons behind this success remain not fully understood. Some studies have explored the impact of segmentation granularity on subword performance (Samuel and Øvrelid, 2023; Novotný et al., 2021), suggesting that each task or language may have its own optimal granularity for maximizing performance. However, other factors, such as the influence of data sources used to construct tokenizers or the role of morphological information, require more comprehensive investigation.

In the context of Indo-European languages, particularly in French, words are composed of a series of morphemes[1] (Touratier, 2012). These morphemes can be categorized as either lexical or grammatical and are analogous to the subword units idea previously mentioned, but here adhere to well-defined linguistic rules.

In specialized domains, such as medical one, meaningful morphemes follow construction rules from Greek and Latin languages. These rules help medical professionals deduce the meanings of unfamiliar terms and remember complex terminology effectively. Despite the agglutinative nature of biomedical terminology, existing PLMs do not explicitly integrate this knowledge into their tokenization processes since they only rely on statistical tokenizers (BPE, SentencePiece, etc.). As a result, common terms are inconsistently tokenized into arbitrary subwords in these models.

In this paper, we investigate the impact of word tokenization strategies in the French biomedical domain and their effectiveness on downstream NLP tasks. Our study delves into the nuances of different tokenization algorithms, aiming to understand why subword tokenization strategies, such as BPE and SentencePiece, outperform other methods (Kudo, 2018). We also identify areas for further

[1] In linguistics, a morpheme is defined as the smallest unit of meaning within a word.

optimization and provide a comprehensive analysis of their performance on a large set of 23 diverse French biomedical NLP tasks, such as named entity recognition (NER), multi-label classification (CLS), or semantic textual similarity (STS). Finally, we propose an original tokenization strategy that integrates morpheme-enriched word segmentation into existing tokenization algorithms. The latter is included in the comparison of tokenizers and makes it possible to study the contribution of subword units constructed from linguistic rules.

Our contributions are as follows:

- We introduce an original tokenization strategy that integrates manually defined morphemes into statistical tokenization algorithms.
- We analyze the ability of statistical tokenizers to segment words regarding their real linguistic segmentation.
- We provide both qualitative and quantitative analyses to assess how word tokenization approaches (statistical methods vs. morpheme-enriched variants) and the data source on which they are trained impact the performance of BERT-based language models.
- We explore the relationship between tokenization granularity and its impact on performance in various downstream NLP tasks.

The morpheme-enriched tokenization strategy, experiment reproduction scripts, and resulting BERT-based PLMs are freely available under the MIT license on GitHub and Hugging Face[2]

2. Related works

Recent research into domain-specific language models has shown that utilizing specialized data during pre-training significantly enhances model performance in that domain. Various strategies, with varying proportions of in-domain and out-of-domain training data, have been proposed across diverse fields, including biomedicine (Lee et al., 2019; Gu et al., 2021; El Boukkouri et al., 2022; Labrak et al., 2023), scientific research (Beltagy et al., 2019) and clinical (Alsentzer et al., 2019).

In the context of biomedical-specific language models, it is widely recognized that training models from scratch using in-domain corpora (Gu et al., 2021) yields noticeable performance improvements compared to other pre-training strategies. The authors also demonstrated the benefits of using domain-specific tokenizers generated through conventional statistical tokenization construction techniques, such as WordPiece (Schuster and Nakajima, 2012), SentencePiece (Kudo and Richardson, 2018), and BPE (Sennrich et al., 2016), on an in-domain corpus, resulting in improved performance in downstream tasks.

Although statistical-based tokenization algorithms are the predominant method employed in recent biomedical language models, some studies have raised questions about the effectiveness of this approach and its suitability for specific downstream tasks or languages (Mielke et al., 2021; Novotný et al., 2021). As a result of these inquiries, various methods for improving tokenization have emerged, some involving training models from scratch (Kudo, 2018), while others do not (Hofmann et al., 2022; Fan and Sun, 2023). One such method involves incorporating linguistic knowledge during the tokenization process by utilizing morphemes (Fujii et al., 2023; Pan et al., 2020; Chen and Fazio, 2021; Toraman et al., 2023), with the aim of mimicking how humans learn and understand languages. However, there have been fewer contributions in the context of biomedical domains and Indo-European languages (Jimenez Gutierrez et al., 2023), despite these fields being highly dependent on an agglutinative terminology.

3. Tokenization Strategies

In this section, we provide a brief overview of the two studied statistical-based tokenization approaches (Section 3.1), followed by the description of our original approach that integrates linguistic knowledge through morphemes into existing tokenizers algorithms (Section 3.2).

3.1. Statistical Tokenization Algorithms

In this study, we compare two statistical-based tokenization methods, BPE and SentencePiece. BPE begins with individual characters and progressively combines them into subword pairs based on their frequency in the training data. In contrast, SentencePiece employs two subword segmentation algorithms, Unigrams and BPE, offering flexibility in terms of segmentation granularity. While SentencePiece is widely used in French biomedical models (Touchent et al., 2023; Labrak et al., 2023; Copara et al., 2020; Berhe et al., 2023), its appropriateness for a specific language and domain may vary, potentially leading to suboptimal subword segmentation.

3.2. Morpheme-enriched Tokenization

In our study, focusing on improving the modeling of specialized medical terminology in the medical

[2]Repositories are currently private to respect the double-blind review process, but an anonymized version is available: https://anonymous.4open.science/r/BioMedTok-D665/

field and reducing the impact of unseen words during model pre-training, our primary emphasis is on lexical morphemes (Touratier, 2012). To achieve this, we created a manual list of around 600 frequently used lexical morphemes in the French medical domain, sourced from the book by Cottez (1980). Examples of these morphemes include terms like `céphal-`, `clinico-`, `-thérapie`, `thoraco-`, `-ome` and `-gène`.

We trained our morpheme-enriched tokenizers by modifying both the BPE and SentencePiece algorithms. During training, we introduced a predefined list of language-specific morphemes as tokens. These morphemes were enforced selections by the tokenizer when encountered, while the remaining text underwent the standard tokenization process of the chosen algorithm. This approach enabled us to combine traditional BPE and SentencePiece tokenizations with morpheme tokens, mitigating issues related to unseen words during training.

4. Experimental Protocol

In this section, we outline the experimental approach used to evaluate the impact of tokenization strategies on French biomedical PLMs. Firstly, in Section 4.1, we present the set of 23 selected biomedical NLP downstream tasks used in our study. Next, we describe the different training data sources employed to train the statistical tokenizers in Section 4.2. Following this, in Section 4.3, we explain the training procedure for the chosen BERT-based model architecture. Finally, in Section 4.4, we provide a comprehensive description of the evaluation methodology used to assess the performance of these models.

4.1. Downstream Tasks

We summarize the datasets of the 23 NLP biomedical downstream tasks from DrBenchmark (Labrak et al., 2024), including NER, part-of-speech (POS) tagging, STS and classification.

DEFT-2020 (Cardon et al., 2020) is a dataset featured in the 2020 edition of the annual French Text Mining Challenge, known as DEFT. It encompasses clinical cases, encyclopedia, and drug labels, all of which have been annotated for two specific tasks: (i) assessing textual similarity and (ii) performing multi-class classification. The first task is geared towards determining the degree of similarity between pairs of sentences, with a scale ranging from 0 to 5 and involves 1,010 sample pairs. The second task involves identifying, for a given sentence, which among three provided sentences is the most similar. There are 1,102 samples included in this task.

DEFT-2021 (Grouin et al., 2021) is a subset of 275 clinical cases taken from the 2019 edition of DEFT. This dataset is manually annotated in two tasks: (i) multi-label classification with 275 samples and (ii) NER. The multi-label classification task focuses on identifying the patient's clinical profile based on the diseases, signs, or symptoms mentioned in the clinical cases with 4,712 samples. The dataset is annotated with 23 axes derived from Chapter C of the Medical Subject Headings (MeSH). The second task involves fine-grained information extraction for 13 entities.

E3C (Magnini et al., 2020) is a multilingual collection of clinical cases annotated for Named Entity Recognition (NER). It encompasses two types of annotations: (i) clinical entities and (ii) temporal information and factuality. While this dataset spans five languages, our evaluation focuses on the French portion. Since the dataset does not come with predefined subsets for its 1,402 samples, we conducted random splits of 70% for training, 10% for validation, and 20% for testing, as outlined in Table 1.

Subset	Train	Validation	Test
Clinical	87.38% of L2	12.62% of L2	100% of L1
Temporal	70% of L1	10% of L1	20% of L1

Table 1: Description of the sources for E3C.

The QUAERO French Medical Corpus (Névéol et al., 2014), simply referred to as QUAERO in this paper, contains annotated entities and concepts for NER tasks. The dataset covers two text genres (drug leaflets and biomedical titles). 10 entity categories corresponding to the UMLS Semantic Groups (Lindberg et al., 1993) were annotated, for a total of 26,409 entity, which were mapped to 5,797 unique UMLS concepts. Due to the presence of nested entities, we opted to simplify the evaluation process by retaining only annotations at the higher granularity level, following a similar approach to the one described in Touchent et al. (2023), which translates into an average loss of 6.06% of the annotations on EMEA and 8.90% on MEDLINE. Additionally, considering that some documents from EMEA exceed the maximum input sequence length that most current language models can handle, we decided to split these documents into sentences.

MorFITT (Labrak et al., 2023b) is a multi-label dataset that has been annotated with medical specialties. It comprises 3,624 biomedical abstracts sourced from PMC Open Access. These abstracts have been annotated across 12 distinct medical specialties, resulting in a total of 5,116 annotations.

Mantra-GSC (Kors et al., 2015) is a multilingual dataset annotated in biomedical NER for five languages, however we focused only on the French subset. It covers three sources (EMEA, Medline and Patents) and use two distinct annotation schemes. These sources encompass diverse types of documents, including biomedical abstracts/titles, drug labels, and patents. To maintain evaluation uniformity, we randomly divided the dataset into three subsets: 70% for training, 10% for validation, and 20% for testing.

CLISTER (Hiebel et al., 2022) is a collection of French clinical case sentence pairs used for Semantic Textual Similarity (STS) evaluation. It consists of 1,000 sentence pairs, manually annotated by multiple annotators who assigned similarity scores ranging from 0 to 5 for each pair. These individual scores were then averaged to derive a floating-point number that represents the overall similarity of the two sentences.

CAS (Grabar et al., 2018) dataset comprises 3,790 clinical cases that underwent POS tagging with 31 different classes, using automatic tagging through the Tagex tool[3], achieving a 98% precision rate in comparison to manual annotations. This dataset involves tasks like classifying clinical cases for negation and uncertainty, as well as named-entity recognition for identifying markers of negations and speculation within medical histories and patient care. To create subsets, a random split was applied, allocating 70% for training, 10% for validation, and 20% for testing since predefined subsets were not provided.

ESSAI (Dalloux et al., 2021) consists of 7,247 clinical trial protocols that have been annotated with 41 POS tags using the TreeTagger tool (Schmid, 1994). It does also contain a classification and two named-entity recognition tasks similar to those from CAS dataset. As the dataset was not initially separated into three distinct subsets, we opted to apply the same processing methodology as we did for CAS dataset.

PxCorpus (Kocabiyikoglu et al., 2022) is a dataset designed for spoken language understanding in the medical domain, specifically focusing on transcripts related to drug prescriptions. It comprises 4 hours of transcribed dialogues, amounting to 1,981 recordings. These dialogues have been meticulously transcribed and semantically annotated. The primary task involves categorizing the textual utterances into one of four intent classes (prescribe, replace, negate, none). The second task pertains to NER, where each word in a sequence is classified into one of 38 classes, including categories such as drug, dose, or mode.

[3] https://allgo.inria.fr/app/tagex

4.2. Tokenizers Data Sources

To ensure a fair and comprehensive comparison of training data sources used by the statistical tokenizers, we carefully curated a 1GB subset of raw, lowercase text data from a variety of sources, including NACHOS (Labrak et al., 2023), PubMed Central, CC100 (Wenzek et al., 2020), and the French Wikipedia. We then constructed tokenizers using both tokenization algorithms, resulting in a total of 16 tokenizers: 8 with the integration of morphemes and 8 without. These specific data sources were chosen for their diversity: NACHOS focuses on French biomedical content, PubMed Central on English biomedical content, Wikipedia on general French language, and CC100 on general multilingual content. Each tokenizer was configured with a vocabulary size of 32k tokens, consistent with the original hyperparameters used in other French biomedical models such as CamemBERT-BIO (Touchent et al., 2023) and DrBERT (Labrak et al., 2023).

4.3. Language Model Pre-Training

To assess the impact of introducing morphemes into tokenizers on the pre-training process of biomedical language models, we conducted pre-training from scratch using the 16 tokenizer combinations (see Section 4.2). Our choice of architecture was RoBERTa (Liu et al., 2020), which is based on the masked language modeling objective and configured with standard token masking percentages as introduced by the authors.

For the PLMs training data, we utilized the NACHOS corpus created by (Labrak et al., 2023). This corpus, already pre-processed and converted to lowercase, is consistent with the data sources used for training the tokenizers. It comprises 1.1 billion words, equivalent to 7.4GB of raw text data, sourced from a wide range of online resources focusing on the French biomedical and clinical domains.

The pre-training process was conducted uniformly across all models, employing the same hyperparameters and executed over a 20-hour period. We harnessed the computational power of 32 V100 32GB GPUs available on the Jean-Zay supercomputer for this purpose. By maintaining consistent procedures and employing a fixed seed to mitigate randomness during training, we ensured the reliability and reproducibility of our experiments.

			BPE								SentencePiece							
			NACHOS		*PubMed*		*CC100*		*Wiki*		*NACHOS*		*PubMed*		*CC100*		*Wiki*	
Dataset	**Task**	**Metric**	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*
CAS	***CLS***	F1	94.2*	94.9	94.7	94.2*	95.2	**95.3**	94.8	94.8	94.8	94.7*	93.4**	93.6**	94.4	94.1	95.3*	95.1**
	NER Neg	SeqEval	**87.0**	83.3*	82.4**	81.3**	84.9	84.2	84.7	84.5*	86.1	86.4	83.6*	83.9	85.4	84.2**	85.6	83.2
	NER Spec	SeqEval	30.3*	30.6	35.0	28.2*	34.6	32.0	34.4	34.0	**36.1**	29.8	28.4*	22.2**	31.9	28.7*	32.1	27.0*
	POS	SeqEval	97.0**	96.9**	97.1	96.9**	97.1*	97.0**	**97.2**	96.9**	97.1*	97.0*	96.9**	96.9**	97.1	97.1	97.1	97.1
PxCorpus	***CLS***	F1	94.8	94.2	93.6	93.9	94.2	94.6	93.4	93.7	**94.9**	94.1	94.8	94.1	94.8	93.7	93.7	94.5
	NER	SeqEval	95.9	95.9	95.9	95.9	96.1	96.0	96.2	95.9	96.1	96.1	96.0	96.1	95.9	96.1	**96.2**	96.1
DEFT2020	***STS***	MSE	0.71	0.71	0.64*	**0.75**	0.70	0.67	0.71	0.69	0.72	0.71	0.63**	0.63	0.70	0.67*	0.70	0.67*
	CLS	F1	**91.0**	85.9	57.6**	73.7	79.5	76.3	77.1	66.0	83.0	85.3	80.9	66.7**	61.1*	66.3*	75.0*	77.4*
MORFITT	***CLS***	F1	68.6**	68.0**	66.5**	65.9**	68.4**	67.0**	68.7	67.3**	**69.6**	68.8*	66.8**	66.2**	68.2	67.5**	69.1**	67.7**
E3C	***NER Clinical***	SeqEval	**54.2**	53.1	52.4	48.6**	52.7	51.3**	51.1*	52.0*	54.2	52.4	52.1	51.1**	53.8	52.5*	53.2	51.7
	NER Temporal	SeqEval	82.0	81.2	80.9**	80.0**	81.8	81.2	**82.3**	80.6**	82.1	81.6	80.3**	79.8**	80.6**	81.1**	81.6*	81.73*
CLISTER	***STS***	MSE	0.63*	0.63	0.63	0.60**	0.65	0.63	0.62**	**0.66**	0.61*	0.64	0.61**	0.62**	0.62	0.60*	0.64**	0.63**
DEFT2021	***NER***	SeqEval	60.3	59.0**	58.1**	56.2**	59.4**	59.2**	60.1**	59.1**	**61.3**	60.1*	57.0**	56.6**	59.2**	59.9**	59.3**	58.9**
	CLS	F1	32.9	34.5*	33.4	32.3	34.5*	33.9	34.2	32.9	34.3	33.1	34.3	33.1	31.0	31.9*	34.2	**34.9**
ESSAI	***NER Spec***	SeqEval	60.5	60.9	56.4*	59.2	57.9	61.5	63.6	57.4	63.9	62.8	57.6	55.7*	**64.6**	62.0	61.4	63.1
	POS	SeqEval	98.4*	98.3	98.3	98.2**	98.4	98.4	98.3	98.3	98.4	98.4	98.3	98.2*	**98.4**	98.3	98.3*	98.3
	NER Neg	SeqEval	83.0	83.4	79.3	76.4	82.2	83.2	81.8	84.2*	81.3	84.0*	80.2	81.1	83.2	**84.2**	82.1	79.6*
	CLS	F1	97.3	97.1*	97.4	96.6**	97.4	96.7**	97.4	97.0**	97.3	97.3	**97.5**	97.2*	97.0	97.0	97.5*	97.0*
QUAERO	***NER Medline***	SeqEval	57.7	56.2**	55.4**	53.6**	57.9	55.0**	57.3	56.4**	**58.2**	55.5**	54.8**	52.9**	57.5*	55.8**	56.9	54.9**
	NER EMEA	SeqEval	65.6	65.1	63.9	63.1**	62.1**	62.7*	63.1**	62.6*	65.5	**65.9**	62.6**	63.8*	62.8**	63.1*	62.7*	62.0**
MantraGSC	***NER EMEA***	SeqEval	60.9	63.9	58.2*	60.6*	**69.3**	63.0	61.9*	62.3**	66.9	62.5*	56.8**	60.3	60.8*	59.5	64.0*	63.9**
	NER Medline	SeqEval	41.4*	42.9	39.3	36.2**	44.3	41.2	43.8	40.8*	41.9	39.5*	36.4**	37.8	46.4*	39.9	**47.1**	36.1*
	NER Patents	SeqEval	52.1*	53.3*	**57.0**	50.2*	57.0	53.9	53.6	52.3*	52.0	49.6*	50.7**	49.4	52.8*	48.0	50.6*	47.8*
Average performances per tasks																		
CLS		F1	**79.80**	79.10	73.87	76.10	78.20	77.30	77.60	75.28	78.98	78.88	77.95	75.15	74.42	75.08	77.47	77.77
NER		SeqEval	63.92	63.75	62.63	60.73	64.63	63.42	64.15	63.24	**65.05**	63.55	61.27	60.82	64.22	62.69	64.06	62.00
POS		SeqEval	97.70	97.60	97.70	97.55	**97.75**	97.70	97.75	97.60	97.75	97.70	97.60	97.55	97.75	97.70	97.70	97.70
STS		MSE	0,67	0,67	0,64	0,68	**0,68**	0,65	0,67	0,68	0,67	0,68	0,62	0,63	0,66	0,64	0,67	0,65

Table 2: Performance of the tokenization algorithms and different data sources used to train tokenizers (top). Average performance per type of tasks is also reported (bottom). *w/o* and *w/* denote models without and with morphemes. Best models are in bold, and the second-best are underlined. Statistical significance is determined using Student's t-test, where * indicates $p < 0.05$, and ** $p < 0.01$.

4.4. Evaluation

All models undergo fine-tuning following a standardized protocol with identical hyperparameters for each downstream task, enabling a focused evaluation of tokenizers. We ensure robustness and reliability by averaging the results across four independent runs and performing statistical significance assessments using Student's t-test.

For consistent comparisons, especially in sequence-to-sequence tasks like POS tagging and NER, we employ the SeqEval (Nakayama, 2018) metric in conjunction with the IOB2 format. To align with established practices (Touchent et al., 2023), our models are trained to predict only the label for the initial token of each word.

5. Results and Discussions

In this section, we present the results of our tokenization strategies on various biomedical NLP tasks, with a focus on key aspects. We investigate the impact of tokenization granularity (Section 5.1), the introduction of morphological information during tokenizer construction (Section 5.2), and the influence of data sources on tokenizers, including token sparsity, morpheme coverage, and the overall performance of different tokenization algorithms (Section 5.3).

Table 2 summarizes the performance of the BPE and SentencePiece strategies, both with (*w/*) and without our morpheme-enriched approach (*w/o*), across various French biomedical downstream tasks. Average performance per task type is also provided for clarity. It's worth noting that, before delving into detailed analysis, there is no consistent tokenization strategy that consistently yields the best results in all tasks, whether it employs a purely statistical algorithm or a statistical approach coupled with morpheme enrichment.

5.1. Impact of tokenization granularity

To assess the impact of tokenization granularity, Table 3 presents the average number of sub-word units per word for each tokenization strategy and data source used in the studied tasks. While deriving overarching conclusions from these results can be challenging, we calculated Pearson correlation (ρ) between models performances on the downstream tasks from Table 2 and the corresponding average number of sub-word units per word. These correlation scores range from -1 to $+1$, where -1 indicates a complete negative linear correlation, 0 represents no correlation, and $+1$ signifies a strong positive correlation. In the context of tokenization,

		BPE								SentencePiece								
		NACHOS		*PubMed*		*CC100*		*Wikipedia*		*NACHOS*		*PubMed*		*CC100*		*Wikipedia*		
Corpus	**Task**	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	ρ
CAS	**CLS**	1.32	1.38	2.20	2.13	1.49	1.49	1.51	1.50	1.32	1.45	2.18	2.15	1.49	1.56	1.51	1.57	*-0.62*
	NER Neg	1.32	1.38	2.20	2.13	1.49	1.49	1.51	1.50	1.32	1.45	2.18	2.15	1.49	1.56	1.51	1.57	*-0.70*
	NER Spec	1.32	1.38	2.20	2.13	1.49	1.49	1.51	1.50	1.32	1.45	2.18	2.15	1.49	1.56	1.51	1.57	*-0.42*
	POS	1.32	1.38	2.20	2.13	1.49	1.49	1.51	1.50	1.32	1.45	2.18	2.15	1.49	1.56	1.51	1.57	*-0.36*
PxCorpus	**CLS**	1.54	1.62	2.26	2.27	1.76	1.72	1.73	1.72	1.54	1.67	2.24	2.30	1.72	1.77	1.77	1.82	*-0.22*
	NER	1.54	1.62	2.26	2.27	1.76	1.72	1.73	1.72	1.54	1.67	2.24	2.30	1.72	1.77	1.77	1.82	*-0.22*
DEFT2020	**STS**	1.41	1.45	2.27	2.24	1.42	1.45	1.43	1.45	1.41	1.49	2.24	2.23	1.41	1.48	1.42	1.49	*-0.47*
	CLS	1.21	1.26	2.13	2.09	1.31	1.34	1.33	1.36	1.20	1.32	2.05	2.04	1.25	1.34	1.29	1.37	*-0.41*
MorFITT	**CLS**	1.38	1.44	2.45	2.40	1.48	1.50	1.49	1.51	1.37	1.50	2.35	2.33	1.46	1.55	1.48	1.57	*-0.82*
E3C	**NER Clinical**	1.30	1.35	2.23	2.17	1.48	1.48	1.50	1.49	1.29	1.43	2.22	2.18	1.48	1.55	1.49	1.56	*-0.59*
	NER Temporal	1.29	1.35	2.22	2.16	1.48	1.48	1.48	1.49	1.29	1.43	2.22	2.18	1.47	1.54	1.48	1.55	*-0.75*
CLISTER	**STS**	1.52	1.59	2.65	2.57	1.73	1.72	1.74	1.72	1.51	1.65	2.56	2.49	1.71	1.77	1.71	1.77	−0.33
DEFT2021	**NER**	1.31	1.37	2.26	2.19	1.48	1.49	1.50	1.50	1.31	1.44	2.19	2.15	1.48	1.55	1.49	1.56	*-0.88*
	CLS	1.50	1.57	2.63	2.56	1.69	1.70	1.71	1.71	1.46	1.61	2.50	2.46	1.64	1.72	1.66	1.74	*-0.11*
ESSAI	**NER Spec**	1.29	1.34	2.20	2.14	1.42	1.43	1.45	1.45	1.29	1.41	2.21	2.16	1.41	1.49	1.46	1.52	*-0.68*
	POS	1.28	1.33	2.19	2.13	1.41	1.42	1.44	1.44	1.28	1.41	2.19	2.15	1.40	1.48	1.44	1.51	*-0.61*
	NER Neg	1.28	1.33	2.19	2.13	1.41	1.42	1.44	1.44	1.28	1.41	2.19	2.15	1.40	1.48	1.44	1.51	*-0.69*
	CLS	1.28	1.34	2.20	2.14	1.42	1.43	1.45	1.46	1.28	1.41	2.20	2.16	1.41	1.49	1.45	1.52	*-0.02*
QUAERO	**NER Medline**	1.53	1.63	2.35	2.26	1.78	1.78	1.77	1.78	1.52	1.76	2.36	2.35	1.77	1.89	1.76	1.89	*-0.77*
	NER EMEA	1.30	1.34	2.14	2.12	1.44	1.46	1.49	1.51	1.30	1.39	2.06	2.04	1.45	1.51	1.50	1.56	*-0.28*
MANTRAGSC	**NER EMEA**	1.33	1.40	2.47	2.41	1.49	1.51	1.50	1.52	1.32	1.43	2.33	2.30	1.46	1.53	1.49	1.55	*-0.63*
	NER Medline	1.89	2.01	2.84	2.70	2.06	2.13	2.14	2.14	1.89	2.09	2.84	2.78	2.06	2.22	2.10	2.22	*-0.64*
	NER Patents	1.54	1.59	2.34	2.30	1.61	1.63	1.59	1.62	1.43	1.52	2.20	2.20	1.50	1.58	1.51	1.60	*0.06*
Average per model		1.39	1.45	2.30	2.25	1.54	1.55	1.56	1.56	1.38	1.51	2.26	2.24	1.52	1.60	1.55	1.62	*-0.48*
Relative Difference (%)		0.0	4.5	65.9	61.8	11.2	11.8	12.3	12.6	−0.7	8.9	62.8	61.1	9.9	15.5	11.7	16.9	

Table 3: Average number of sub-word units per word for each tokenization strategy and data source training. Their Pearson correlation (ρ) with each task performance is reported (last column). Cells colored in red correspond to lower performing models, while those in green represent higher ones. The last row represents the relative difference in terms of average subwords per word compared to the NACHOS BPE without morpheme baseline. *w/o* and *w/* denote models without and with morphemes.

a negative correlation implies that fewer subword units are associated with higher scores, while a positive correlation suggests that more subword units are linked to higher scores.

In overall, we observe in Table 3 an average ρ correlation of -0.48 between tasks and models, indicating that, in general, higher performance scores tend to be associated with fewer subword units. To our knowledge, this is the first time such a correlation has been experimentally demonstrated. However, it's important to note that this correlation varies across the targeted tasks. Tasks like CLS show correlation close to zero, suggesting that they are less affected by the granularity of tokenization. In contrast, STS and sequence-to-sequence tasks, particularly NER, appear to be more influenced by tokenization granularity, likely due to their heavy reliance on immediate context for making predictions.

While the RoBERTa model's embeddings capture semantic meaning and the encoder module captures contextual information (Rogers et al., 2020), we aimed to determine whether the observed correlations are attributed to a specific part of this architecture. To investigate this, we isolated and froze the embeddings and/or encoder of our BERT-based model, based on the NACHOS SentencePiece, during fine-tuning for various tasks. The experimental approach, as detailed in Table 4, involved several stages. Initially, we established a baseline for each task with no frozen components. Subsequently, we conducted experiments by freezing only the embedding layer, only the encoder, and both the embeddings and encoders. Our findings indicate a stronger dependence on RoBERTa's encoder for tasks such as POS tagging and STS, in contrast to other tasks, which corroborate the context dependency as an explanation to the correlation scores between segmentation granularity and models performances for these tasks but not for NER.

	CAS POS	PxCorpus NER	PxCorpus CLS	CLISTER STS
Full Fine-tuning	97.10	96.10	94.82	0.61
Embedding	97.03 ↓ 0.07	96.10 ↑ 0.00	94.73 ↓ 0.09	0.62 ↑ 1.63
Encoder	65.97 ↓ 32.05	83.95 ↓ 12.64	84.78 ↓ 10.58	0.45 ↓ 26.22
Embedding + Encoder	60.04 ↓ 38.16	79.62 ↓ 17.14	84.78 ↓ 10.58	0.44 ↓ 27.86

Table 4: Performance and relative loss (in %) of the PLMs based on SentencePiece NACHOS without morpheme with parts of the models being frozen.

As shown in Table 3, higher performance scores are associated with fewer subword units. To gain a linguistic perspective on how tokenization strategies behave, we analyzed the segmentation of 150 biomedical terms equally distributed across cardiology, dermatology, obstetric-gynecology, and

ophthalmology, as presented in Table 5. Most models, except for those using SentencePiece NACHOS, struggle to precisely align with the official morphological segmentation established by the Académie Française (French Academy). However, upon closer examination, it is evident that these models often come very close to the desired segmentation. While the segmentations may exhibit slight variations, such as the relocation of a letter from one token to another, they maintain the same number of tokens as the official morphological segmentation. This observation is further supported when we analyze actual tokenizer outputs (see Table 6) and assess the segmentation statistics in Table 5. For example, BPE NACHOS tokenizes the term "ophtalmoscope" into the units "ophtalm oscope," whereas the morphological segmentation should be "ophtalmo scope," a segmentation achieved by its morpheme-enriched counterpart.

					Type of errors	
			EM*	Exact # Tok.	Under Seg.	Over Seg.
BPE	***NACHOS***	w/o	21.3	41.3	9.3	49.3
		w/	34.6	50.0	6.0	44.0
	PubMed	*w/o*	2.6	12.0	2.6	85.3
		w/	17.3	28.6	2.6	68.6
	CC100	*w/o*	8.0	28.0	2.6	69.3
		w/	23.3	38.6	2.6	58.6
	Wikipedia	*w/o*	8.6	24.6	3.3	72.0
		w/	22.0	36.6	4.6	58.6
SP	***NACHOS***	w/o	56.6	74.6	7.3	18.0
		w/	61.3	70.6	2.6	26.6
	PubMed	*w/o*	14.6	26.6	2.6	70.6
		w/	32.0	42.0	2.6	55.3
	CC100	*w/o*	24.0	42.0	4.0	54.0
		w/	36.6	49.3	2.6	48.0
	Wikipedia	*w/o*	18.0	42.0	3.3	54.6
		w/	34.0	54.0	4.6	41.3

Table 5: The average Exact Match (EM*) and portion of terms aligned with the official segmentation length (Exact # Tok.), both in %, are based on the gold segmentation from 150 biomedical terms. Both last columns are referring to the portion of terms suffering from under and over segmentation. *w/o* and *w/* denote without and with morphemes respectively. SP stands for SentencePiece.

In Table 5, we observed various types of errors in segmentation, with the most common issue being over-segmentation of units that are not present in our biomedical lexical morphemes list. This over-segmentation results in smaller, more numerous, and sparser tokens, which can impact the efficiency of pre-training. The reduced frequency of tokens and the faster filling of RoBERTa's 512-token context window with less meaningful tokens can be problematic.

Finally, Table 5 reveals an interesting distinction between BPE and SentencePiece using NACHOS training data. SentencePiece outperforms BPE in achieving segmentations that closely resemble correct ones, both in terms of the number of tokens and their semantic accuracy. SentencePiece excels at matching correct segmentations, particularly for medical terminology, in 56.6% of cases without morphemes and 61.3% when morphemes are used, while BPE NACHOS achieves only 34.6% accuracy.

Base	*cancérigène*	*ophtalmoscope*	*angiographie*
Correct	**cancér i gène**	**ophtalmo scope**	**angio graphie**
BPE Wiki	c anc éri gène	oph tal mos cope	ang i ographie
BPE PubMed	can c é rig è ne	o ph tal m oscope	angi ograph ie
BPE NACHOS	cancé rig ène	ophtalm oscope	angiographie
SentencePiece NACHOS	cancérigène	ophtalm oscope	angiographie
BPE NACHOS `+Morpheme`	**cancér i gène**	**ophtalmo scope**	**angio graphie**
SentencePiece NACHOS `+Morpheme`	**cancér i gène**	**ophtalmo scope**	**angio graphie**

Table 6: Instances of tokenization juxtaposed with their correct segmentation.

5.2. Impact of morphemes

One of our primary objectives was to approximate the correct morphological segmentation of words in the French biomedical language. Our analysis reveals that tokenizers, such as BPE and SentencePiece trained on NACHOS, enriched with morphemes, can often achieve this goal. Notably, SentencePiece NACHOS enriched with morphemes achieved the best performance, with a 61.3% exact match. Our morpheme-enriched approach offers the advantage of obtaining a tokenization that closely resembles what could be achieved through a complex rule-based method. This approach is easily adaptable to other languages with a list of lexical morphemes and similar principles.

As shown in Table 2, the introduction of morphemes (*w/*) may lead to performance enhancements in approximately 25% of the studied downstream tasks. However, it is noteworthy that the best results are primarily achieved by classical statistical tokenizers, BPE and SentencePiece, when not using morphemes, and when trained on our biomedical-specific data, NACHOS. This observation is intriguing because NACHOS-based tokenizers inherently contain a higher proportion of morphemes, as shown in Table 7, which presents the portion of correct morphemes already present in the tokenizers without introducing additional morphological information based on their length ranges. This suggests that introducing morphemes and other forms of morphological knowledge, such as grammatical endings, may have a more substantial impact in contexts that do not align directly with the target domains and languages. However, we can note that the results of this method are inconsistent and do not ensure an overall performance boost across all models or tasks.

Furthermore, it is worth noting that morphemes are often already present in the tokenizers in their complete form, as illustrated in Table 7, or with

		Coverage of the morphemes (%)			
Tokenizer	**Source**	**1 - 3**	**4 - 6**	**7 - 10**	**Global**
BPE	***NACHOS***	83.33	45.38	31.00	47.23
	PubMed	65.15	39.32	15.00	38.06
	CC100	78.78	34.46	7.00	34.77
	Wikipedia	87.87	34.95	10.00	36.67
SP	***NACHOS***	83.33	41.01	28.00	43.59
	PubMed	60.60	37.13	14.00	35.81
	CC100	83.33	34.70	8.00	35.64
	Wikipedia	93.93	37.37	12.0	39.44

Table 7: Percentage of the morphemes already present in the tokenizers vocabularies per range of morphemes lengths. SP stands for SentencePiece.

minor modifications based on token probabilities, as shown in Table 6. Notably, tokenizers based on NACHOS contain a significantly higher percentage of morphemes, with 47.23% for BPE and 43.59% for SentencePiece. Conversely, the source with the fewest morphemes is CC100, with percentages of 34.77% for BPE and 35.64% for SentencePiece. This observation aligns with the fact that CC100 has fewer connections to both the target language and domain.

In general, we observe that despite the significant improvement in segmentation quality (as shown in Table 5), tokenizers enriched with morphemes do not exhibit a strong correlation with the results achieved in downstream tasks, as evident in Table 2. The ability to deliver satisfactory results despite encountering suboptimal segmentations, as seen in the case of PubMed, which frequently oversegments words, underscores the robustness of RoBERTa's architecture in handling noise and its capacity to compensate for such challenges.

5.3. Impact of data sources

As indicated in Table 2, the average performance across tasks demonstrates a significant impact of the training data source on the results obtained by the models. It becomes apparent that using data that is more suitable for the target language, even if it originates from various domains such as Wikipedia and CC100, is more effective than utilizing data from the target domains but from a different language. This is particularly evident in the CLS, NER, and STS tasks, where BPE PubMed achieves an average of 70.16% for classification, 0.63 MSE for STS, and 62.62% for NER, whereas CC100 outperforms with 74.14%, 0.67 MSE, and 64.62%, respectively.

The decrease in performance from PubMed can be attributed to over-segmentation, as seen in Table 3. This over-segmentation is primarily due to the significant differences between the data used to build the tokenizer and the language of the model's pre-training. These differences stem from distinct lexicons, writing styles, and morphological structures in French compared to English, particularly for specialized words like "Péricardite" (French) and "Pericarditis" (English), or "Orthophoniste" (French) and "Speech Therapist" (English). Furthermore, variations in alphabets, such as special French characters like "é" or "è," can lead to token sparsity when encountered in positions not seen during tokenizer construction on PubMed. This results in a lack of both language and domain-specific information for French, as only limited tokens can be used to form sentences.

Some data sources are surprisingly less affected by the introduction of morphemes. For instance, the CC100 source is not positively impacted by morphemes, despite having a lower proportion of morphemes in its original version, as shown in Table 7. This behavior may be explained by the increased granularity introduced by morphemes, which reduces the probabilities of other tokens appearing. This can lead to a poorer representation of words.

6. Conclusion

In this study, we conducted a comprehensive investigation into the influence of various word tokenization strategies within a BERT-based masked language model across diverse French biomedical NLP tasks. Notably, we observed that existing methods for tokenizing biomedical text often fall short of aligning with morphological rules and how humans learn these specialized terms. This suboptimal segmentation can impact the agglutinating nature of biomedical terminology. To assess the effects of this segmentation on downstream applications, we developed a set of novel biomedical tokenizers that adhere more closely to morphological rules. These tokenizers combine various automatic tokenization approaches and vocabularies to enrich segmentation with morphemes. We employed these enhanced tokenizers in the pre-training of multiple RoBERTa-based models, which we then evaluated across a wide array of 23 French biomedical tasks, including POS, NER, STS, and CLS.

Our findings show that integrating morphemes into automatic tokenization approaches can achieve parity or improve performance in certain tasks, such as NER and POS tagging. However, this enhancement is not consistent across all tasks. While there is a correlation between segmentation granularity and downstream task performance, we also observe that pre-training processes exhibit robustness to suboptimal tokenization, yielding surprisingly good results even with very short and sparse subword units. To conclude, our study reveals that achieving optimal tokenization involves a combination of factors, including minimizing word segmentation and having access to domain-specific data in the target language.

7. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715). This work was financially supported by ANR MALADES (ANR-23-IAS1-0005) and Zenidoc.

8. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Issam Bazzi and James R Glass. 2002. A multi-class approach for modelling out-of-vocabulary words. In *Interspeech*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pretrained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31, Virtual. Association for Machine Translation in the Americas.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized French language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France. ATALA et AFCP.

H. Cottez. 1980. *Dictionnaire des structures du vocabulaire savant: éléments et modèles de formation*. Collection Les usuels du Robert. Le Robert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Retrain or train from scratch? comparing pretraining strategies of BERT in the medical domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2626–2633, Marseille, France. European Language Resources Association.

Allison Fan and Weiwei Sun. 2023. Constructivist tokenization for English. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 36–40, Washington, D.C. Association for Computational Linguistics.

Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita, and Yasuhiro Sogawa. 2023. How do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research*

Workshop), pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362, Toronto, Canada. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickael Rouvier, Pacome Constant dit Beaufils, Natalia Grabar, Beatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-Antoine Gourraud, and Richard Dufour. 2024. Drbenchmark: A large language understanding evaluation benchmark for french biomedical domain.

Martha Larson. 2001. Sub-word-based language models for speech recognition: implications for spoken document retrieval. *Whorkshop on Language Modeling and Information Retrieval*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp.

Vít Novotný, Eniafe Festus Ayetiran, Dalibor Bačovský, Dávid Lupták, Michal Štefánik, and Petr Sojka. 2021. One size does not fit all: Finding the optimal subword sizes for FastText models across languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1068–1074, Held Online. INCOMA Ltd.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

David Samuel and Lilja Øvrelid. 2023. Tokenization with factorized subword encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14143–14161, Toronto, Canada. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Igor Szoke, Lukás Burget, Jan Cernocky, and Michal Fapso. 2008. Sub-word modeling of out of vocabulary words in spoken term detection. In *2008 IEEE Spoken Language Technology Workshop*, pages 273–276. IEEE.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Rian Touchent, Laurent Romary, and Eric de la Clergerie. 2023. Camembert-bio: a tasty french language model better for your health.

Christian Touratier. 2012. Chapitre V. Les classes de morphèmes. In *Morphologie et morphématique : Analyse en morphèmes*, Langues et langage, pages 78–114. Presses universitaires de Provence, Aix-en-Provence.

Yining Wang, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2017. Word, subword or character? an empirical study of granularity in chinese-english nmt. In *Machine Translation*, pages 30–42, Singapore. Springer Singapore.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

9. Language Resource References

Rémi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France. ATALA et AFCP.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 1–7, Brussels, Belgium.

Cyril Grouin, Natalia Grabar, and Gabriel Illouz. 2021. Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021 (clinical cases classification and automatic evaluation of student answers : Presentation of the DEFT 2021 challenge). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 1–13, Lille, France. ATALA.

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2022. CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4306–4315, Marseille, France. European Language Resources Association.

Alican Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *13th Language Resources and Evaluation Conference (LREC 2022)*.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023a. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Yanis Labrak, Mickaël Rouvier, and Richard Dufour. 2023b. MORFITT : A multi-label corpus of French scientific articles in the biomedical domain. In *30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) Atelier sur l'Analyse et la Recherche de Textes Scientifiques*, Paris, France. Florian Boudin.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The Unified Medical Language System. *Methods Inf Med*, 32(4):281–291.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Rian Touchent, Laurent Romary, and Eric de la Clergerie. 2023. Camembert-bio: a tasty french language model better for your health.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.