A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages

Lisa Raithel[1,2,3,4*], **Hui-Syuan Yeh**[4,*], **Shuntaro Yada**[5], **Cyril Grouin**[4], **Thomas Lavergne**[4], **Aurélie Névéol**[4], **Patrick Paroubek**[4], **Philippe Thomas**[3], **Tomohiro Nishiyama**[5], **Sebastian Möller**[1,2,3], **Eiji Aramaki**[5], **Yuji Matsumoto**[6], **Roland Roller**[3], **Pierre Zweigenbaum**[4]

[1]BIFOLD, Ernst-Reuter Platz 7, 10587 Berlin, Germany;
[2]Quality & Usability Lab, TU Berlin, Ernst-Reuter Platz 7, 10587 Berlin, Germany;
[3]German Research Center for Artificial Intelligence (DFKI), Alt-Moabit 91c, 10559 Berlin, Germany;
[4]Université Paris-Saclay, CNRS, LISN, Rue du Belvédère, 91405 - Orsay, France;
[5]Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan;
[6]RIKEN, Nihonbashi 1-chome Mitsui Building, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

Abstract

User-generated data sources have gained significance in uncovering Adverse Drug Reactions (ADRs), with an increasing number of discussions occurring in the digital world. However, the existing clinical corpora predominantly revolve around scientific articles in English. This work presents a multilingual corpus of texts concerning ADRs gathered from diverse sources, including patient fora, social media, and clinical reports in German, French, and Japanese. Our corpus contains annotations covering 12 entity types, four attribute types, and 13 relation types. It contributes to the development of real-world multilingual language models for healthcare. We provide statistics to highlight certain challenges associated with the corpus and conduct preliminary experiments resulting in strong baselines for extracting entities and relations between these entities, both within and across languages.

Keywords: biomedical NLP, information extraction, adverse drug reactions, multilingual

1. Introduction

An adverse drug reaction (ADR) is a "harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product" (Edwards and Aronson, 2000). ADRs constitute a significant problem in pharmacovigilance. No medication is without side effects, and even though there are clinical trials for each drug, the pool of patients included in the trials can never represent an entire population with respect to, e.g., age, gender, health, or ethnicity (Hazell and Shakir, 2006). Even post-release surveillance campaigns might fail to reach the patients who have issues with the released medication (Hazell and Shakir, 2006). Therefore, medication use and effects must be monitored constantly.

Consequently, biomedical and clinical texts are a much-used resource for supporting pharmacovigilance since they contain information about patients, their medication intake, and, potentially, their medical history. For example, researchers extract information from electronic health records (EHRs), scientific publications, public health or treatment guidelines, search logs (White et al., 2016), and any other text dealing with medical issues. However, all of these are written from the physician's perspective of treating the patient.

In contrast, social media, such as X (formerly Twitter) or Facebook, are created from the patient's perspective. Taken collectively, social media content can provide population-level signals for ADRs and other health-related topics. Internet and social media engage many users and offer the means to access data at scale for specific topics of interest. Previous studies have shown that despite the large online user communities, they are not necessarily a representative sample of the population at large (Hargittai and Walejko, 2008; Wagner et al., 2015). Nonetheless, people can use social media to anonymously discuss health issues in their own words without the fear of not being taken seriously, which is, in fact, one of the reasons for the under-reporting of ADRs, alongside a general mistrust of clinical providers (Yang et al., 2012; Palleria et al., 2013). Another factor making social media useful for detecting ADRs is the variety of languages provided on the internet, making health-related information more accessible to laypeople. They, therefore, often turn to patient fora to research and collect information on topics they are concerned with, following "translations" from technical terminology to lay language provided by other members of the respective communities. Sometimes, there are even clinicians involved in these fora. This, again, highlights the necessity to extract relevant information not only from texts written by experts but also to listen to

*Shared first authorship; Corresponding author: raithel@tu-berlin.de

LREC-COLING 2024, pages 395–414

the patients' voices and process texts written by "normal" people.

Although the number of non-English and multilingual datasets is rising in the clinical and biomedical domain (Névéol et al., 2018), there is still a need for shareable corpora for particular tasks. Especially the detection of ADRs, which is important across all countries and, therefore, languages, still shows much room for improvement, even in English data (Magge et al., 2021), but more so in other languages (Klein et al., 2020; Raithel et al., 2022). Furthermore, shifting the perspective to the patient with the help of social media might support clinicians and other practitioners to understand their patients and the experienced ADRs better, react more appropriately, and meet the patients' needs more precisely (Arase et al., 2020). This also allows patients to participate actively in their treatment (Segura-Bedmar et al., 2014). Finally, the collected information from these crowd signals can be used for drug re-purposing and the development of new medications (Scaboro et al., 2022). Therefore, with the presented work, we aim to broaden the access to resources for pharmacovigilance across languages and switch the perspective on health issues to the one of the patients. We contribute to the development of real-world and multilingual models for patient-centric healthcare as follows:

- We provide a new multilingual corpus focused on ADRs in three languages: German, French, and Japanese. It is annotated with entities, attributes, and relations to describe experiences with ADRs from a patient's perspective.[1]
- We describe the characteristics of the presented data and highlight challenges associated with the extraction of ADRs.
- We provide annotation guidelines, which aim to be robust across a variety of languages.[2]
- We provide baseline models for named entity extraction, attribute classification, and relation extraction.[1]

2. Related Work

2.1. Methods

Since approximately 2010, with one of the first publications on the extraction of ADRs from social media by Leaman et al. (2010), the interest in and the number of social media datasets has been growing slowly. By now, researchers, health-related industries, and authorities recognize the value of patient-generated data with respect to improving medication products and public health monitoring (Sarker et al., 2015).

Detecting and extracting ADRs from social media is done similarly to other information extraction tasks in NLP. With the success of Transformer-based models (Vaswani et al., 2017) like `BERT` (Devlin et al., 2019) and `XLM-RoBERTa` (Conneau et al., 2020) in almost all areas of NLP, these also started to dominate in the task of ADR extraction (Tutubalina et al., 2021; Weissenbacher et al., 2022). Even so, there are still quite a few challenges in need of being addressed. The detection of ADRs in user-generated texts often suffers from small corpora (see Section 2.2), imbalanced label distributions, spelling mistakes, and colloquial language in general, and only a few language-specific medical Transformer models exist. Further, documents can contain ambiguous content and speculated statements or patients worrying about things that have not yet happened. These need to be distinguished from actual occurrences of ADRs.

In the context of the Social Media Mining for Health[3] (SMM4H) 2022 shared task, Portelli et al. (2022) address the limits of Transformers concerning document classification, entity extraction, and normalization. They show that ensembling methods and architectures can improve the performance of these models, but also by using generative models like `GPT-2` (Radford et al., 2019).

Miftahutdinov et al. (2020) compare different model and data setups and demonstrate that a Convolutional Neural Network in combination with `fastText` embeddings (Bojanowski et al., 2017) can outperform `mBERT` on Russian ADR texts in binary classification. When using both English and Russian tweets for fine-tuning an English-Russian `BERT` model (`EnRuDRBERT`) they achieved the best result (within their experiments), especially when compared to only fine-tuning on Russian data. However, the authors also note that adding Russian data to the English data only improved the results on the English test set by one percentage point. Gencoglu (2020) uses sentence embeddings (Reimers and Gurevych, 2019) to represent the tweets from SMM4H 2020, based on `RoBERTa` (Liu et al., 2019) for English and multilingual `DistilBERT` (Sanh et al., 2019) for Russian and French to perform document classification. They further weigh the contribution of positive samples to the loss function higher than the one of the negative samples to account for the label imbalance. With this, they achieve the best result within the shared task on the French dataset ($F_1 = 17\%$).

Chowdhury et al. (2018) simultaneously classify

[1] Data and code can be found here: https://github.com/Dotkat-dotcome/KEEPHA-ADR

[2] https://github.com/DFKI-NLP/keepha_annotation_guidelines/blob/main/KEEPHA_annotation_guidelines.pdf

[3] https://healthlanguageprocessing.org/smm4h-2022/

posts and extract ADR and indication mentions from social media data in a multi-task setting. They combine additive attention (Bahdanau et al., 2015) and a coverage mechanism (See et al., 2017) in a Recurrent Neural Network (RNN) and show that with this, mentions of ADRs are captured more accurately. Raval et al. (2021) model the tasks of ADR classification and extraction in a generative setting and use `T5` (Raffel et al., 2020) for a sequence-to-sequence approach. Adding temperature scaling (Devlin et al., 2019) and proportional mixing to account for different dataset sizes and languages improves performance in the binary classification of the French SMM4H 2020 dataset ($F_1 = 20\%$) compared with earlier results.

Similar work focuses on other types of text or medically relevant information. For example, Meoni et al. (2023) and Agrawal et al. (2022) study multilingual medical entity extraction using large language models (LLMs) with `InstructGPT` (Ouyang et al., 2022) and Feng et al. (2023) propose `DKADE`, a framework incorporating a knowledge base that allows extracting ADRs and associated medication mentions in Chinese medical texts.

2.2. Existing Datasets

While there are many social media datasets related to the extraction of ADRs for English, e.g., the CADEC (Karimi et al., 2015) and PsyTar (Zolnoori et al., 2019) corpora, non-English datasets are rare. We show those published in recent years in Table 6 in Appendix A and describe them in more detail below.

Spanish The SpanishADR corpus (Segura-Bedmar et al., 2014) was the first non-English social media dataset focused on ADRs. The data originates from the patient forum "ForumClinic". 400 forum posts were randomly picked for annotation, and two annotators annotated adverse events and drug mentions. Segura-Bedmar et al. (2014) report an inter-annotator agreement (IAA) based on F_1 score of 0.89 for the drug mentions and 0.59 for adverse events.

Russian Alimova et al. (2017) provided the first corpus in Russian. They crawled the drug review forum Otzovik and created a corpus based on 580 reviews. The reviews were annotated sentence-wise with one out of four labels: *Indication, Beneficial effect, Adverse drug reaction, Other*.

Tutubalina et al. (2021) created RuDReC, also originating from Otzovik. The data is divided into two parts: one containing annotations on the entity level and the other one without annotations, comprising about 1.4 million texts from various online sources focused on health-related user posts. The annotated part comprises 500 documents. The labels of the annotated corpus are sentence-based, marking whether or not health-related issues are mentioned using five different sentence labels. Those that contain health problems were further annotated on the entity level, distinguishing six different entity types. IAA was determined to be "approximately 70%", using a relaxed agreement for entities following earlier work (Metke-Jimenez et al., 2014; Karimi et al., 2015).

Sboev et al. (2022) again harness reviews from Otzovik. 2,800 drug reviews are annotated with 18 entity types and additional attributes for drug and disease mentions. Further, specific mentions are normalized to their respective concepts ICD-10 and MedDRA. The accuracy achieved for the ADR entity type is 61.1% using exact F_1 score.

Klein et al. (2020) present a Twitter dataset made from Russian tweets with binary annotation. The training set (the only one available) contains 7,612 tweets of which 666 describe an ADR. For the test set, Klein et al. (2020) list 1,903 tweets with 166 of those expressing an ADR. The data was prepared for the fifth SMM4H shared task in 2020.

French For the same shared task, Klein et al. (2020) further provide a French corpus based on data collected from Twitter. The publicly available training set contains 2,426 tweets with 39 ADR examples.

Japanese Arase et al. (2020) published a corpus based on the Japanese patient forum TOBYO. The authors crawled all entries related to lung cancer and containing one to five drugs from a precompiled dictionary. The final corpus provides 169 documents annotated with drug effect spans, related drug mentions, types of reactions, and the ICD-10 codes for those. IAA was calculated using Fleiss' κ, resulting in $\kappa = 0.52$ for span and type agreement.

German Finally, the corpus provided by Raithel et al. (2022) contains data from the German patient forum lifeline.de and is annotated with binary labels, expressing whether or not a document contains a mention of an ADR. Of the 4,169 documents, only 101 contain ADRs, showing a similar distribution of labels as other binary annotated corpora.

3. The Corpus

Our new corpus contains data in three languages, German, French, and Japanese, based on the project collaborators' major languages. The three languages belong to different families, and the data

originates from different sources, representing diverse ways of expressing health-related issues written by laypeople. It is annotated with entities, attributes, and relations. With this, we aim to capture relevant medical mentions from a patient's perspective. We further add relationships between these entities to model, e.g., interactions between drugs and symptoms (that is, ADRs), body parts, or the patient's assessment of their well-being.

3.1. Data Collection

The general requirements we set for the data were as follows: (i) The data should be health-related but not specific to any drug or disease, (ii) the data should be de-identifiable or already de-identified, (iii) the data should be distributable to other research teams.

German For the German data, we obtained permission from the administrators of the forum Lifeline[4], to download and share the data. On Lifeline, people discuss their experiences with specific diseases or medication and help each other in various life situations. We built a crawler and downloaded all posts available in the user forum in July 2021, containing posts between 2000 and 2021. All messages were filtered for Covid-19-related posts, to remove potential vaccine-related reactions and discussions and avoid biasing our dataset towards this topic. Ten thousand texts were randomly sampled and annotated with a binary label, expressing whether the text mentioned an ADR or not. Of these 10,000, 324 contained ADR mentions and were subsequently further annotated.

French We found it very difficult to receive access to French patient fora. For every potential resource, requirement (iii) would not have been met. Thus, we translated some of the already de-identified German texts and annotated them with binary labels to reduce the annotation and curation effort. We used the DeepL[5] machine translation service to translate German texts into French. Then, we provided the texts to native French speakers who checked the texts for comprehensibility. Minor issues were corrected by our annotators and texts that were not comprehensible due to an erroneous translation were discarded. Finally, we chose 100 translated documents containing ADRs for further annotation. Due to the relatively lower number of annotations in the French data compared to the other two languages, we designate French as a low-resource target language for our cross-lingual experiment in Section 4.2. The set of French texts is distinct from the German texts to prevent data leakage in cross-lingual experiments. See Appendix B for more details on the French and German data.

entity type	**attributes**
drug	increase, decrease, stopped, started, unique_dose
change_trigger	
disorder	negated
function	negated
anatomy	
test	
opinion	positive, negative, neutral
measure	
time	frequency, duration, date, point in time
route	
doctor	
other	
user	
url	
personal_info	

Table 1: The different entity types and attributes. The bottom three are only for de-identification purposes.

Japanese The Japanese texts were collected from both Twitter and Yahoo! JAPAN Chiebukuro (YJQA)[6], a Japanese Q&A forum including health care issues. For this, we had to relax requirement (i) for Twitter since searching for tweets without keywords is not possible due to the limitations of the Twitter API. We collected tweets that mention the drug "Lexapro" and its ADR-related keywords (i.e., nausea, sleepiness, and appetite) from June 2017 to May 2020. This drug is popular enough to be mentioned in social media and known as causing ADRs; we plan to extend the variety of drugs in future work. For YJQA, we selected the Q&As labeled "concerns about ADR" from an existing YJQA breast-cancer corpus (Kamba et al., 2021), which is a labeled corpus of 1,000 randomly selected questions on breast cancer posted to YJQA from January 2018 to June 2020. See Appendix C for more details on the Japanese data.

3.2. Annotation

The annotation guidelines were first developed using English examples from CADEC (Karimi et al., 2015) and other English corpora related to ADRs. Annotation was conducted using BRAT[7]. After several pilot rounds of annotating these, the guidelines were applied to texts in the target languages and further refined. Then, our annotators, all (near-

[4] https://fragen.lifeline.de/forum/
[5] https://www.deepl.com/translator
[6] https://chiebukuro.yahoo.co.jp/
[7] https://brat.nlplab.org/

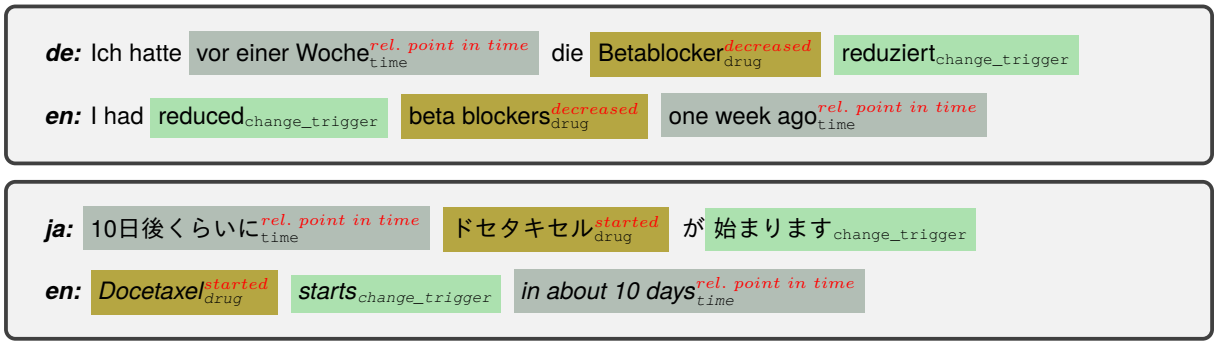


Figure 1: Example annotation of a German (top) and Japanese (bottom) text, with their respective English translation.

) native speakers of the respective languages[8], were trained the same way, i.e., by first annotating English examples. These annotations were subsequently discussed and any constructions that could not be modeled with our annotation scheme were further investigated to determine if these were language-agnostic or applied to one language only. Ultimately, we decided to annotate 12 entity types, four attribute types, and 13 relation types. They are shown in Table 1 and Table 2.

De-identification For de-identifying private user information, such as names and addresses, we first tried to find as many identifiers as possible using regular expressions specific to the respective sub-dataset. Very frequent occurrences were, for instance, user names, with which patients greet each other and/or refer to each other using nicknames (or nicknames of nicknames, for example, "Mohnblümchen" for the user name "Mohnblume", a diminutive of "poppy"). We collected the regular expression matches and replaced them with a mask (`<user>`) to keep the structure of the text intact. However, since users are very creative in inventing names, greetings, and goodbyes, not all of them were captured. Therefore, one of the tasks for the annotators was to add an entity label `user` to all still-existing names during the entity annotation process. Those were then replaced after the annotation was completed. We did the same for URLs, (e-mail) addresses, and other potentially de-identifying information.

Entities We annotate entities in the form of noun- or verbal phrases together with their modifying parts, e.g., adjectives and adverbs. Complex modifiers often occurring in Japanese are excluded to support the language-independence of the guidelines. We always prefer the smallest core noun phrase or, otherwise, the whole verb phrase. Metaphors, descriptive language, and spelling mistakes are annotated as well. Discontinuous entities are allowed if necessary.

In more detail, we annotate drug mentions (`drug`) and any description of medical signs or symptoms, no matter whether or not they are an ADR (`disorder`). We further annotate trigger words or phrases that mark a change in medication intake (`change_trigger`) as well as mentions describing any part of the body (`anatomy`). Next, we mark body functions (`function`), i.e., normal processes of the body, like "sleep" or "appetite", which can sometimes be negated, similar to disorders. Medical tests (`test`) and resulting measurements or medication dosages (`measure`) are labeled, too, as well as the means of medication intake (`route`). Further, we mark the assessment and evaluation of patients with respect to drugs, disorders, or functions using the entity type `opinion`. Since timelines are also an essential concept in medication intake, we apply a label called `time` to any mention expressing a time, e.g., a duration or a frequency. Finally, doctors' professions are labeled with `doctor`, and all remaining entities that seem relevant to the annotator can be marked with `other`.

Attributes Some entities are extended by attributes. For example, the `drug` entity can be further specialized by adding a marker that represents the current state of the drug, e.g., whether it was recently started or stopped by the patient (or by prescription). Mentions of body functions and disorders can be negated, for example, in the case the medication helped the patient and the described symptoms do not exist anymore. Patients' opinions on drugs or disorders can be attributed as positive, negative, or neutral. Lastly, time expressions can be marked as describing, e.g., a frequency or a duration.

Relations The most important relation type is `caused` and differentiates our corpus from those of other work described in Section 2.2: We do not

[8] See Appendix D for more information on the annotation process and our annotators.

relation type	head	tail
caused	drug, disorder	disorder, function
treatment_for	drug	disorder, function
has_dosage	drug	measure
experienced_in	disorder	anatomy
examined_with	disorder, anatomy, function	test
has_result	test	measure, disorder, function
refers_to	disorder	disorder, function^{negated}
refers_to	drug	drug
refers_to	anatomy	anatomy
refers_to	function	function
interacted_with	drug	drug
signals_change_of	change-trigger	drug
has_time	drug, disorder	time
has_route	drug	route
is_opinion_about	opinion	drug, disorder, function
misc	ANY	ANY

Table 2: Overview of available relation types and the entity types they associate. 'ANY' stands for any entities we defined.

specifically mark mentions of ADRs with an "ADR" label but only express ADRs with the `caused` relation between medications and symptoms (or body functions). This relation type can also be used to mark disorders that are the reason for other disorders or body functions. Further, we represent treatments of medical signs or symptoms with the `treatment_for` relation. Medications and their routes and dosages can be connected via the types `has_route` and `has_dosage`, respectively. To connect medical symptoms with a body part, we introduce the relation type `experienced_in`. Moreover, disorders, body parts, and functions can be `examined_with` a test. Those tests can have results (`has_result`) in the form of measures (like a certain cholesterol value), but also in the form of diagnoses, expressed as disorders or functions. In case there is evidence that a medication interacted with another one, this can be modeled using the `interacted_with` relation type. Triggers of medication change can be associated with the `signals_change_of` relation to a drug mention. Furthermore, drugs and disorders can be connected to time expressions with the `has_time` relation, to mark the time of medication intake or the duration of a symptom. To represent assessments by patients concerning drugs, disorders, and functions, we introduce the `is_opinion_about` type. Finally, we add a `refers_to` relation type to connect co-referring mentions, e.g., in case patients first mention a medication name and afterward only an abbreviation of it. All associations that seem relevant to the annotators but are not represented in our annotation scheme can be modeled with the `misc` relation. We *do not* annotate relations if the relevant entities are part of a hypothetical or speculative statement or a question. See examples of annotated texts in Figure 1 and Figure 2.

3.3. Final Dataset

In total, the corpus contains 118 texts in German with a minimum of 55 tokens per text, 100 texts in French with at least 42 tokens per text, and 619 texts in Japanese, with the shortest text containing 15 characters. See Figure 3 for the distribution text lengths per language. The number of entities, relations, and attributes per language is shown in Table 3, and more detailed statistics are shown in Table 8 (Appendix E). The German data were annotated by two annotators and subsequently consolidated, while the other two datasets were annotated by one annotator each.

language	#doc	#ent	#rel	#attr
de	118	3,487	2,163	1,141
fr	100	1,939	1,129	537
ja	619	9,464	5,083	2,364

Table 3: Number of documents (#doc) with the number of entities (#ent), relations (#rel), and attributes (#attr) of each type for each language (lang.).

German The by far most frequent entity type for the German part of the data is `disorder` (1,151 annotations), followed by `drug` (642 annotations). The entity type with the lowest frequency for German is `route`. Furthermore, `caused` is the most often annotated relation type. For attributes, we find that the `time` attribute was used quite often by the annotators (622 times), mostly with `duration` or `point in time` as attribute values. The inter-annotator agreement for the German data was a micro average (relaxed) F_1 score of 0.77 for entities (with `drug` and `disorders` showing an agreement of 0.93 and 0.84, respectively), 0.38 for relation types (the annotators agreed on the `caused` relation with a score of 0.60) and 0.41 for attributes. Note that the relation and attribute annotation was conducted in the same session as the entity annotation, so disagreements in the entity annotation were propagated to the other layers.

French For French, the most often annotated entity type is `disorder` (588 mentions). Similar to

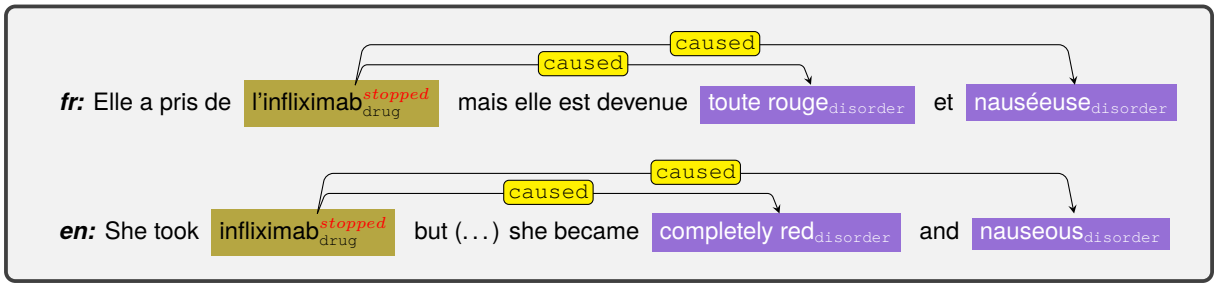


Figure 2: An example annotation of a `caused` relation taken from the French dataset and translated. According to the writer of this message (patient), the medication *infliximab* is likely to have caused the symptoms *toute rouge* and *nauséeuse*.

the German data, it is followed by `drug`. `test` is the entity type with the fewest annotations. The `caused` relation is annotated in the French data 342 times, showing the highest frequency. Also, `has_time` has the second-highest frequency for French. Attributes for time expressions and medication mentions are used the most.

Japanese The distribution of the entities in the Japanese data follows the same pattern as German and French: `disorder` and `drug` are the most frequent types, with 2,843, and 2,031 occurrences, respectively. Doctors' names are listed last in terms of frequency. Regarding relations, we again see the same pattern as in the other two languages: `caused` and `has_time` are the most frequent relation types with 979 and 753 annotations, respectively. For Japanese, attributes for `opinion` and `time` occur most often.

drug	disorder (de)	translation (en)
ads	Gelenkschmerzen	joint pain
estreva gel	vermehrte, starke Kopfschmerzen	increased severe headaches
cerazette	3kilo runter	3 kilos down
opipramol	Watte im Kopf	"cotton in my head"
mtx	Haarausfall	hair loss
venaflaxin	Unwirklichkeits-gefühle	feelings of unreality
utrogest	wilde Träume	wild dreams

Table 4: A random selection of ADRs (i.e., medication mentions that have caused disorders) extracted from the German part of the corpus.

In summary, we can see similar patterns in the annotations across languages, e.g., the distribution of entity types. However, the Japanese dataset is much larger than the German and French datasets, which is also evident in the number of entity and relation annotations in total. Apart from being less distant languages, the French and German data originate from the same source, and this also shows in the distribution of annotations.

Due to the colloquial style of the text, it was sometimes difficult for the annotators to pinpoint the boundaries of entities since some descriptions are rather "creative" or even metaphorical (see Table 4). Similarly, finding the exact attribute for, e.g., a medication mention, was challenging as well since patients often describe starting and stopping their medication intake in one and the same sentence. For illustration, we show some extracted information from the German part of the corpus in Table 4, together with their translations. Note that these extracted phrases still need to be verified by a pharmacovigilance expert.

4. Experiments

The following describes our baseline models, which will be published together with the data and are meant to serve as a starting point for future work. We show experiments on named entity recognition (NER), attribute classification (AC), and relation extraction (RE).

We evaluate the performance of the models detailed below using BRAT format to account for span boundaries independently of the tokenizer. Therefore, for NER, we convert the predictions of the models back to BRAT and evaluate them using "brat eval"[9]. We then report micro and macro average F_1 scores for all tasks, calculated on relaxed boundaries for NER.

4.1. Task Setup

We carried out the three tasks independently. The hyper-parameter tuning is performed for each dataset combination for each task.

Name Entity Recognition (NER): The dataset includes discontinuous and overlapping entity annotations (see Table 9 in Appendix E for details), and preparing these annotations for model fine-tuning requires complex methods (Baldini Soares et al.,

[9] https://perso.limsi.fr/pz/blah2015/.

2019; Dai et al., 2020; Dirkson et al., 2021; Li et al., 2021). Since these special entity annotations are infrequent, we remove sentences containing them during model fine-tuning. This helps prevent models from encountering and potentially struggling with special annotations without employing more complex handling methods. We convert the BRAT annotations to *BILOU* format (Ratinov and Roth, 2009) for fine-tuning. For evaluation, we convert the predicted *BILOU* tags back to BRAT format.

Attribute Classification (AC): Similar to RE, we extract the sentences covering the entity corresponding to the attribute and use the special token pair `[E]`, `[/E]` to mark the entity.

Relation Extraction (RE): To prepare each relation sample from the document-level annotations, we extract only the sentences containing entities from the documents. We use special token pairs `[E1]`, `[/E1]` and `[E2]`, `[/E2]` (Baldini Soares et al., 2019) to enclose the head and tail entities.

4.2. Experiment Setup

For all experiments, we fine-tune XLM-RoBERTa$_{large}$ (Conneau et al., 2020), henceforth XLM-R, on the respective downstream task. The model supports French, German, and Japanese, among other languages. The different settings are aimed at investigating the performance within and across languages.

Mono-lingual: We fine-tune and test XLM-R on each language of the dataset separately, for French, German, and Japanese, respectively.

Multi-lingual: We mix the languages while fine-tuning; each batch samples from each language proportionally to the size of this language in the training set. The fine-tuned multilingual models are evaluated on each language separately and across languages.

Cross-lingual: We apply the model in a zero-shot cross-lingual transfer setting, i.e., we (1) fine-tune XLM-R on the source language(s) and (2) directly test the model on the target languages.

4.3. Results

The results are described in the following and shown in Table 5.

Mono-lingual: Regarding the results for AC and RE, we see that the **ja** models perform the best and are closely followed by **fr** with **de** being last with a larger difference in scores. This, in turn, follows the pre-training data size of XLM-R[10]. However, for NER, the performance is **fr** $>$ **de** $>$ **ja**, with **ja** falling to the last place.

Multi-lingual: In general, the multilingual models fine-tuned on all languages boost performance across all tasks and languages, except for **ja** in RE. When comparing with the monolingual experiments, **fr** outperforms marginally in AC and RE, benefiting from the contributions of the other two languages.

Cross-lingual: We observe that the models fine-tuned on **de** and evaluated on **fr** work well and only show a modest decrease from the monolingual models trained on **fr** (-1% for RE; -3% for NER; -10% for AC macro F_1). The models fine-tuned on **de** and evaluated on **ja** are still far behind the monolingual model trained on **ja** only ($-20 \sim 30\%$ macro F_1). When comparing models fine-tuned on **de+ja** and evaluated on **fr** to models fine-tuned on **de** only, we observe consistent improvements in micro F_1 across the three tasks, but a drop in macro F_1 for NER and RE.

5. Discussion & Conclusion

With this work, we provide a new corpus of texts in German, French, and Japanese to support pharmacovigilance across languages by extracting information on ADRs from user-generated content. Training models on this corpus might facilitate information aggregation across countries, which is important for detecting rare diseases or adverse reactions. Furthermore, gathering and analyzing data globally can help develop new medications or treatments and benefit minorities. The corpus is annotated based on annotation guidelines carefully designed to apply to German, French, Japanese, and also English, potentially allowing the guidelines to be used for other languages as well. Annotations are conducted on entity, attribute, and relation levels to cover as much information as possible. By choosing languages from different language families and cultures, we provide a challenging resource with which we hope to advance the detection of ADRs and other medically relevant expressions.

To initiate future work, we provide baseline models for all three tasks, i.e., named entity recognition, attribute classification, and relation extraction, highlighting the difficulties of state-of-the-art Transformer models when faced with complex domain-specific data.

[10] Based on the CommonCrawl Corpus (Wenzek et al., 2020), the order in terms of data size is **ja** (69.3 GiB) $>$ **de** (66.6 GiB) $>$ **fr** (56.8 GiB).

Experiments	*train*	*test*	NER (%)		AC (%)		RE (%)	
			micro F1	macro F1	micro F1	macro F1	micro F1	macro F1
Mono-lingual	de	de	75.8	65.4	76.8	56.9	79.3	75.7
	fr	fr	82.5	71.9	84.4	73.8	87.0	78.2
	ja	ja	61.0	58.5	85.8	81.0	87.2	80.4
Multi-lingual	de+fr+ja	de	77.3	67.6	80.4	66.9	83.4	79.2
	de+fr+ja	fr	83.9	75.3	90.8	82.8	88.3	82.0
	de+fr+ja	ja	64.5	65.1	88.0	82.6	86.5	78.0
	de+fr+ja	de+fr+ja	74.1	69.3	85.8	71.7	85.9	76.7
Cross-lingual	de	fr	77.3	68.8	69.5	63.6	78.7	79.3
	de	ja	48.8	38.8	53.7	41.3	62.2	54.5
	de+ja	fr	77.5	66.7	80.8	71.2	83.2	75.9

Table 5: Average scores of models fine-tuned on five different seeds on the KEEPHA dataset with different language combinations. The underlying pre-trained model for all experiments was XLM-RoBERTa$_{large}$.

Further future work will focus on improving the cross-lingual performance of available models, for example in combination with few-shot approaches and/or large language models, such as Llama (Touvron et al., 2023). A more detailed investigation into the impact of the different data sources on the overall performance of the models might further deepen the understanding of the data, too. Moreover, investigating cross-cultural differences in how people discuss their health issues online is an exciting topic to explore. Building on the work of Scaboro et al. (2022), who analyzed negation and speculation constructions, examining specific syntactic structures and linguistic phenomena, as well as potential biases in the new corpus would be interesting, too. Finally, we are already working on extending our corpus with more data. For instance, we are annotating Japanese case reports and social media messages containing a more diverse pool of medications. Moreover, we aim to gain access to more (original) French data to diversify the dataset even more. Normalizing disease descriptions to medical ontologies will be one of the next steps as well.

6. Ethical Considerations

When using data from social media, we commit to a particular sub-group of people: Those who have access to and actively participate on these platforms. Depending on the platform, the age range might vary, too. Again, this introduces a bias, which can be learned by the respective language models fine-tuned on these data.

Also, the presented dataset is only a small glimpse of ADR-related topics discussed online. The German and French parts of the corpus, particularly, are very similar due to the translation. More different sources and languages need to be considered to make the dataset more diverse.

Several ethical aspects need to be considered when creating the dataset. First, the de-identification might not be perfect, i.e., even if usernames, etc., are masked, it might still be possible to identify the users since the fora are publicly accessible. The corpus will only be distributed via a data protection agreement and only within the research community. Second, the extracted information should not be further processed as is but instead verified by a pharmacovigilance expert. One mention of a potential ADR in a user post does not make an ADR per se, but this information should be further investigated. Related to this, normalizing user descriptions to medical ontologies would also make it easier for experts to analyze potential health risks. Also, the automatic translation of the German texts into French might have introduced some biases.

Regarding the language model we used to conduct the baseline experiments (XLM-RoBERTa$_{large}$), we cannot rule out the existence of sensitive contents in the pre-training data, which might also have introduced biases into the models.

7. Acknowledgements

First and foremost, a heartfelt 'thank you' to our annotators Alon Drobickij, Selin Yeginer, Garance Forestier, Emiliano Valdes Menchaca, and Narumi Tokunaga, for doing amazing work. We further thank the anonymous reviewers for their constructive feedback on this paper. Our work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, by JPMJCR20G9, ANR-20-IADJ-0005-01, and DFG-442445488 under the trilateral ANR-DFG-JST AI Research project KEEPHA, and by the German Federal Ministry of Education and Research under the grant BIFOLD24B.

8. Bibliographical References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Shaika Chowdhury, Chenwei Zhang, and Philip S. Yu. 2018. Multi-Task Pharmacovigilance Mining from Social Media Posts. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 117–126, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An Effective Transition-based Model for Discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

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*, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2021. FuzzyBIO: A Proposal for Fuzzy Representation of Discontinuous Entities. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 77–82, online. Association for Computational Linguistics.

I. Ralph Edwards and Jeffrey K. Aronson. 2000. Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet*, 356(9237):1255–1259.

Ze-Ying Feng, Xue-Hong Wu, Jun-Long Ma, Min Li, Ge-Fei He, Dong-Sheng Cao, and Guo-Ping Yang. 2023. DKADE: A novel framework based on deep learning and knowledge graph for identifying adverse drug events and related medications. *Briefings in Bioinformatics*, page bbad228.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*, 1 edition. Wiley.

Oguzhan Gencoglu. 2020. Sentence Transformers and Bayesian Optimization for Adverse Drug Effect Detection from Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 161–164, Barcelona, Spain (Online). Association for Computational Linguistics.

Eszter Hargittai and Gina Walejko. 2008. The participation divide: Content creation and sharing in the digital age1. *Information, Communication & Society*, 11(2):239–256.

Lorna Hazell and Saad A. W. Shakir. 2006. Underreporting of adverse drug reactions : A systematic review. *Drug Safety*, 29(5):385–396.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts in Health-Related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden. Association for Computational Linguistics.

Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A Span-Based Model for

Joint Overlapped and Discontinuous Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4814–4828, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.

Zulfat Miftahutdinov, Andrey Sakhovskiy, and Elena Tutubalina. 2020. KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 51–56, Barcelona, Spain (Online). Association for Computational Linguistics.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in Bioinformatics*, 22(1):146–163.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Caterina Palleria, Christian Leporini, Serafina Chimirri, Giuseppina Marrazzo, Sabrina Sacchetta, Lucrezia Bruno, Rosaria M. Lista, Orietta Staltari, Antonio Scuteri, Francesca Scicchitano, and Emilio Russo. 2013. Limitations and obstacles of the spontaneous adverse drugs reactions reporting: Two "challenging" case reports. *Journal of Pharmacology & Pharmacotherapeutics*, 4(Suppl1):S66–S72.

Beatrice Portelli, Simone Scaboro, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. AILAB-Udine@SMM4H'22: Limits of Transformers and BERT Ensembles. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 130–134, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL '09*, page 147, Boulder, Colorado. Association for Computational Linguistics.

Shivam Raval, Hooman Sedghamiz, Enrico Santus, Tuka Alhanai, Mohammad Ghassemi, and Emmanuele Chersoni. 2021. Exploring a Unified Sequence-To-Sequence Transformer for Medical Product Safety Monitoring in Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3534–3546, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *EMC$\hat{2}$: 5th Edition Co-located with NeurIPS'19*.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Simone Scaboro, Beatrice Portelli, Emmanuele Chersoni, Enrico Santus, and Giuseppe Serra. 2022. Increasing adverse drug events extraction robustness on social media: Case study on negation and speculation. *Experimental Biology and Medicine (Maywood, N.J.)*, page 15353702221128577.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics (Oxford, England)*, 37(2):243–249.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017*, page 11.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ryen W. White, Sheng Wang, Apurv Pant, Rave Harpaz, Pushpraj Shukla, Walter Sun, William DuMouchel, and Eric Horvitz. 2016. Early identification of adverse drug reactions from search log data. *Journal of Biomedical Informatics*, 59:42–48.

Christopher C. Yang, Haodong Yang, Ling Jiang, and Mi Zhang. 2012. Social media mining for drug safety signal detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, SHB '12, pages 33–40, New York, NY, USA. Association for Computing Machinery.

9. Language Resource References

Ilseyar Alimova, Elena Tutubalina, Julia Alferova, and Guzel Gafiyatullina. 2017. A machine learning approach to classification of drug reviews in Russian. In *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*, pages 64–69.

Yuki Arase, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Annotation of adverse drug reactions in patients' Weblogs. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6769–6776.

Masaru Kamba, Masae Manabe, Shoko Wakamiya, Shuntaro Yada, Eiji Aramaki, Satomi Odani, and Isao Miyashiro. 2021. Medical Needs Extraction for Breast Cancer Patients from Question and Answer Services: Natural Language Processing-Based Approach. *JMIR Cancer*, 7(4):e32005.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In *Fifth Social Media Mining for Health Applications(#SMM4H) Shared Tasks at COLING 2020*, page 10.

Alejandro Metke-Jimenez, Sarvnaz Karimi, and Cecile Paris. 2014. Evaluation of text-processing algorithms for adverse drug event extraction from social media. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 15–20, New York, NY, USA. Association for Computing Machinery.

Lisa Raithel, Philippe Thomas, Roland Roller, Oliver Sapina, Sebastian Möller, and Pierre Zweigenbaum. 2022. Cross-lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3637–3649, Marseille. European Language Resources Association.

Alexander Sboev, Sanna Sboeva, Ivan Moloshnikov, Artem Gryaznov, Roman Rybka, Alexander Naumov, Anton Selivanov, Gleb Rylkov, and Vyacheslav Ilyin. 2022. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models. *Applied Sciences*, 12(1):491.

Isabel Segura-Bedmar, Ricardo Revert, and Paloma Martínez. 2014. Detecting drugs and adverse events from Spanish social media streams. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 106–115, Gothenburg, Sweden. Association for Computational Linguistics.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics (Oxford, England)*, 37(2):243–249.

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.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, Hamideh Moayyed, and Somaieh Goudarzvand. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90:103091.

A. Related Datasets

Other non-English social media corpora focused on ADRs. The number of documents (#documents) refers to the definition of documents per corpus, i.e., some are sentence-based, some are post-based, etc. Some test sets are unavailable to the public since they are/were part of a shared task.

language	#documents	type	annotation	authors
es	400	forum	entities	Segura-Bedmar et al. (2014)
ru	*279	drug reviews	multi-label	Alimova et al. (2017)
fr	3,033	Twitter	binary	Klein et al. (2020)
ru	9,515	Twitter	binary, entities	Klein et al. (2020)
ja	169	forum	entities, normalization	Arase et al. (2020)
ru	**500	drug reviews	multi-label, entities	Tutubalina et al. (2021)
ru	2,800	drug reviews	entities	Sboev et al. (2022)
de	4,169	forum	binary	Raithel et al. (2022)
de, fr, ja	837	forum, Twitter, YJQA	entities, attributes, relations	ours

Table 6: Other non-English social media corpora focused on ADRs. es=Spanish, fr=French, ru=Russian, ja=Japanese, de=German. *Number of documents containing ADRs. **This is only the annotated part of the RuDReC corpus.

B. Details on German and French Data

To avoid potential confusion, we provide an explanation for the relation between the Lifeline corpus (German) provided by Raithel et al. (2022) and the data presented in this paper.

French The Lifeline corpus (Raithel et al., 2022) contains 4,169 documents in German, crawled from the patient forum lifeline.de. These documents are labeled with binary classes, i.e., either a document mentions an adverse drug reaction or does not. We took these documents and automatically translated them into French. The translations were validated and improved (if necessary) by French speakers (our annotators). We then took the 100 positive documents that remained after validation (those containing mentions of ADRs) and prepared them for further annotation. These documents, therefore, overlap with the positive documents in Raithel et al. (2022) but are semi-automatically *translated*.

German The German documents presented in this paper *do not* overlap with the data in Raithel et al. (2022), but they are extracted from the same forum.

De-Identification Regarding the masking of identifying information, for French and German, the following applies: The German data is from a public anonymous forum, so they can be found already publicly on the web. Of course we have to acknowledge that often, people on any kind of social media are not necessarily aware of the potential reach of their posts. However, by masking details, we try to cut the connection between our documents and the original forum posts to make it more difficult to trace the documents back to their original.

As the French were a translation of part of the already de-identified German data, these documents were subject to the same de-identification procedure, plus an additional modification of the documents by translation. Even with a translation of the documents back to German, it is difficult to trace the documents back to their original.

C. Details on Japanese Data

The Japanese tweet documents contain 20 tweets per document, and each Q&A document includes one question and one answer text, resulting in 99 full documents as shown in Table 3. The definition of "token" or "word" in Japanese may change according to different Japanese grammar theories. We can estimate what space-delimited languages call 'the number of "words"' based on the average character counts in Japanese "words", which more-or-less span 2.5–3.5 characters.

De-Identification We have two sources for Japanese: Twitter (currently X) and an online forum. We applied basic regular expressions to identify Twitter user names and URLs in our data. Then, we manually de-identified all potential mentions of private information, such as names of persons, hospitals, and organizations.

D. Annotation Process and Annotators

Annotation Process For annotation, we used the widely known annotation tool BRAT, since almost everyone on the team was familiar with it, it allows the annotation of attributes, and it was furthermore favorably reviewed by Neves and Ševa (2021).

The guidelines design and pilot annotation were mainly done using English data since we did not have many positive documents in Japanese, French, or German. For all languages, the first annotations were therefore conducted by the annotators and some of this paper's authors on English texts, with data taken, e.g., from the CADEC corpus (Karimi et al., 2015). We annotated examples in two rounds for three months and discussed/refined the entity and relation scopes. When these pilot annotations reached a satisfactory IAA, we asked our annotators to label the data in the respective languages. During annotation, whenever problems occurred, the annotator discussed with the annotation instructors (the first and third author), which involved a small number of working-level modifications to the guidelines. We generally followed the state-of-the-art methodology recommended by Fort (2016).

Annotators Table 7 lists our annotators. Each annotator except one (A5) in Table 7 is an enrolled student and employed as a student assistant with varying working hours, depending on their availability. Annotators A1 to A4 earn(ed) 12,95€ per hour. Annotators can distribute their working hours freely, with the recommendation to annotate not more than two hours continuously. Annotator A5 is a full-time employee who annotates Japanese corpora.

annotator	working language	knowledge of languages	study program	entry	working hours
A1	de, en	German and Russian bilingual, good knowledge of English	Pharmacy, Freie Universität Berlin, Germany	November 2021	10
A2	de, en	German and Turkish bilingual, good knowledge in English, basics in French and Spanish	Pharmacy, Freie Universität Berlin, Germany	March 2022	10
A3	fr, de	French (native), German (C1), English (C1)	Human Medicine, Charité Berlin, Germany	May 2023	8
A4	fr, de	Spanish (native), French (C2), German (C1), good knowledge of English	Life Science Engineering, HTW Berlin, Germany	August 2023	20
A5	ja	Japanese (native) with good knowledge of English	(MSc in Biomedicine)	April 2021	full-time

Table 7: The background information of our annotators. The table shows the languages they were working on, the languages they know in general, their study programs (or obtained degree), the time they were hired, and their working hours per week.

E. Dataset Statistics

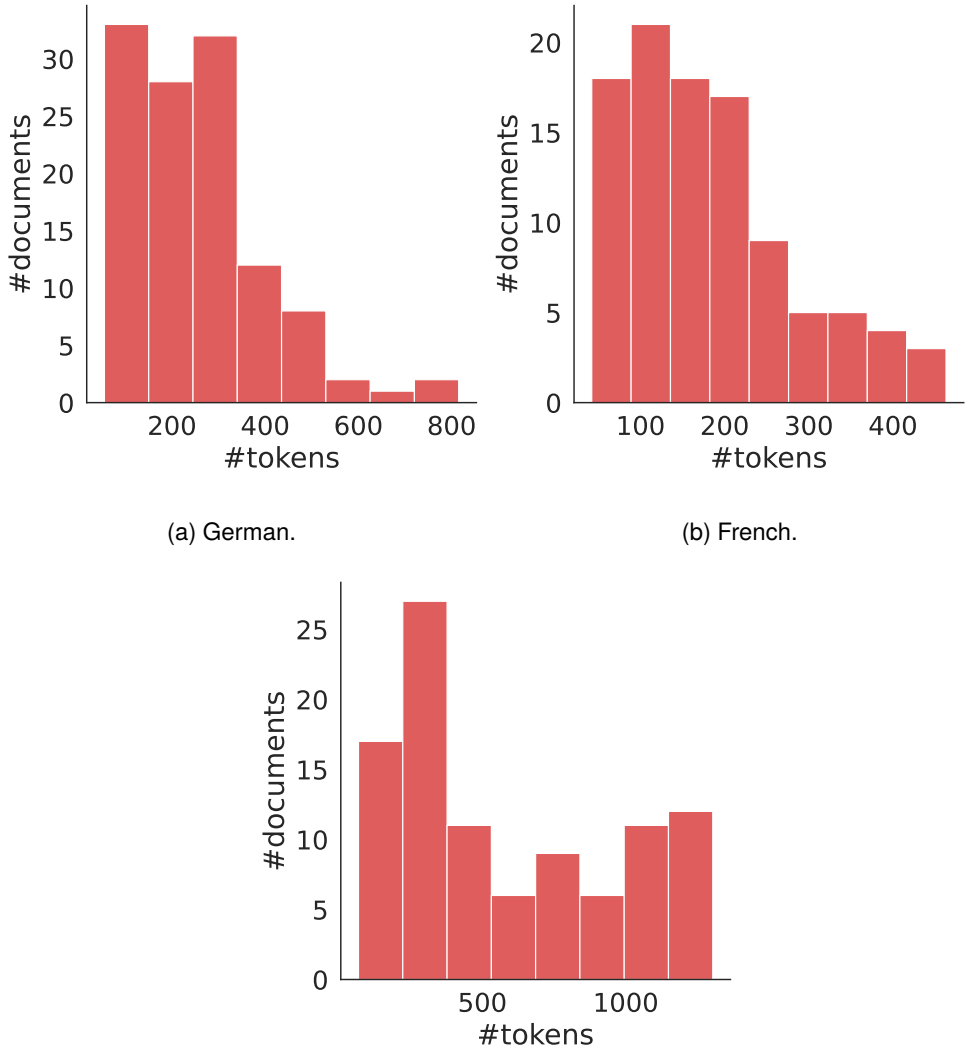
Table 8 shows detailed statistics of the presented dataset. Table 9 shows the number of complex entity annotations. We further show the distribution of document length in Figure 3, the number of mentions per entity type in Figure 4, the span lengths per entity type in Figure 5, the number of relations per type in Figure 6, and the distribution of attribute values in Figure 7.

	#docs	#tokens total	#tokens mean	#tokens max	#tokens min	#sentences total	#sentences mean	#sentences max	#sentences min
German	118	29,032	246.03	815	55	1,674	14.19	50	1
French	100	18,184	181.84	463	42	969	9.69	25	1
Japanese	99	58,024	586.10	1,303	68	2,165	21.87	58	4

Table 8: An overview of the currently annotated data in German, French, and Japanese. It shows the number of documents for each language, the total number of tokens and sentences, and the mean, minimum, and maximum number of tokens and sentences per document.

	#Discontinuous	#Overlapping	#Total
German	51	5	5346
French	24	0	9454
Japanese	25	10	9818

Table 9: The number of discontinuous and overlapping entity annotations in German, French, and Japanese.

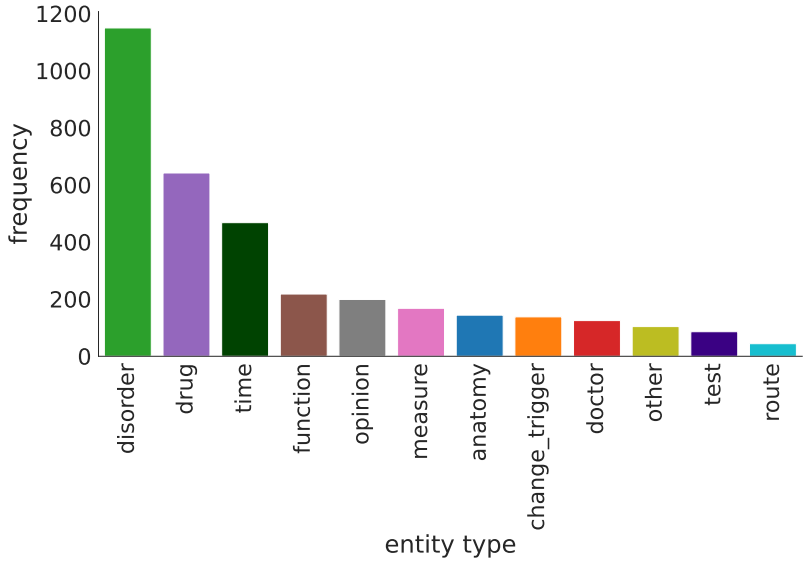


(a) German.

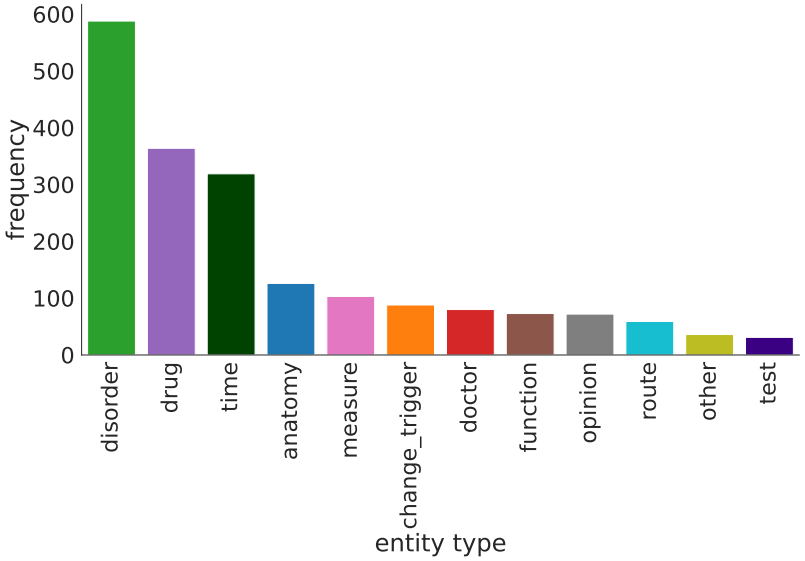
(b) French.

(c) Japanese.

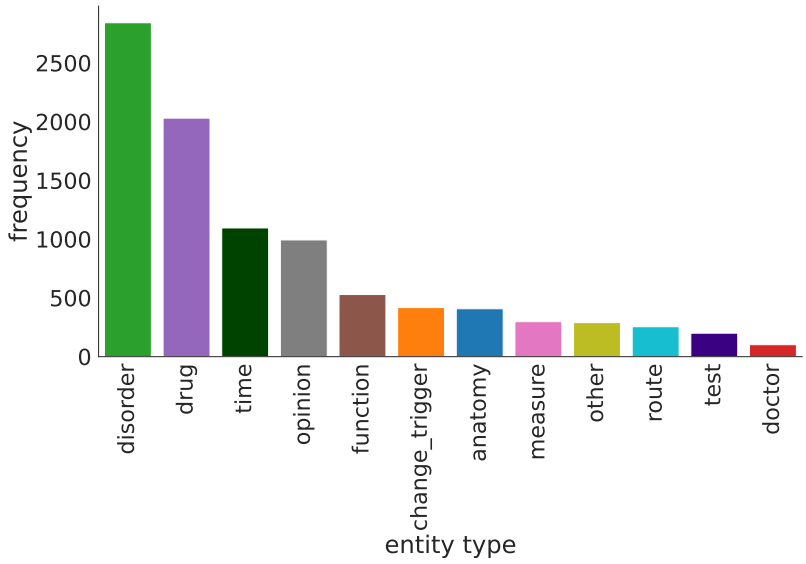
Figure 3: The distribution of document length of the German (a), French (b), and Japanese (c) data using the number of tokens. Note the different scaling on the axes.



(a) German.

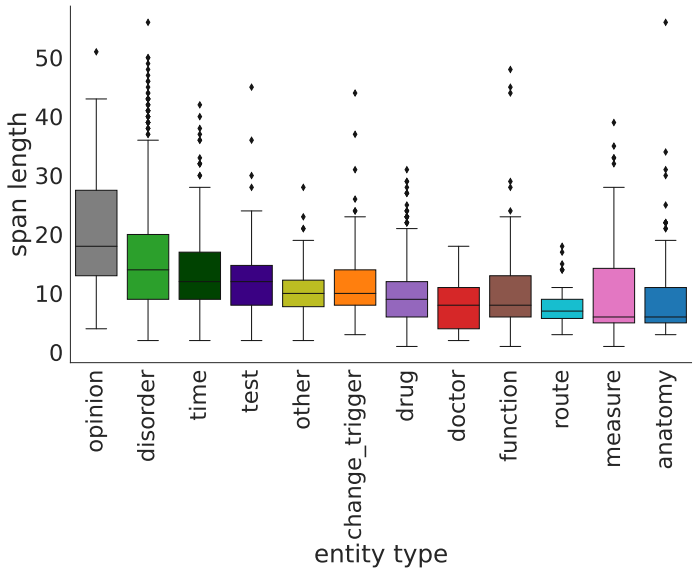


(b) French.

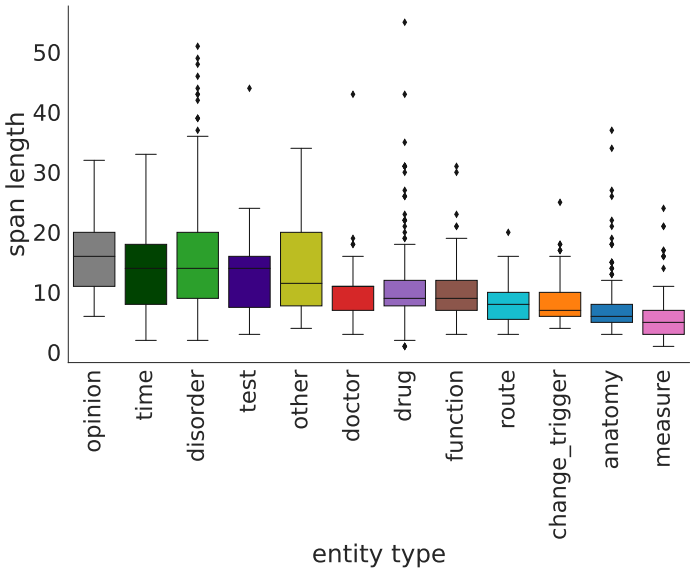


(c) Japanese.

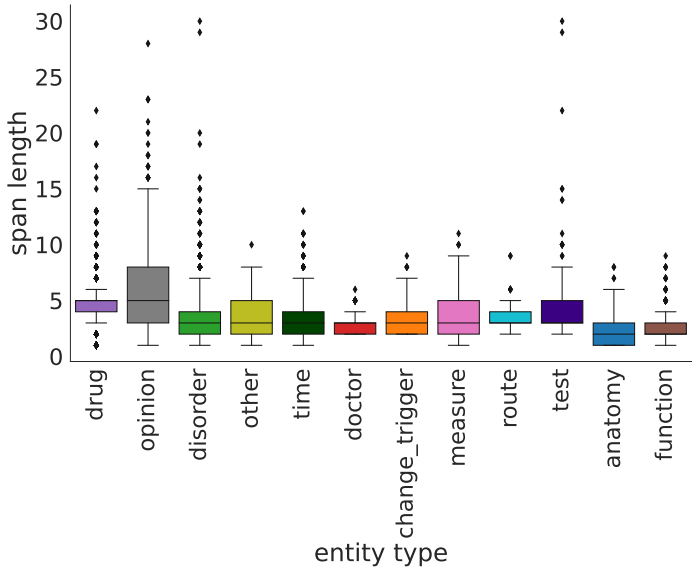
Figure 4: The distribution of entity types across all documents for German (a), French (b), and Japanese (ja). Note the difference in scale when comparing the three languages.



(a) German.

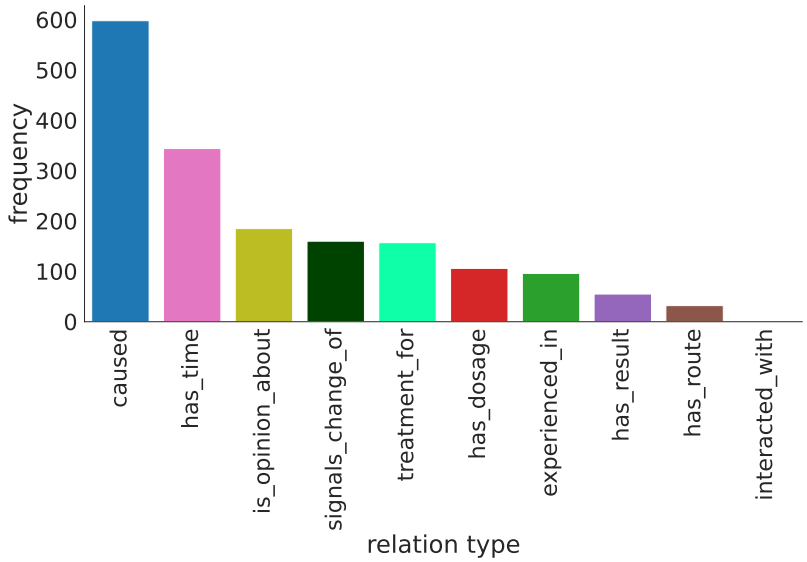


(b) French.

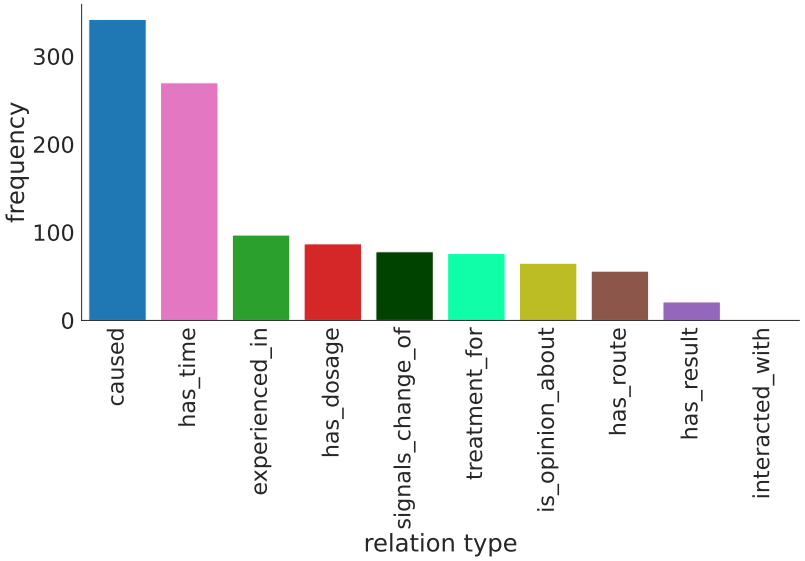


(c) Japanese.

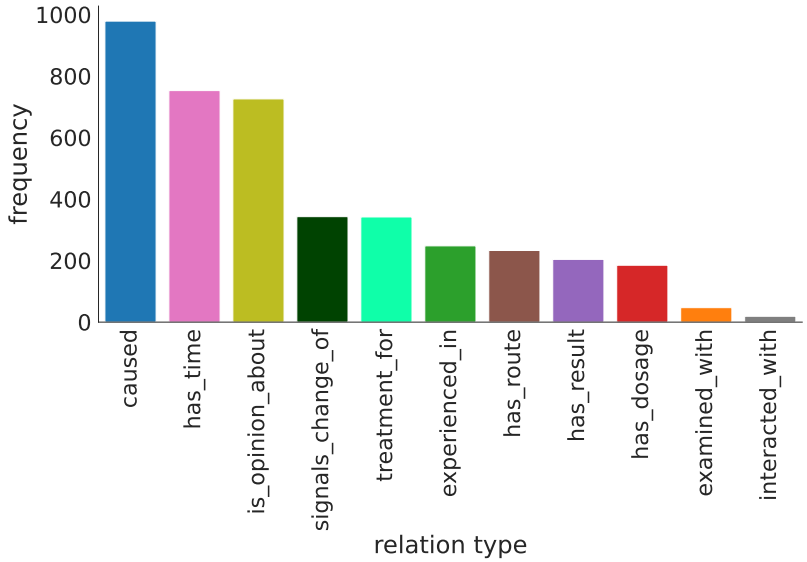
Figure 5: The distribution of span length per entity type for German (a), French (b), and Japanese (c).



(a) German.

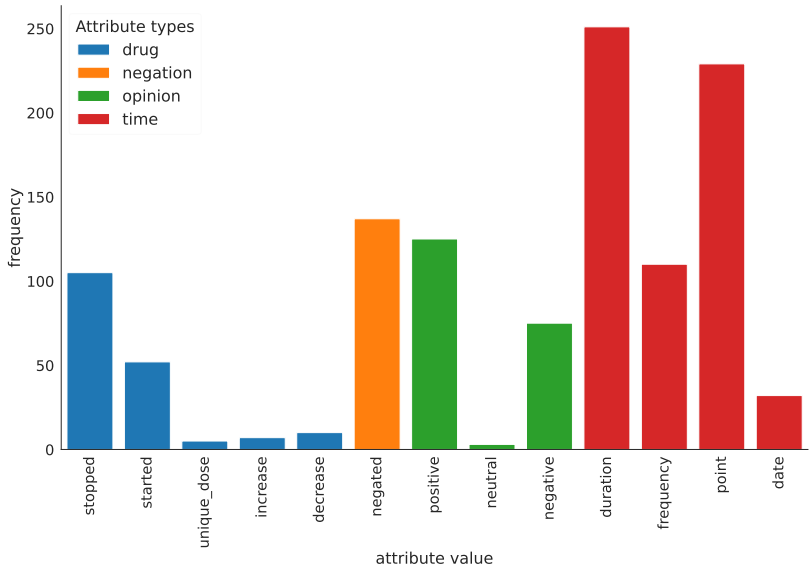


(b) French.

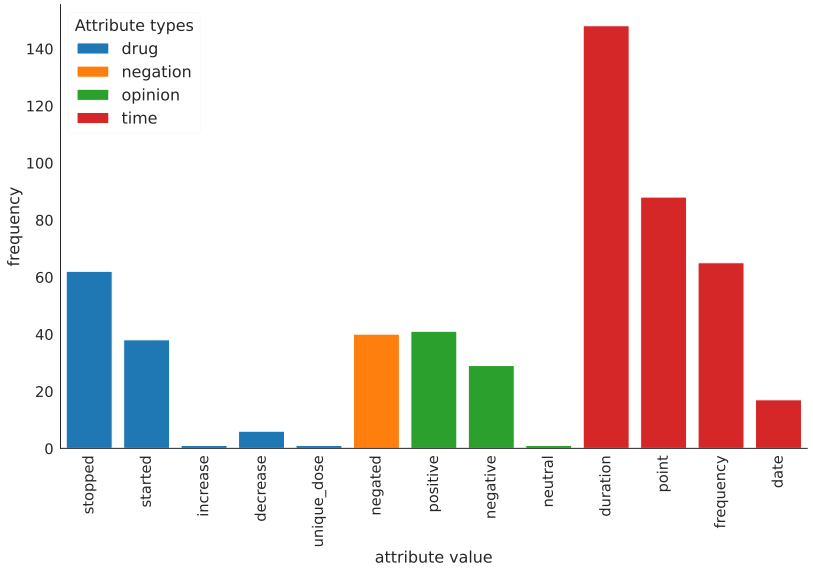


(c) Japanese.

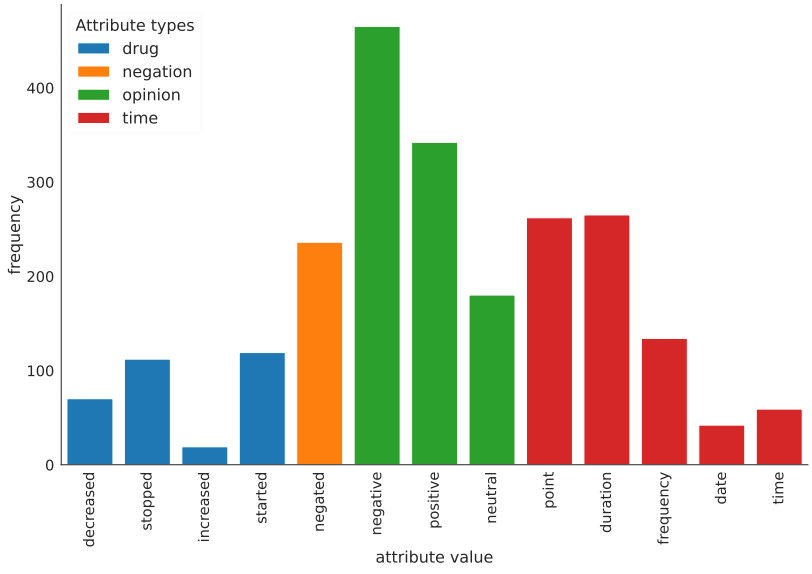
Figure 6: The distribution of relation types for German (a), French (b), and Japanese (c). Note the difference in scale when comparing the three languages.



(a) German.



(b) French.



(c) Japanese.

Figure 7: The distribution of attribute values for each attribute type and for each language: `time` (duration, frequency, point in time, date), `opinion` (positive, neutral, negative), `drug` (stopped, started, unique dose, increase, decrease) and negation (only shown if an expression is negated).