

A Dataset for Pharmacovigilance in German, French, and Japanese

Introduction & Motivation

- Adverse Drug Reactions (ADRs) are a major problem.
- User-generated data is important to cover ADRs worldwide.
- Existing data are mostly scientific articles in English.
- Information extraction can help to discover new ADRs and improve pharmacovigilance.

Contributions

- An annotated corpus in three languages: German, French, and Japanese.
 - The texts come from different sources and were written by laypeople.
 - Our corpus contains annotations covering 12 entity types, 4 attribute types, and 13 relation types.
- Preliminary experiments resulting in strong cross-language baselines for extracting entities and relations between these entities.

Annotation Guidelines for Three Languages

Mentions referring to a drug

Further, mentions referring to a drug but not clearly stating the drug's name are also annotated (see also example 13 (en)).

- (12) (en) The **pillars** does its job **opinion**, (...)
 (fr) (...), j'ai pris **deux** **par j'ai pris**, (...)
 I took **two** **pillars** **of data science**
 (de) Ich nehme die **Tabletten** seit 2 Tagen **opinion**
 I take those **pillars** for two days **opinion**
 (ja) 薬方薬 **を購入して** **飲み始めた** **change.trigger** 所、(...)
 [Once I **started** **change.trigger** taking purchased **Chinese medicine**, ...]

2.3 Disorder

A **disorder** annotation denotes any disease, sign or symptom related to the patient's health, including mental issues. Sometimes a disorder may be expressed as a parameter in combination with a value: e.g., **high LDL** (parameter=LDL, value=high)². When the value is outside the normal range, this describes a disorder. Sometimes, disorders are only referred to very broadly, e.g., it might happen that the patient simply says "I do not feel well"; These expressions are also treated as disorders.

- (13) (en) I **tried** the advertised **Arthritis medicines** with **severe side-effects** and only tried this one because the **doctor** had samples.



paper



guidelines

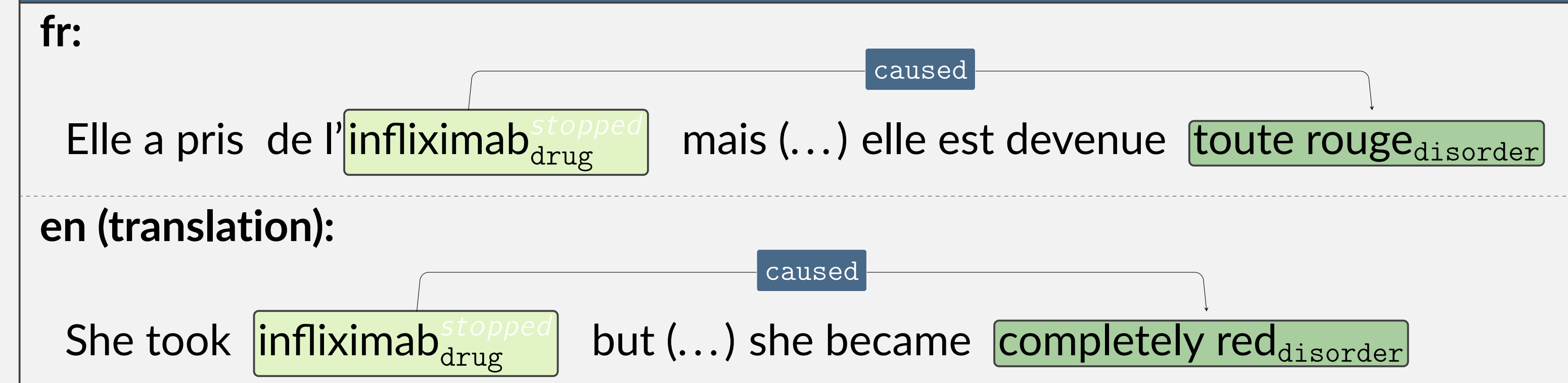
Conclusion & Discussion

- Annotation across languages is possible and feasible: Our guidelines are designed to be applied to other languages as well.
- Our corpus can also be used for tasks other than detection of ADRs.
- Training models on this corpus might facilitate information aggregation across countries for improved pharmacovigilance.
- The baselines demonstrate decent performance:
 - Within languages, performance follows the training and fine-tuning data distribution;
 - Multilingual fine-tuning boosts performance except for Japanese RE;
 - Cross-lingual works better for closer languages.
- However, the dataset seems challenging and calls for further development of methods both within and across languages.

Annotating Adverse Drug Reactions (ADRs) across Languages



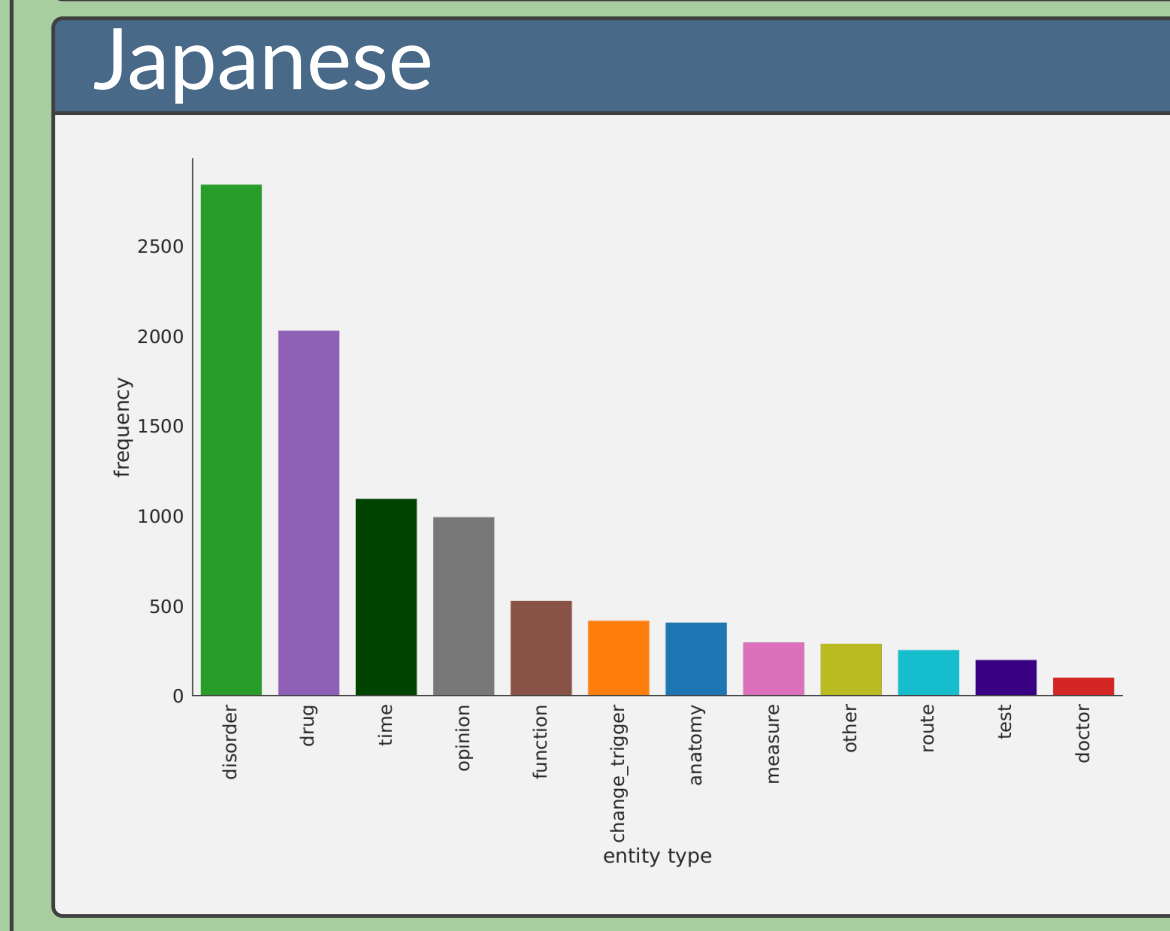
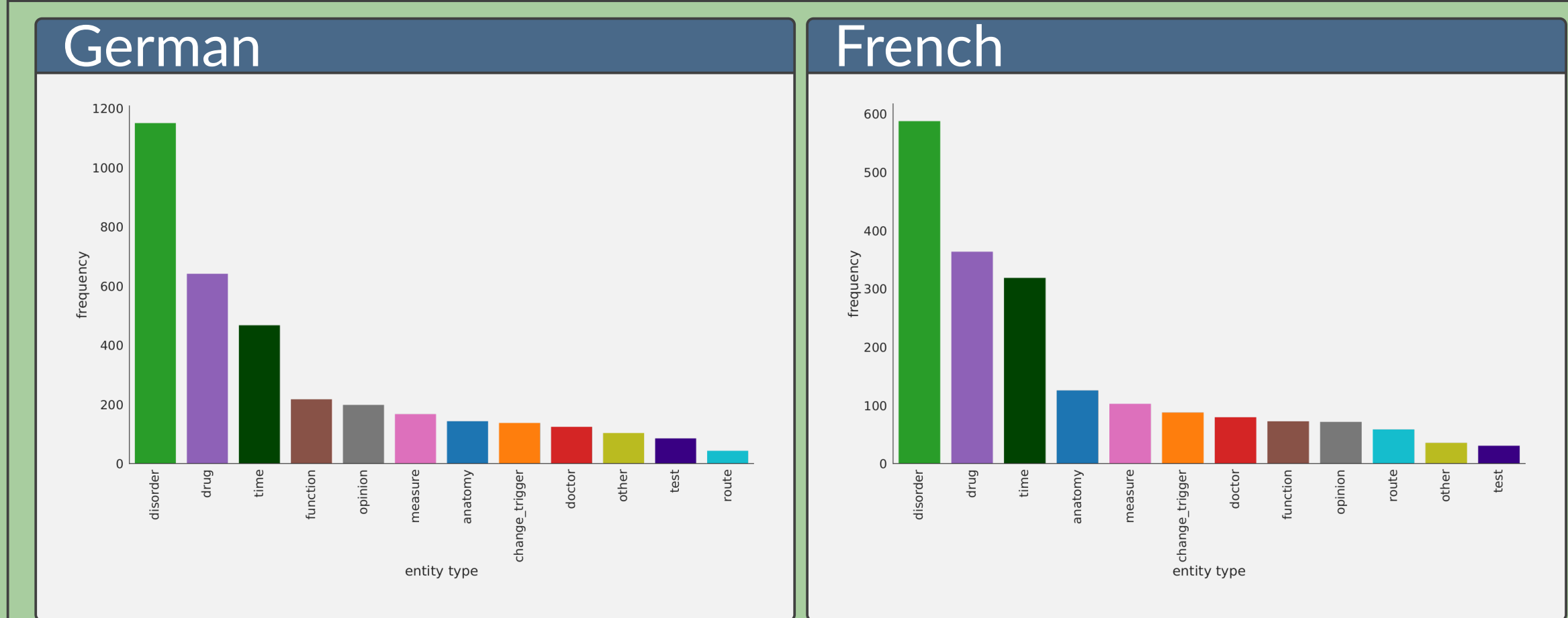
Example Annotation



Overview Corpus

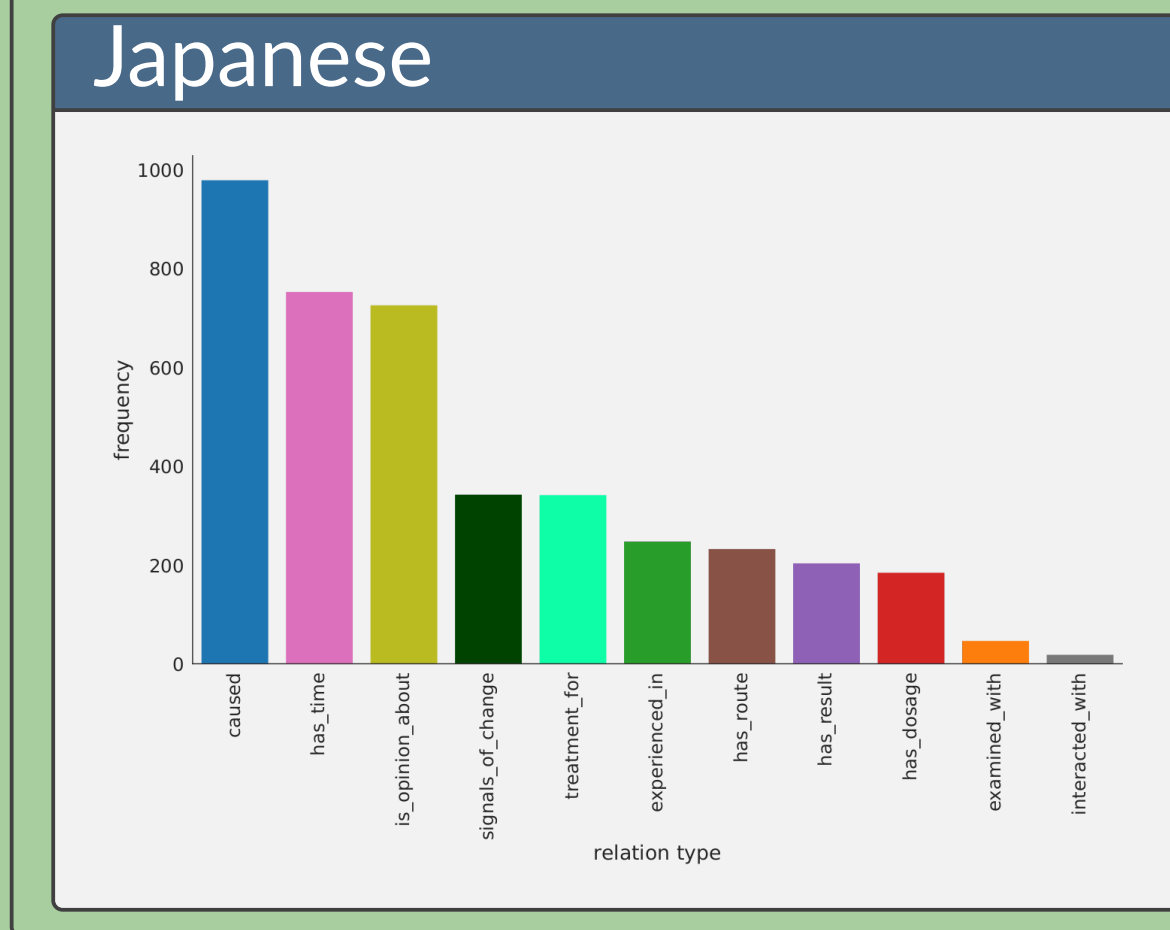
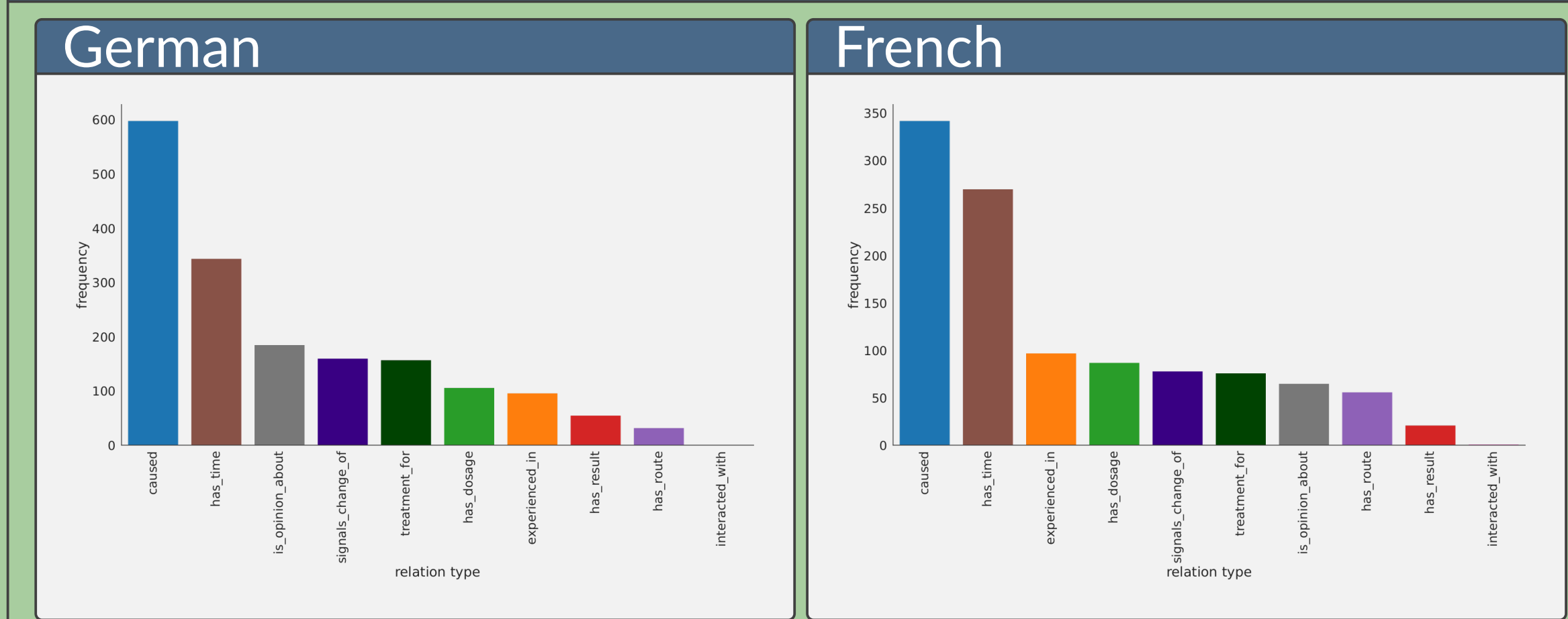
language	#documents	#entities	#relations	#attributes
German	118	3,487	2,163	1,141
French	100	1,939	1,129	537
Japanese	619	9,464	5,083	2,364

Distribution of Entity Types



- Entities are the "bits of information" relevant for us to capture the health state of the patient.
- The distribution of entities is quite similar in all three languages.
- Most important for the detection of ADRs are disorder, function, and drug mentions.
- We also include mentions that describe time expressions, personal opinions of the patients, body parts, medical tests, and other phrases that determine the health of a patient further.

Distribution of Relation Types

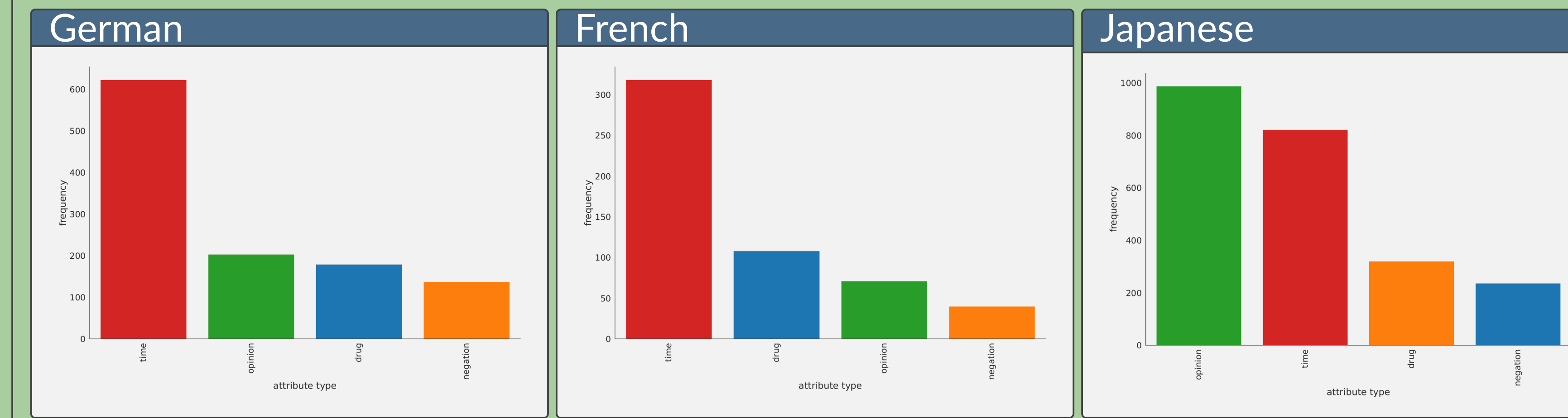


- Relations associate entities with each other.
- ADRs are represented by a **caused** relationship between a drug and a disorder or function.
- Treatments for medical signs and symptoms are represented via a **treatment_for** relation between a drug and a disorder or function.
- Other relations include those defining medication dosages, opinions of patients about a certain treatment, medical test results, etc.

Extracted ADRs (selection)

drug	disorder (de)	translation (en)
ads	Gelenkschmerzen	joint pain
estrema 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
venafloxin	Unwirklichkeitsgefühle	feelings of unreality
utrogest	wilde Träume	wild dreams

Distribution of Attribute Types



- Attributes add even more information to the given entities, e.g., if a medication was just started or stopped or if a patient is content or not about their treatment.

Baseline Results (XLM-Roberta):

Named Entity Recognition (NER) - Attribute Classification (AC) - Relation Extraction (RE)

Setup	train	test	NER (%)		AC (%)		RE (%)	
			micro F1	macro F1	micro F1	macro F1	micro F1	macro F1
mono	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	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
cross	de+fr+ja	de+fr+ja	74.1	69.3	85.8	71.7	85.9	76.7
	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