A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, Aurélie Névéol

Lang Resources & Evaluation
DOI 10.1007/s10579-017-9382-y

ORIGINAL PAPER

A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)

Leonardo Campillos[1] · Louise Deléger[1] · Cyril Grouin[1] · Thierry Hamon[1] · Anne-Laure Ligozat[1] · Aurélie Névéol[1]

Abstract Quality annotated resources are essential for Natural Language Processing. The objective of this work is to present a corpus of clinical narratives in French annotated for linguistic, semantic and structural information, aimed at clinical information extraction. Six annotators contributed to the corpus annotation, using a comprehensive annotation scheme covering 21 entities, 11 attributes and 37 relations. All annotators trained on a small, common portion of the corpus before proceeding independently. An automatic tool was used to produce entity and attribute pre-annotations. About a tenth of the corpus was doubly annotated and annotation differences were resolved in consensus meetings. To ensure annotation consistency throughout the corpus, we devised harmonization tools to automatically identify annotation differences to be addressed to improve the overall corpus quality. The annotation project spanned over 24 months and resulted in a corpus comprising 500 documents (148,476 tokens) annotated with 44,740 entities and 26,478 relations. The average inter-annotator agreement is 0.793 F-measure for entities and 0.789 for relations. The performance of the pre-annotation tool for entities reached 0.814 F-measure when sufficient training data was available. The performance of our entity pre-annotation tool shows the value of the corpus to build and evaluate information extraction methods. In addition, we introduced harmonization methods that further improved the quality of annotations in the corpus.

Electronic supplementary material The online version of this article (doi:10.1007/s10579-017-9382-y) contains supplementary material, which is available to authorized users.

Leonardo Campillos and Louise Deléger have contributed equally to this study.

✉ Aurélie Névéol
aurelie.neveol@limsi.fr

[1] LIMSI-CNRS, Université Paris Saclay, 91403 Orsay, France

Published online: 15 February 2017

Keywords Semantic annotations · Personal health information · Inter-annotator agreement · Clinical narrative

1 Introduction

Corpora with high-quality reference annotations for specific linguistic or semantic phenomena are precious resources for the scientific community. Annotated corpora can be used to develop and evaluate Natural Language Processing (NLP) methods within a defined experimental setting. A number of evaluation campaigns (also called shared tasks or challenges) are regularly carried out for stimulating research in specific areas, thereby providing valuable resources and experimental frameworks. Evaluation campaigns over the past decades have covered research fields such as information retrieval (Text Retrieval Conferences, TREC),[1] semantic annotation (e.g. SemEval tasks),[2] named entity extraction (Message Understanding Conference, MUC),[3] cross-lingual tasks (Cross-Language Evaluation Forum, CLEF)[4] and information extraction in specialized domains (e.g. Informatics for Integrating Biology and the Bedside,[5] from here on i2b2, and Critical Assessment of Information Extraction in Biology, BioCreAtIvE).[6] Resources from evaluation campaigns contribute to validating approches and facilitate replicating experiments by allowing several groups to work with the same data.

Annotated corpora have become available for several genres and subfields in the biomed ical domain. However, very few resources are available for languages other than English. To address this need, we introduce a large high-quality corpus of clinical documents in French, annotated with a comprehensive scheme of entities, attributes and relations: the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). The annotation features good inter-annotator agreement values, which is proof of resource quality.

Herein, we describe the contents of the corpus and the development methodology (pre-annotation, harmonisation, criteria and difficulties found). Section 2 reviews related work and describes our annotation scheme. Sections 3 and 4 explain, respectively, how texts were prepared and selected. Section 5 details the types of annotations and the annotation protocol, and Sect. 6 reports statistics and evaluation metrics. Section 7 discusses the impact of this work.

2 Representation of clinical information contained in text corpora

The availability of clinical corpora is scarce as compared to corpora in the biological domain (Roberts et al. 2009; Cohen and Demner-Fushman 2014). Ethical and privacy issues arise when working with Electronic Health Records (hereafter,

[1] http://trec.nist.gov/.

[2] See Pradhan et al. (2014), Bethard et al. (2015, 2016), Elhadad et al. (2015).

[3] Grishman and Sundheim (1996) report some background on the MUC campaigns.

[4] http://www.clef-campaign.org/.

[5] https://www.i2b2.org/.

[6] http://www.biocreative.org.

EHRs). These require supplementary measures to de-identify patient data—e.g. by removing Personnal Health Identifier or replacing them with surrogates (Grouin and Névéol 2014) before releasing the corpus for research.

Notwithstanding these constraints, annotation efforts have taken hold in the biomedical NLP community, predominantly on English data. Notable research initiatives, in collaboration with health institutions, have annotated clinical texts: the Mayo Clinic corpus (Ogren et al. 2008), the Clinical E-Science Framework (CLEF) (Roberts et al. 2009), the THYME (Temporal Histories of Your Medical Events) project (Styler et al. 2014),[7] the SHARP Template Annotations (Savova et al. 2012), the MiPACQ (Multi-source Integrated Platform for Answering Clinical Questions) (Albright et al. 2013), the IxA-Med-GS (Oronoz et al. 2015) or the Harvey corpus (Savkov et al. 2016). Research challenges have also fuelled the annotation of resources or enrichment of available texts. Well-known corpora come from the i2b2 challenges (Uzuner et al. 2010, 2011; Sun et al. 2013), SemEval (Bethard et al. 2016) and the Shared Annotated Resources (ShARe)/CLEF eHealth labs.[8]

Overall, two levels of annotations have been applied in clinical texts. The first (and more widespread) is a low-level annotation focusing on defining what mentions of clinical and linguistic interest need to be marked in text, and what linguistically and clinically grounded representations to use. The second is a high-level annotation aimed at formally integrating all this information—i.e. linguistic and clinical data—for reasoning over the whole EHR in a computationally actionable way. This is the case of Wu et al. (2013) and Tao et al. (2013), who used a higher-level formal (OWL) clinical EHR representation implemented in cTakes, but relying on a low-level annotation (Savova et al. 2012). The Biological Expression Language (BEL)[9] seems to be a mix between the low and high-level of annotation for life science text (vs. clinical).

Our work has carried out a low-level annotation, but our scheme can be compatible with a high-level representation in the long-run. We checked the aforementioned projects to devise the scheme used in MERLOT, which built on prior work as much as possible while trying to avoid some of the caveats reported and adapt to the nature of our data (Sect. 5.2.1). The final scheme was intended to be suitable for many clinical subfields. In preliminary work (Deléger et al. 2014a), we tested its applicability to clinical notes covering a range of specialities, including foetopathology.

3 Corpus preparation

The original corpus documents were converted from Word to text format using Antiword.[10] A simple rule-based algorithm was used to reconstruct split lines within a paragraph or sentence. The remainder of this section details the processes of de-identification (3.1) and zone detection (3.2).

[7] https://clear.colorado.edu/TemporalWiki/index.php/Main_Page.

[8] http://clefehealth2014.dcu.ie/task-2/2014-dataset.

[9] http://www.openbel.org/.

[10] http://www.winfield.demon.nl/.

3.1 Corpus de-identification and pseudonymization

Due to privacy issues, clinical notes cannot be released in their original form. Protected health information (PHI) (e.g. person names) must be removed (*de-identification*) and replaced with realistic surrogates (*pseudonymization*).

We de-identified clinical notes using a protocol devised by Grouin and Névéol (2014). A set of 100 documents from a corpus of 138,000 documents was pre-annotated with the MEDINA rule-based de-identification system and revised independently by two annotators. Gold standard annotations were obtained through consensus. This gold-standard corpus was used to train a conditional random field (from here on, CRF) model that was in turn used to pre-annotate the 500 documents in our set. Each document was double-checked sequentially by two annotators (three annotators participated in total, A2, A3 and A5, so that each annotator worked with two thirds of the data). PHI elements were then replaced with plausible surrogates.

The annotations for re-introduced PHIs are available for all documents in the corpus, and were used to inform the automatic pre-annotation process.

3.2 Zone detection

We defined a typology of the sections occurring in clinical notes to characterize the contents of documents in our corpus as medical vs. non-medical. We considered four (high-level) section types: (1) a generic header, with contact information for the health care unit in which the note was created (this header is the same for all notes from the same unit); (2) a specific header, with information such as the patient's name, date of birth, admission and discharge dates; (3) the core medical content of a note; and (4) a footer, with the physician's signature and greetings (this latter only if the text is a letter).

Two annotators (A2, A5) manually annotated two samples of 100 randomly-selected notes, by marking the beginning of each section type. Inter-annotator agreement (IAA hereafter) for identifying main content lines had an F-measure of 0.980. Sample 1 was used as a development corpus to design and improve our system. Sample 2 was used as a test set to evaluate the final system. We trained a CRF model to identify the sections and extract the main content of clinical notes. We classified each line of text as belonging to a section, using the BIO (Begin, Inside, Outside) format. Features include the length of a line, the first or second tokens of a line, or the presence of blank lines before a line. This approach draws on previous work on medical section identification from clinical notes (Tepper et al. 2012) and scientific abstracts (Hirohata et al. 2008). More details about the zone detection system are given in Deléger and Névéol (2014), Deléger et al. (2014b).

The resulting model was then applied to the 500 texts in our set. Two annotators revised sequentially the zones segmented in each file. The annotations for manually-validated zones are available for all documents and were shown to annotators in the entity and relation annotation phase (Sect. 5).

4 Corpus design

We used texts from a corpus of 138,000 clinical notes from French healthcare institutions (approximately, 2000 patient EHRs). It covers numerous medical specialities and several text types: discharge summaries, physician letters, medical procedure reports and prescriptions.

4.1 Document selection process

Previous work in corpus linguitics has established good practices for corpus development (Sinclair 2005). A corpus should include *complete documents*, be *representative* (i.e. cover all relevant characteristics of the language) and *balanced* (i.e. all linguistic aspects should be distributed similarly to the natural distribution). The construction of specialized domain corpora, which might exhibit specific properties, is inherently different to that of general language. However, corpus representativity can be achieved by selecting texts that cover the variety of language uses from the relevant domain (Habert et al. 2001).

We restricted to a sample of 500 documents from the Hepato-gastro-enterology and Nutrition ward, to account for the variety of clinical language while keeping the project feasible (Deléger et al. 2014a). We assumed that the corpus is sufficiently homogeneous for its size to train machine learning models.

Accordingly, we considered the four criteria listed below (Sects. 4.1.1–4.1.4) and selected 10 sets of 500 documents through random sampling. We computed the distribution of Semantic Groups among UMLS concepts in each set (see Sect. 4.1.4). We compared these distributions to those of the whole corpus and chose the set with the most similar distribution. Distributions were compared using the Kullback-Leibler value (also called KL-divergence, Kullback and Leibler 1951). The KL-divergence is a measure describing the dissimilarity between two probability distributions and is defined as follows:

$$D(P\| Q) = \sum_{i=1}^{t} p_i \log \frac{p_i}{q_i}$$

P and Q being two probability distributions. The two distributions are identical when the KL-divergence is equal to zero. We thus chose the file set with the smallest KL-value (i.e., with the distribution closest to the whole corpus).

4.1.1 Note type

We selected clinical notes based on the four note types present in the whole corpus (discharge summaries, procedure reports (e.g., radiology reports), physician letters, and prescriptions), keeping the same proportional distribution.

4.1.2 Document length

We divided the notes into three categories based on their length: short notes (word count in the 1st–25th percentile), medium notes (word count in the 26th–75th percentile), and long notes (word count in the 76th–100th percentile). We oversampled medium notes (80%) compared to short (10%) and long (10%) notes. In this way, the majority of notes were close to the average text length.

We did not compute document length based on the whole content of the clinical notes. Clinical notes often include header and footer sections (with, for example, contact information for physicians) that bear little medical interest compared to the main content. In earlier work, we built an automatic tool that identifies zones within clinical notes (see Sect. 3.2). We used this tool to automatically detect the main medical content of clinical notes and computed text length based on the content identified automatically. These zones detected were then manually validated in selected documents.

4.1.3 Gender of patients

We kept the same proportional distribution of male and female patients as in the whole set of notes.

4.1.4 Semantic content

We also checked the semantic content of texts, based on medical concepts from UMLS metathesaurus (Bodenreider 2004). UMLS Concepts are organized in 15 Semantic Groups (SGs) (Bodenreider and McCray 2003). We identified UMLS concepts in the corpus by using a dictionary-based exact-match approach. Then, we looked at the distribution of SGs among those concepts.

5 Corpus annotation

5.1 Annotation tools

We used the BRAT Rapid Annotation Tool (BRAT) developed by Stenetorp et al. (2012).[11] A review of annotation tools Neves and Leser (2012) showed that BRAT was easy to use and could support both our annotation scheme and automatic pre-annotations. Configuration files were set-up to ensure that annotation labels were sorted in the order reflecting their relative frequency, based on a small sample of annotated texts. The most frequent entities (e.g. Anatomy and Procedures) appear at the top of the list while less frequent ones (e.g. medication attributes) are lower in the list and require scrolling for selection. Also, the color scheme for entities was chosen in an attempt to have distinctive colors next to one another and reduce the

[11] This tool is freely available from http://brat.nlplab.org/.

hazard of confusion when annotating entities. The BRAT configuration files are supplied as supplementary material.

We used the open-source companion tool brateval developed by Verspoor et al. (2013) to compute the IAA values (in terms of F-measure) on entity and relation annotations. We extended brateval to compute IAA of attributes.

5.2 Entity and relation annotation

5.2.1 Annotation scheme

The annotation scheme was designed to provide a broad coverage of the clinical domain, in order to allow for the annotation of medical events of interest mentioned in the clinical documents. Semantic annotations in the scheme include entities, attributes, relations between entities, and temporal annotations. We presented in Deléger et al. (2014a) the first version of the schema used to train the annotators.

The annotation scheme for entities comprises 12 elements (Table 1). Our scheme was derived in part from the UMLS semantic groups described in McCray et al. (2001) and Bodenreider and McCray (2003). We included 9 of the 15 UMLS SGs: Anatomy, Chemicals and Drugs, Concepts and Ideas, Devices, Disorders, Genes and Molecular Sequences, Living Beings, Physiology and Procedures. Note that the semantic type (hereafter, STY) Findings was not included in the Disorder class, because prior work has shown this category to yield many false positives (Mork et al. 2010; Névéol et al. 2009). We also created four additional categories for annotating elements of clinical interest:

- SignOrSymptom: Signs/Symptoms and Disorders are separate categories.
- Persons: we created a category for human entities and excluded them from the Living Beings group.
- Hospital: we added an entity type for healthcare institutions.
- Temporal: we created a separate category for temporal expressions and excluded them from the Concept and Ideas group.

We have not restricted the annotation to UMLS entities or specific syntactic classes (e.g. noun or adjective phrases). For example, we have annotated verbs when required, mapping them semantically to the relevant category (e.g. *saigner*, 'to bleed', was annotated as a Disorder entity).

The annotation scheme also defines some attributes (Table 2), which are linked to entities and/or other attributes.

The following are the attributes related to any event entity:

- Aspect: They are anchors of aspect relations to entities (see below).
- Assertion: Textual anchors of assertion relations to entities (see below).
- DocTime: temporal data of events with regard to the moment when the text was created: After, Before, Before_Overlap and Overlap.
- Measurement: Qualitative or quantitative descriptions of entities. This category gathers adverbs (e.g. *progressivement*, 'progressively'), relational and

Table 1 Entities

Entity type	Definition	UMLS semantic type(s)	Examples
Anatomy	Any part or component of the body	Anatomical Structure, Body Location or Region, Body Part Organ or Organ Component, Body Space or Junction, Body Substance, Body System, Cell, Cell Component, Embryonic Structure, Fully Formed Anatomical Structure, Tissue	*Foot; right femoral artery*
Biological process or function	A process or state occurring naturally or as a result of an activity	Biologic Function; Cell Function; Genetic Function; Molecular Function; Natural Phenomenon or Process; Organ or Tissue Function; Organism Function; Physiologic Function	*Transit*
Chemicals_ Drugs	Matter of particular or definite chemical constitution; a substance used as a medication or in the preparation of medication	Antibiotic; Biomedical or Dental Material; Carbohydrates; Chemical; Chemical Viewed Functionally; Chemical Viewed Structurally; Clinical Drug; Hazardous or Poisonous Substance; Inorganic Chemical; Pharmacological Substance; Vitamin	*Insulin; steroids; Percocet*
Concept_Idea	An abstract or generic idea generalized from particular instances	Classification, Conceptual Entity, Functional Concept, Group Attribute, Idea or Concept, Intellectual Product, Language, Qualitative Concept, Quantitative Concept, Regulation or Law, Spatial Concept	*Weight; length*
Devices	An object for diagnosis or treatment	Devices	*Insulin pomp; pacemaker*

Table 1 continued

Entity type	Definition	UMLS semantic type(s)	Examples
Disorder	A condition of the patient or of one of its parts that impairs normal functioning and is manifested by distinguishing signs and symptoms	Acquired Abnormality; Anatomical Abnormality; Cell or Molecular Dysfunction; Congenital Abnormality; Disease or Syndrome; Experimental Model of Disease; Injury or Poisoning; Mental or Behavioural Dysfunction; Pathologic Function; Neoplastic Process	*Diabetes; myocardial infarction*
Genes/Proteins	A gene is defined as the portion of DNA encoding the blueprint for constructing a protein	Amino Acid, Peptide or Protein; Enzyme, Lipid; Immunologic Factor; Indicator, Reagent, or Diagnostic Aid; Gene or Genome; Nucleic Acid, Nucleoside or Nucleotide; Receptor	*PTX1; fibrin*
Hospital	Health care facility, office or ward	_	*Mercy Hospital; ER*
LivingBeings	An individual form of life that is not human	Alga; Amphibian; Animal; Archeon; Bacterium; Bird; Fish; Fungus; Invertebrate; Mammal; Organism; Plant; Reptile; Rickettsia or Chlamydia;Vertebrate; Virus	*Salmonella, HIV*
MedicalProcedure	An activity relating to the practice of medicine or the care of patients	Diagnostic procedures; Health care activity; Laboratory procedure; Therapeutic or preventive procedure	*Angiography, psychiatric consult*
Persons	Human living beings	Human	*Patient; Dr Smith*
Sign/symptom	A manifestation of a condition	Sign or symptom	*Pain; cough*
Temporal expression	Temporal expressions	Temporal concept	*Weekly, 1984*

qualitative adjectives (e.g. *sévère*, 'severe') and quantifiers (*quelques*, 'some'). We also consider measurement units for results of clinical tests.
– Localization: This category expresses spatial details about entities (e.g. *droite*, 'right', or *inférieur*, 'inferior'), which are often mapped to the UMLS Spatial concept type.

Another subset of attributes are specific to some event entities:

– Drug attributes: we consider four types: AdministrationRoute, Dosage, DrugForm and Strength. Temporal attributes (e.g. frequency and duration) are expressed by means of temporal relations (not specific to drug entities). Frequency and dosage data are not split in atomic attributes for measurement units and values.
– Person attributes: we define five types: Donor, HealthProfessional, FamilyMember, Patient and Other. These attributes are only applied to Person entities, but relate to other entities through the Experiences relation.

Our scheme for relations were derived in part from the UMLS Semantic Network and also drew on previous annotation work of clinical texts (e.g. Savova et al. 2012). MERLOT comprises 37 types of relations (Tables 3, 4):

– Aspect relations: they encode a change (or lack of change) with regard to an entity: Continue, Decrease, Improve, Increase, Recurrence_StartAgain, Start, Stop and Worsen (Table 4).
– Assertion relations: there are four types: Negation, Possible, Presence and SubjectToCondition (Table 4). We annotated assertions as relations to make clearer the association between a concept and the type of assertion.
– Drug-attribute relations: four types of links to medication attributes (Table 4): HasAdministrationRoute, HasDosage, HasDrugForm and HasStrengh.
– Temporal relations: there are six types: Before, Begins_on, During, Ends_on, Overlap and Simultaneous (Table 4)
– Event-related relations (Table 3): there are 15 types: Affects, Causes, Complicates, Conducted, Experiences, Interacts_with, Localization_of, Location_of, Measure_of, Performs, Physically_related_to, Prevents, Reveals, Treats and Used_for. Localization_of and Measure_of are links to the attribute entities Localization_of and Measurement_of, respectively.

The temporal scheme for annotation was derived from TimeML (Pustejovsky et al. 2003), but in a slightly different way to previous work (Tapi Nzali et al. 2015) as signals were annotated together with temporal expressions instead of being annotated separately. For instance, the entire expression *il y a 5 ans* (five years ago) was annotated as a time expression of the type duration, while strict TimeML guidelines would require *5 ans* ('5 years') to be annotated as a Duration and *il y a* ('ago') to be annotated as a signal.

Table 2 Attributes

Attribute type	Definition	Involved entities	Involved relation(s)	Examples
Aspect	A phrase that represents a change or an evolution (movements of object are not covered)	All entities	Start	*Started on;*
			Stop	*Interrupted;*
			StartAgain	*Relapse;*
			Continue	*Continued;*
			Increase	*Increase in;*
			Decrease	*Decreased*
			Improve	
			Worsen	
Assertion	A phrase indicating a statement of fact or possibility regarding an entity	All entities and Aspect, Measurement and Localization	Negation	*No;*
			Presence	*Presence of;*
			Possible	*Suspected;*
			Subject to condition	*In case of*
Localization	Precise area where an entity is located (e.g. body side)	All entities	Localization_of	*Left; bilateral*
Measurement	A figure, extent, attribute or amount obtained by measuring or observing, including subjective qualifications. Two subtypes: Quantitative and Qualitative	All entities	Measure_of	*3 cm; normal*
Person type	Person entity type; the predefined options are Patient, PatientFamily, HealthProfessional, Donor and Other	Persons	Experiences	*Dr. Colin*

Table 2 continued

Attribute type	Definition	Involved entities	Involved relation(s)	Examples
DocTime	Temporal data of an annotated event with regard to the moment when the document was authored; the predefined options are Before, After, Overlap and Before_Overlap	Events		*Operation in 1984* [Before]
TemporalType	Type of temporal expression; the predefined options are Date, Time, Duration and Frequency	Temporal	Temporal relations	*Twice a day, 1981*
Medication attributes				
Administration route	Route or method of administering a medication	Chemicals/Drugs	HasAdminis-trationRoute	*Oral; IV*
Dosage	How many of each drug the patient is taking	Chemicals/Drugs	HasDosage	*3 Tablets; two puffs*
DrugForm	Form of a medication	Chemicals/Drugs	HasDrugForm	*Tablet; cream*
Strength	Strength number and unit of a prescribed drug	Chemicals/Drugs	HasStrength	*10 mg; 5 mg/ml*

Table 3 Event-related relations

Relation	Definition	Involved entities
Affects	Produces a direct effect on a process or function	Disorder → BiologicalProcess
		SignOrSymptom → BiologicalProcess
		MedicalProcedure → BiologicalProcess
		Chemicals_Drugs → BiologicalProcess
Causes	Brings about a condition or an effect. Implied here is that an agent, such as a pharmacologic substance or an organism, has brought about the effect. This includes induces, effects, evokes and etiology	LivingBeings → Disorder
		LivingBeings → SignOrSymptom
		Chemicals_Drugs → Disorder
		Chemicals_Drugs → SignOrSymptom
		MedicalProcedure → Disorder
		MedicalProcedure → SignOrSymptom
		Disorder → Disorder
		SignOrSymptom → SignOrSymptom
		SignOrSymptom ↔ Disorder
Complicates	Causes to become more severe or complex or results in adverse effects	Disorder → Disorder
		Chemicals_Drugs → Disorder
		MedicalProcedure → Disorder
		SignOrSymptom → SignOrSymptom
		Chemicals_Drugs → SignOrSymptom
		MedicalProcedure → SignOrSymptom
		SignOrSymptom ↔ Disorder
Conducted	When a test is conducted to investigate a disorder and the outcome is unknown	MedicalProcedure → Disorder
		MedicalProcedure → SignOrSymptom

Table 3 continued

Relation	Definition	Involved entities
Experiences	When a human is affected by an event (e.g. a disorder or a medical procedure).	Persons → Disorder Persons → SignOrSymptom Persons → MedicalProcedure Persons → Chemicals_Drugs Persons → BiologicalProcess Persons → Concept_Idea
Interacts_with	Acts, functions, or operates together with	Chemicals_Drugs → Chemicals_Drugs
Localization_of	The spatial or relative localization of an entity	Localization → Entity
Location_of	The position, site, or region of an entity or the site of a process	Anatomy → Anatomy Anatomy → Disorder Anatomy → SignOrSymptom Anatomy → MedicalProcedure Anatomy → LivingBeings Hospital → MedicalProcedure
Measure_of	The relation between a measurement value and an entity	Measurement → event entity
Performs	A person conducts a procedure	Persons → MedicalProcedure
Physically_ Related_to	Related by virtue of some physical attribute or characteristic	Concept_Idea → Anatomy Concept_Idea → Persons Concept_Idea → Disorder Concept_Idea → SignOrSymptom

Table 3 continued

Relation	Definition	Involved entities
Prevents	Stops, hinders or eliminates an action or condition	Chemicals_Drugs → Disorder
		Chemicals_Drugs → SignOrSymptom
		MedicalProcedure → Disorder
		MedicalProcedure → SignOrSymptom
		Devices → Disorder
		Devices → SignOrSymptom
Reveals	When a test is conducted and the outcome is known or leads to a diagnosis	MedicalProcedure → Disorder
		MedicalProcedure → SignOrSymptom
		SignOrSymptom → Disorder
Treats	Applies a remedy with the object of effecting a cure or managing a condition	Chemicals_Drugs → Disorder
		Chemicals_Drugs → SignOrSymptom
		MedicalProcedure → Disorder
		MedicalProcedure → SignOrSymptom
		Devices → Disorder
		Devices → SignOrSymptom
Used_for	When a device is used (e.g. to conduct a treatment or to administer a drug)	Devices → MedicalProcedure
		Devices → Chemicals_Drugs
		Devices → LivingBeings

Table 4 Aspect, assertion, drug-attribute and temporal relations

Aspect	Definition	Involved entities
Continue	Shows the continuation of an event	Aspect → event entities
Decrease	A lowering value (e.g. of dose)	
Improve	An improvement (e.g. in condition)	
Increase	A rising value (e.g. of dose)	
Recurrence_ StartAgain	Indicates that an event begins occurring again	
Start	Indicates the initiation of an event	
Stop	Indicates the ending of an event	
Worsen	A negative change (e.g. in health)	

Assertion	Definition	Involved entities
Negation	An event is negated	Assertion → event entities
Possible	An event may occur	
Presence	An event occurs	
SubjectToCondi- tion	An event may occur on condition that another event occurs	

Drug-attribute	Types	Involved entities
	HasAdministrationRoute	Chemical_Drugs → drug attributes
	HasDosage	
	HasDrugForm	
	HasStrength	

Temporal	Definition	Involved entities
Before	An event precedes another event/temporal expression	Event entity → Event/ Temporal entity
Begins_on	The event starts on an event or temporal expression	
During	The temporal span of an event is completely contained within the span of another event or temporal expression	
Ends_on	The event finishes on an event or temporal expression	
Overlap	An event happens almost at the same time, but not exactly, as another event/temporal expression	
Simultaneous	An event happens at exactly the same time as another event/ temporal expression	

Lastly, we have flagged ambiguous annotations, abbreviations and acronyms (e.g. *SC* stands for *surface corporelle*, 'body surface'). We have also flagged coreferent pronouns referring to Person entities. An example is shown in Fig. 2 (first sentence), where the entity *votre* (annotated as Persons, PERS) bears the mark *Yes*. Other types of coreference are not annotated. The annotation format makes it possible to remove these flags easily and include or exclude them as a feature according to the training needs of a specific machine learning system.

5.2.2 Annotation process

We first carried out preliminary work to establish the annotation guidelines[12] and annotation method (Deléger et al. 2014a). Then we found that higher IAA values and higher annotation quality could be achieved when the annotation process was carried out in two steps: first perform entity and attribute annotation, then proceed with relation annotation.

To make the staging of annotation work easier, the 500 documents in the corpus were distributed in 100 sets of 5 documents each. Annotators were instructed to work with one set of documents at a time, and to record the annotation time per set. Entities and attributes were annotated before relations.

The annotation work was staged into three phases: a training phase, a consensus phase and an independent phase.

During the training phase, all annotators worked on the same sets of documents (set 0 and 1) to familiarize themselves with the annotation guidelines and discuss any disagreements with other annotators. As a result, 2% of the corpus was annotated by all annotators and consensus annotations were obtained through discussion. The level of training of each annotator was measured through IAA values between each annotator and the consensus annotations. The training was sequential. Annotators worked with set 0, then they could compare their annotations to the gold-standard consensus, before proceeding to set 1.

During the consensus phase, annotators were paired to carry out the double annotation of 19 sets (about two sets per annotator pair). Annotators worked independently in entity and attribute annotations. Then, consensus annotations were obtained jointly by resolving any conflicts. Again, annotators worked independently to add relation annotations. A consensus was finally achieved jointly. We computed the IAA for each of these sets for entities, attributes and relations. In this way, 11% of the corpus was double-annotated.

During the independent phase, the remaining 79 sets were distributed to annotators 2, 4 and 5, who performed the annotation task independently. We did not double-annotate all documents due to time constraints and the fact that we got good IAA values for the 19 double-annotated sets (0.793 for entities, 0.775 for attributes, and 0.789 for relations, exact match). Furthermore, previous work showed that, when inter-annotator agreement values are high, there is no statistically significant difference in the performance of models trained on single-annotated vs. double-annotated training data (Grouin et al. 2014).

5.2.3 Pre-annotation methods

Two types of pre-annotation methods were applied: (1) a lexicon-based approach, used to pre-annotate the first sets of documents; (2) a machine-learning-based approach, used after a sufficient sample of documents was annotated.

[12] The MERLOT annotation guidelines is available at: https://cabernet.limsi.fr/annotation_guide_for_the_merlot_french_clinical_corpus-Sept2016.

Lexicon-based pre-annotation was first used to supply the annotators with entities pre-annotated automatically. This method applied an exact-match strategy based on a French UMLS dictionary and a lexicon derived from small samples of previously manually annotated documents. The pre-annotation process consisted of the following steps: sentence segmentation and tokenization, lemmatization with the French lemmatizer Flemm (Namer 2004), generation of spelling and derivational variants (using the Unified Medical Lexicon for French, UMLF (Zweigenbaum et al. 2005)), application of regular expressions to detect measurements (e.g., *3 cm*) and durations (e.g., *2 weeks*), and matching with the two lexicons. This matching was first applied to the original token and then to the lemma and variants when no match was found. Entities annotated using the lexicon from previous manual annotations had precedence over entities annotated using the larger, UMLS-derived lexicon.

For *machine-learning-based pre-annotation*, we trained CRF models on annotated documents, using Wapiti (Lavergne et al. 2010) with these features:

- **Lexical features:**
  - 1-grams, 2-grams and 3-grams of tokens (−1/+1 window)
  - 1-grams and 2-grams of lemmas
- **Morphological features:**
  - the token is uppercase
  - the token is a digit
  - the token is a punctuation mark
  - 1 to 4-character suffixes of the token
  - 1 to 4-character prefixes of the token
- **Syntactic features:** 1-grams and 2-grams of POS tags of tokens, as provided by the TreeTagger tool (Schmid 1995) (−2/+2 window)
- **Semantic features:**
  - UMLS CUIs of the current token and the previous token
  - 1-grams, 2-grams and 3-grams of UMLS STYs of tokens (−1/+1 window)
  - 1-grams, 2-grams and 3-grams of UMLS SGs of tokens (−1/+1 window)
  - current token was identified as a measurement using regular expressions
  - current token was identified as a duration using regular expressions

Because our annotation scheme includes embedded entities, we built several CRF models, one for each layer of embedment (Alex et al. 2007). Figure 1 shows a sentence with two annotation levels. This required a first CRF layer to capture the Disorder concept *envahissement ganglionnaire* ('ganglionar invasion') and a second layer to capture the embedded Anatomy concept *ganglionnaire* ('ganglionar'). Our

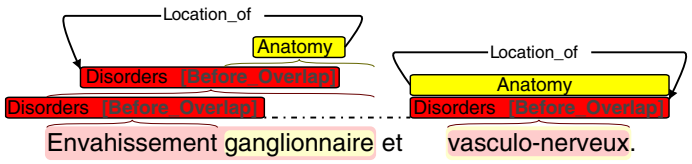

Fig. 1 Sample annotation from the MERLOT corpus

pre-annotation could not match discontinuous entities (e.g. *envahissement vasculo-nerveux*, 'neurovascular invasion').

After using the CRF models to recognize entities (as well as textually anchored attributes), we applied a simple rule-based postprocessing to identify a number of non textually-anchored attributes including Person attributes, Measurement attributes and Temporal type attributes.

5.2.4 Annotation homogenization process

As the annotation process spanned over the course of three years, and because the guidelines went through a few rounds of updates, we performed a final homogenization of annotations. The harmonisation step addressed two points:

- Consistency of annotations over the course of the annotation work: the same entity within a similar context in two documents might have been annotated either with two distinct categories, or annotated only in one document. These inconsistencies depend on the moment the annotation was performed (at the beginning or end of the annotation process), but also on the context meaning, which needed to be checked.
- Consistency of annotation rules: some annotators considered that information between two entities could be inferred without tagging any relation. Inconsistencies in relations especially affected the Aspect and Assertion markers, as annotators interpreted their meanings differently.

We designed scripts to automatically track inconsistencies in entity and attribute annotations across texts and to make the harmonization easier. Two types of inconsistencies were addressed: (1) those involving different annotations for the same text mention (possible annotation error); and (2) inconsistencies where an entity annotated in a document was not marked in another (possible missing annotation). Relation inconsistencies were not addressed. Due to time and human availability constraints, we set up a frequency threshold of 10 mismatches for correcting annotations. That is, we checked and unified (if necessary) entities mismatching their types/attributes up to 10 times.

More efforts were required to fix disagreements on entity types. Mismatches of the same string involved checking each context to understand semantic nuances. Indeed, some entities needed further discussion in the harmonisation stage due to the lack of clear mapping to any UMLS entity. Unifying Assertion and Aspect

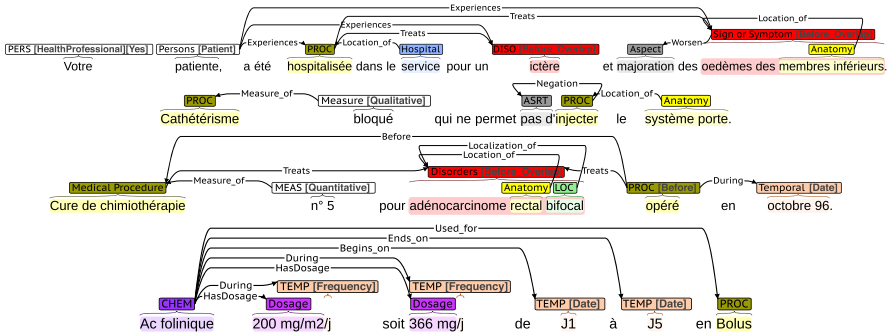

Fig. 2 Sample annotations from the MERLOT corpus

entities took longer in strings where we finally decided to mark two labels (e.g. *redoser*, 'to dose again', MedicalProcedure, was also in the end labelled as Aspect to mark the StartAgain relation). Harmonising attributes was quicker and straightforward. Mismatches were mostly due to missing flags in the annotations, especially of abbreviations (e.g. *hb*, 'hemoglobine'). Attribute annotation mistakes were less frequent and easy to spot and correct.

6 Results: corpus statistics

This section presents the results of the corpus development, which spanned over the course of three years. Figure 2 shows sample annotated excerpts.

6.1 Number of annotations

After harmonising the annotated documents, the annotations amounted to a total of 44,740 entities (including 419 discontinuous entities) and 26,478 relations. The mean (M) number of entities per text was 89.48, and the mean of relations per document was 52.96. Table 5 breaks down the word count[13] and compares the number of annotations before and after the harmonisation process. Figures show that 91 entities and 159 relations were added to the final documents. Both entities and relations increased after the texts were harmonised, due to missing items. Nevertheless, these changes did not require deep and time-consuming changes with regard to the texts produced by annotators. The average IAA value between sets before and after the harmonisation had a 0.988 F-measure with regard to entity annotations. That is, annotations produced by six different annotators were fairly consistent across documents and did not require much effort towards harmonisation.

Figures 3 and 5 depict, respectively, the frequency distribution of annotations of entity and relation types. The most frequent event entity type is MedicalProcedure. This may be partly explained by our annotation criteria, since we annotated verb

[13] Word counts here presented were obtained by means of *wc* Unix commands.

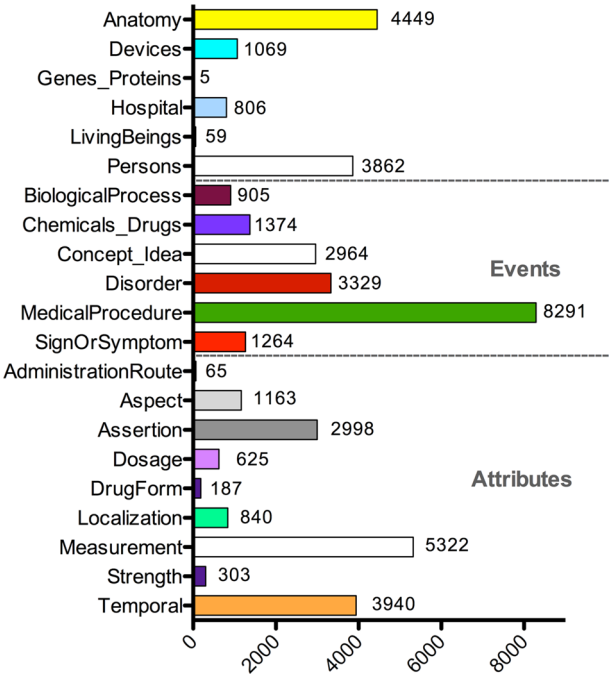

Fig. 3 Frequency of entity types

phrases (e.g. *opérer*, 'to perform a procedure') in addition to noun phrases. Other frequent event entities correspond to Persons and Anatomy. Most medical conditions are Disorder entities instead of Signs or Symptoms. This can be both due to the entity types in our texts and also to annotators' choice of marking Disorder instead of Sign or Symptom events. Genes and Proteins, nevertheless, are infrequent. Regarding attribute entities, Measurement and Temporal entities are widespread, whereas drug-related attributes such as AdministrationRoute and DrugForm occur rarely.

6.2 Inter-annotator agreement (IAA)

In the training sets, IAA values had an average F-measure of 0.681 for the first batch of documents (set 0), but improved to 0.717 in set 1. To assess the soundness of our annotation, a medical doctor annotated set 0 and achieved an F-measure of 0.740 with regard to the consensus annotations of relations.

The average F-measure of the remaining 19 double-annotated sets (i.e. excluding the training sets 0 and 1) was 0.793 for entities, 0.775 for attributes, and 0.789 for relations. These are *good* IAA values—using the term suggested by Altman (1990). We computed our IAA values requiring an exact match between annotations, which is generally lower than a partial match. For example, Albright et al. (2013) achieved an F1 measure of 0.697 in exact match, but of 0.750 in partial match. Overall, our

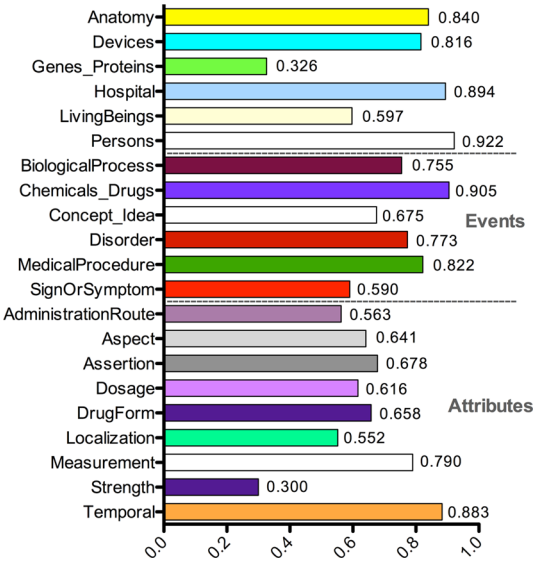

Fig. 4 F-measure per entity type

Table 5 Overall (Total) and average per text (M) number of annotations and word count

	Total	M
Entities		
Before harmonisation	44,649	89.30
After harmonisation	**44,740**	**89.48**
Relations		
Before harmonisation	26,319	52.64
After harmonisation	**26,478**	**52.96**
Words	148,476	296.95

The highest number of annotations is bolded

results are in line with other clinical annotations. Gains in IAA values after a round of consensus have also been reported by Ogren et al. (2008) for English (from 75.7 to 81.4% in entity annotation, exact match) and Oronoz et al. (2015) for Spanish (from 88.63 to 90.53% in term annotation). In a POS annotation task of clinical texts, Savkov et al. (2016) also obtained similar results (0.76% of F-measure). We also obtained higher IAA values in entity annotation than in relation annotation, as other teams have reported (cf. Roberts et al. 2009). We would like to highlight that other work has evaluated annotation quality using annotator-reviser (or adjudicator) agreement, which usually yields higher agreement values. For example, Bada et al. (2012) achieved 90+% annotator-reviser agreement for biomedical concept annotation in the CRAFT corpus. In the THYME corpus, Bethard et al. (2016) reported an interannotator agreement of 0.731 (F1) for temporal expressions, and an annotator-adjudicator agreement of 0.830. Tables 6 and 7 report the figures of the IAA values between pairs of annotators, computed as the average F-measure of both sets that were double-annotated.

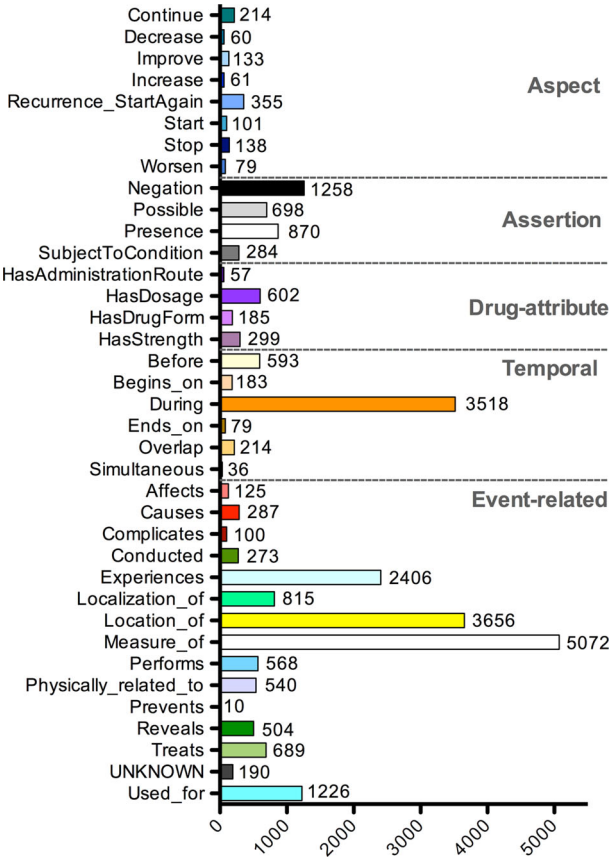

Fig. 5 Frequency of relation types

Figures 4 and 6 break down the average F-measure values corresponding to the IAA of each type of entity and relation, respectively. With regard to event entities, the higher IAA values correspond to Chemical and Drugs, Hospital, Persons, Medical procedures and Devices. Signs or Symptoms have lower IAA values than Disorders—probably due to the fact that annotators had difficulties in distinguishing them. Genes and Proteins and Living Beings had the lower values. Attribute entities with the higher IAA values are Temporal and Measurement. Strength has a poor IAA value, which accounts for the fact that several annotators could not discriminate it clearly from Dosage.

As for the relations annotated, the higher IAA values are those involving drug-attribute entities—i.e. HasAdministrationRoute, HasDosage and HasStrength. Likewise, Negation, Measure_of and Performs relations have high values. Relations such as Affects, Causes, Conducted_for and Complicates have low values. This is probably due to the annotators' lack of medical knowledge to ascertain the cause-

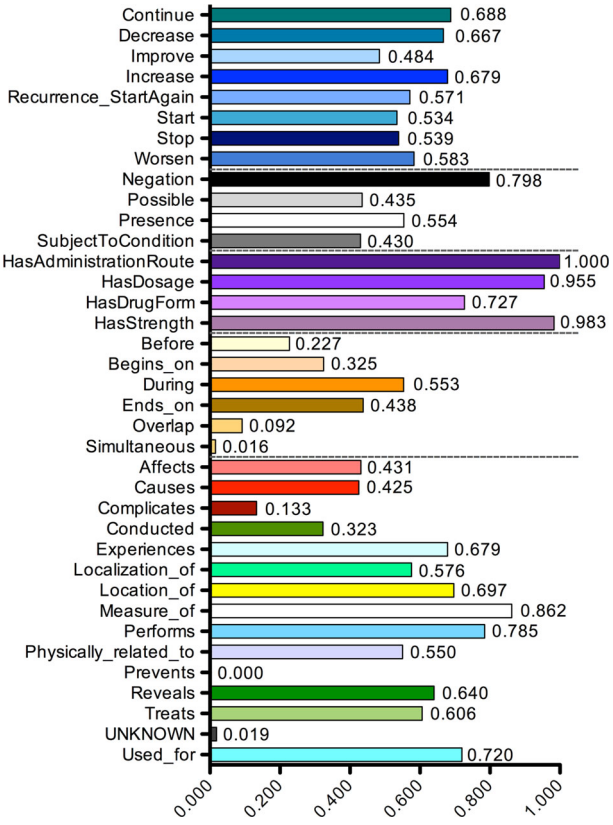

Fig. 6 F-measure per relation type

Table 6 Inter-annotator agreement for entities and relations

	Entities					Relations				
	A2	A3	A4	A5	A6	A2	A3	A4	A5	A6
A1	0.817	0.779		0.810	0.794	0.866	*0.724*		**0.868**	0.792
A2		*0.750*	**0.844**	0.794	0.771		0.762	0.782	0.806	0.775
A3				0.800	0.801				0.756	0.748
A4					0.787					0.797

The lower values are shown in italics

The higher values are shown in bold

affect relationship between entities. Having health professionals annotating the same texts would be an interesting replication experiment to compare results. Lastly, temporal relations (e.g. Overlap, Before, Simultaneous) show the lower IAA values. The lack of context information to understand the timeline of patient's events might explain these poor values.

Table 7 Inter-annotator agreement for attributes

	Attributes				
	A2	A3	A4	A5	A6
A1	**0.844**	0.751		0.819	0.757
A2		0.746	0.829	0.753	*0.736*
A3				0.804	0.755
A4					0.764

The lower values are shown in italics

The higher values are shown in bold

Table 8 Mean and average F-measure per number of training documents in batch

Preannotation	Sets in batch	Training docs. in batch	Mean F-measure	SD
Lexicon	1	15	0.483	–
Lexicon	11	20	0.546	0.042
CFR	10	55	0.718	0.062
CFR	10	95	0.774	0.074
CFR	68	130	0.814	0.045

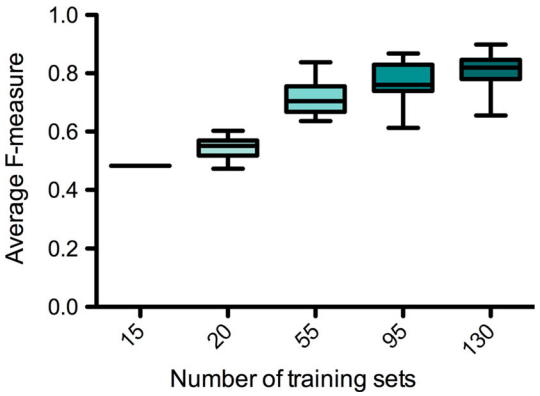

Fig. 7 Performance of pre-annotation (F-measure per number of training documents)

6.3 Performance of automatic pre-annotation

The average performance of the automatic pre-annotation in terms of F-measure was 0.768, a figure close to our IAA values (0.793 for entities).

A lexicon-based approach was used to preannotate the first batch (1 set of documents preannotated after training on 15 documents) and the second batch (11 sets of documents preannotated after training on 20 documents). The average F-measure values of this method were low: respectively, 0.483 and 0.546 (Table 8). The following sets were preannotated using the CRF models trained on 55, 95 and 130 documents. With the machine-learning-based preannotation, the F-measures increased steadily: respectively, 0.718, 0.774 and 0.814 (Fig. 7). Figure 8 shows the

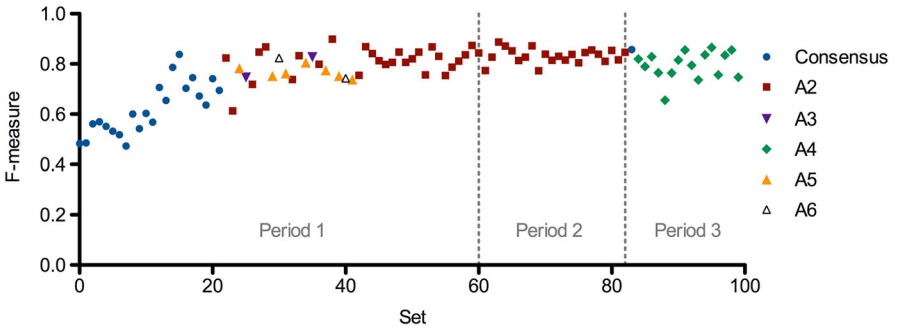

Fig. 8 Performance of the pre-annotation in terms of F-measure over the entire corpus

F-measure of the preannotation of entities for each set and the time-line of the annotation task. Period 1 corresponds to the time when all annotators (except annotator 4) worked on the multiple- and double-annotated sets (2014); period 2, to the interval when annotator 2 pursued the annotations (from 2014 through 2015); and period 3 corresponds to annotator 4, who took over of the final stages (2016). Circles represent consensus annotations (sets 0 to 21, and set 83); and triangles and squares, single annotations. When comparing F-measures across sets, we can observe that the performance of the preannotation increases with the number of training documents, but until a certain amount of data, where the F-measure values reaches a plateau.

A one-way ANOVA showed that the difference between the four types of batches (i.e. respectively having used 20, 55, 95 and 130 training documents) was statistically significant: $F(3,96) = 97.25$, $p < 0.0001$ (***). The effect size was nonetheless very large (eta squared = 0.75). Note that we did not consider the first set trained on 15 documents in this ANOVA test, due to the scarce data.

6.4 Annotation time

6.4.1 Training stage

Figure 9 presents the annotation time in minutes each human annotator spent to annotate the first two sets of five documents (set 0 and set 1) in entities (left) and relations (right). Those two sets were annotated during the training stage by each human annotator (A2–A6), followed by a consensus stage (C). We only report the annotation times of five annotators, due to the availability of the data. Note that annotation times were longer for annotator 4, who took hold of the training annotation task after the guidelines were fixed.

Annotation time for entities and attributes in the training sets range from 90 to 300 minutes in set 0, and from 120 to 180 minutes in set 1. The maximum time was spent during the consensus stage, which involved several annotators. Consensus took much more time for set 0 than for set 1. This observation corresponds to a progression in the training process, as the number of inconsistencies and decisions

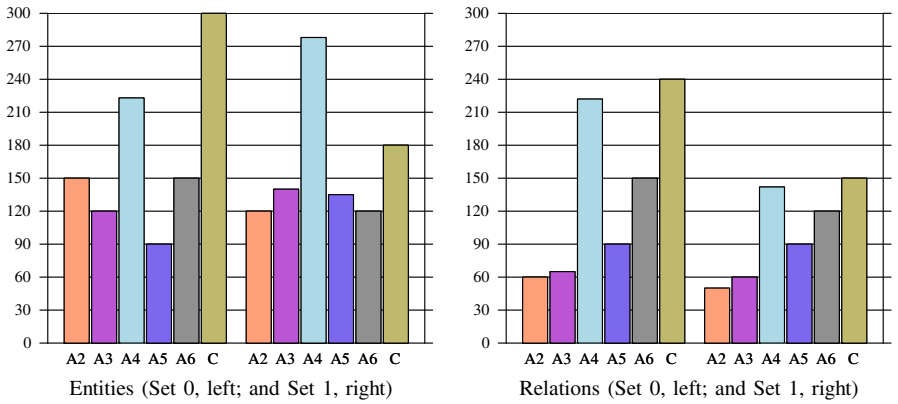

Fig. 9 Annotation time in minutes for entities (*left*) and relations (*right*) on set 0 and 1 for single annotations (A2 = annotator 2, ..., A6 = annotator 6) and consensus (C)

Table 9 Average annotation times per set (in minutes) corresponding to each annotator

	A1	A2	A3	A4	A5	A6
Entities	62.50	55.15	55.00	156.59	76.25	91.30
Relations	36.88	49.33	33.00	87.24	38.00	73.70

to make decreased as guidelines were assimilated. The average annotation times in both sets was 167.17 for entities, and 119.92 for relations. Annotation time for relations was lower than for entities, and annotators' times in set 0 were close to those in set 1. Again, the consensus time decreased when annotating relations in set 1. As mentioned, a medical doctor also worked on the first batch of documents (set 0). His annotation times were in a similar range to other annotators (75' for entities, 150' for relations).

6.4.2 Production stage (double and independent annotations)

The mean annotation time (per set of five documents) in the production stage was 82.73 for entities and 53.02 for relations. As expected, annotators spent less time in the production than in the training stage. However, differences across annotators appeared (Table 9), especially regarding annotator 4.

Figures 10 and 11 represent the annotation time in minutes each human annotator spent to annotate each set of five documents during the production stage. Sets are presented in the order each annotator processed them, from the first two from the training stage to the more recent ones. In Fig. 10 (representing double-annotated sets: from set 2 to set 21, and also set 83), dark bars show entities, and light bars indicate relations. In Fig. 11 (all sets), full-coloured symbols represent the

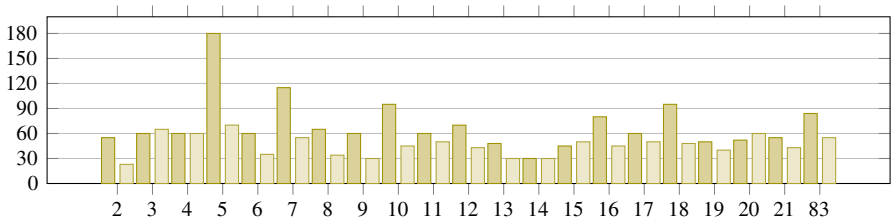

Fig. 10 Annotation time in minutes for entities (*dark bars*) and relations (*light bars*) for each set of double-annotated documents (set numbers are placed on the x axis)

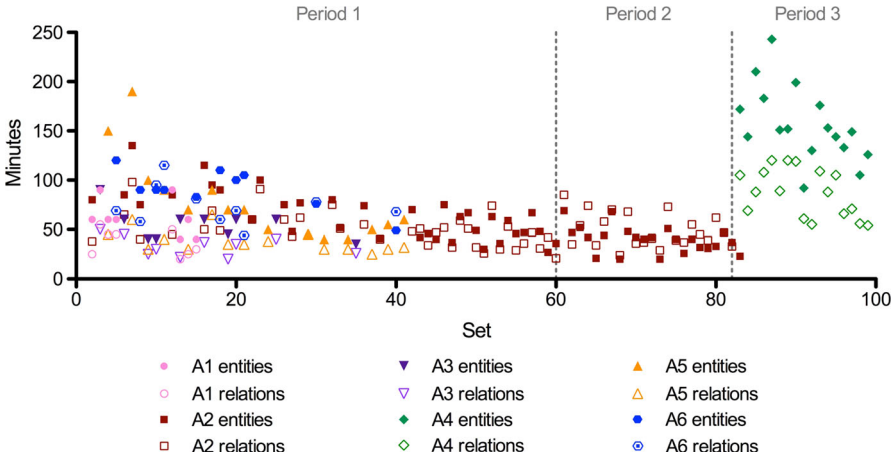

Fig. 11 Annotation times of entities and relations (in minutes) per set of five documents

annotation times of entities, and empty symbols, those of relations. The number of sets per annotator differs; annotators 2 and 4 carried out most of the annotation task.

These histograms show that, overall, more time was needed to annotate entities than relations. Exceptions were some difficult sets with relations with semantically difficult nuances or where domain knowledge was needed. The graphs suggest different annotator profiles. A first group (A1, A4 and A6) spent consistently as much time during the training stage as in the consensus stage. These annotators might have been careful and looked up the guidelines and supporting resources consistently throughout the annotation. A second group (A2, A3 and A5) spent more time during the first stages but annotated the other sets more rapidly. Those annotators might have taken time to get acquainted with the guidelines before feeling comfortable with the task.

Concerning the consensus stage (double-annotation), a lot of time was needed for setting up the annotation guidelines in the first two sets. For the remaining sets, however, consensus took annotators less time than single annotation did. Exceptions are set 5 (consensus of 180 minutes) and set 7 (consensus made in 115 minutes). Set 7 was annotated by annotators 2 and 5, who designed the annotation guidelines. As

this set is the first they processed out of the training stage, another discussion took place to enrich the guidelines based on this new annotated set.

A final remark is to be made on annotator 4, whose times were longer both for entities and relations. Two reasons might explain this. First, this annotator was not a native-speaker of French. Second, they worked after the annotation guidelines were fixed, without the option to contribute to the guidelines according to their annotation experience, as was the case for the other annotators.

7 Concluding remarks

We have presented the development of a large French clinical corpus annotated with a complex scheme of entities, attributes and relations. To our knowledge, this is the first clinical corpus in a language other than English to provide clinical annotations of this scale and complexity, and featuring good IAA values. In future work, we plan to exploit the annotations to develop and evaluate methods for the automatic extraction of entities, attributes and relations from French clinical text. The corpus may also be used for building clinical information extraction systems or clinical decision support systems by leveraging clinical knowledge encoded in the text of EHRs with entities and relations.

The patient records were obtained through a use agreement with a French hospital whereby data would be restricted to research carried out by the partners entering into this agreement. As a result, the corpus cannot be distributed freely. However, the annotation scheme, guidelines and harmonization tools are available to the community.[14] The texts are, moreover, all related to the Hepatogastroenterology and Nutrition specialities. While this ensures coherence within the corpus, it could limit the applicability of models trained on the corpus to other medical areas.

We would like to highlight that this work has yielded notable results together with the corpus construction. A comprehensive annotation scheme has been designed, applied and fine-tuned to encode entities, attributes and relations in clinical narrative. Automatic techniques to identify sections in clinical notes and preannotate entities have been set up with demonstrated efficiency. Lastly, we have designed a work methodology involving training, consensus and independent annotation stages with a final harmonisation stage. These procedures ensure high-quality annotations, as our IAA values show, and are potentially extensible to other languages and domains.

Acknowledgements This work was supported by the French National Agency for Research under grant CABeRneT (Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle) ANR-13-JS02-0009-01. The authors thank the Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work. The authors specially wish to acknowledge Dr. Griffon for his critical appraisal of the annotation scheme and its application to the corpus.

[14] The BRAT configuration files for the annotation are available at https://cabernet.limsi.fr/MERLOTscheme.zip. The annotation guidelines are available at https://cabernet.limsi.fr/annotation_guide_for_the_merlot_french_clinical_corpus-Sept2016.

References

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W., Warner, C., Hwang, J., et al. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *JAMIA*, *20*(5), 922–930.

Alex, B., Haddow, B., & Grover, C. (2007). Recognising nested named entities in biomedical text. In *Proceedings of BioNLP 2007 workshop* (pp. 65–72). ACL.

Altman, D. G. (1990). *Practical statistics for medical research*. Boca Raton: CRC Press.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., et al. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, *13*(1), 1.

Bethard, S., Derczynski, L., Savova, G., Savova, G., Pustejovsky, J., & Verhagen, M. (2015). SemEval-2015 task 6: Clinical tempeval. In *Proceedings of the international workshop on semantic evaluation (SemEval)* (pp. 806–814). ACL.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2016). SemEval-2016 task 12: Clinical tempeval. In *Proceedings of SemEval Workshop* (pp. 1052–1062). ACL.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270.

Bodenreider, O., & McCray, A. T. (2003). Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, *36*(6), 414–432.

Cohen, K., & Demner-Fushman, D. (2014). *Biomedical natural language processing*. Natural Language Processing Series. John Benjamins.

Deléger, L., Grouin, C., Ligozat, A.-L., Zweigenbaum, P., & Névéol, A. (2014a). Annotation of specialized corpora using a comprehensive entity and relation scheme. In *Proceedings of LREC 2014* Reikjavik, Iceland.

Deléger, L., Grouin, C., & Névéol, A. (2014b). Automatic content extraction for designing a french clinical corpus. In *Proceedings of AMIA annual symposium* Washington, DC: American Medical Informatics Association (AMIA).

Deléger, L., & Névéol, A. (2014). Identification automatique de zones dans des documents pour la constitution d'un corpus médical en français. In *Traitement Automatique de la Langue Naturelle–TALN* (pp. 568–573).

Elhadad, N., Pradhan, S., Chapman, W., Manandhar, S., & Savova, G. (2015). SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of SemEval workshop.* (pp. 303–10). ACL.

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A brief history. In *Proceedings of the 16th COLING conference, vol. 1* (pp. 466–471). Stroudsburg, PA: Association for Computational Linguistics.

Grouin, C., Lavergne, T., & Névéol, A. (2014). Optimizing annotation efforts to build reliable annotated corpora for training statistical models. In *LAW VIII–the 8th linguistic annotation workshop*, 2014 (pp. 54–58).

Grouin, C., & Névéol, A. (2014). De-identification of clinical notes in French: Towards a protocol for reference corpus development. *Journal of Biomedical Informatics*, *50*, 151–161.

Habert, B., Grabar, N., Jacquemart, P., & Zweigenbaum, P. (2001). Building a text corpus for representing the variety of medical language. In *Proceedings Corpus Linguistics, Lancaster.*

Hirohata, K., Okazaki, N., Ananiadou, S., & Ishizuka, M. (2008).Identifying sections in scientific abstracts using conditionalrandom fields. In *Proceedings of the IJCNLP 2008.*

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 49–86.

Lavergne, T., Cappé, O., & Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL)* (pp. 504–513). Association for Computational Linguistics.

McCray, A. T., Burgun, A., & Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In *Proceedings of MedInfo*, (vol. 10, pp. 216–20).

Mork, J., Bodenreider, O., Demner-Fushman, D., Islamaj Doğan, R., Lang, F., Lu, Z., et al. (2010). Extracting Rx information from clinical narrative. *JAMIA*, *17*(5), 536–9.

Namer, F. (2004). Flemm: un analyseur flexionnel de français à base de règles. *Traitement Automatique des langues naturelles (TALN)*, *41*, 523–47.

Neves, M., & Leser, U. (2012). A survey on annotation tools for the biomedical literature. *Brief Bioinformatics*, (pp. 523–47).

Névéol, A., Kim, W., Wilbur, W., & Lu, Z. (2009). Exploring two biomedical text genres for disease recognition. In *Proceedings of the NAACL HLT 2009 BioNLP workshop* (pp. 144–52). ACL.

Ogren, P., Savova, G., & Chute, C. (2008). Constructing evaluation corpora for automated clinical named entity recognition. In *Proceedings of LREC 2008* Marrakech, Morocco.

Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D., & Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, *56*, 318–332.

Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., & Savova, G. (2014). SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of SemEval workshop* Vol. 199, no. 99 (p 54).

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., et al. (2003). TimeML: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, *3*, 28–34.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., et al. (2009). Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Semantics*, *42*, 950–966.

Savkov, A., Carroll, J., Koeling, R., & Cassell, J. (2016). Annotating patient clinical records with syntactic chunks and named entities: The Harvey corpus. *Language Resources and Evaluation*, (pp. 1–26).

Savova, G., Styler, W., Albright, D., Palmer, M., Harris, D., & Zaramba, G., et al. (2012). *SHARP template annotations: Guidelines*. Technical report, Mayo Clinic.

Schmid, H. (1995). Treetaggerl a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, *43*, 28.

Sinclair, J. (2005). Corpus and text–basic principles. *Developing Linguistic Corpora: A Guide to Good Practice*, (pp. 1–16).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). BRAT: A Web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012* (pp. 102–7).

Styler, W, I. V., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P., et al. (2014). Temporal annotation in the clinical domain. *Transactions of the ACL*, *2*, 143–154.

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, *20*(5), 806–813.

Tao, C., Jiang, G., Oniki, T., Freimuth, R., Zhu, Q., Sharma, D., et al. (2013). A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *JAMIA*, *20*(3), 554–62.

Tapi Nzali, M.D., Névéol, A., & Tannier, X. (2015). Analyse d'expressions temporelles dans les dossiers électroniques patients. In *Actes de TALN 2015* (pp. 144–52).

Tepper, M., Capurro, D., Xia, F., Vanderwende, L., & Yetisgen-Yildiz, M. (2012). Statistical section segmentation in free-text clinical records. In *Proceedings of LREC 2012* Istanbul, Turkey.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *JAMIA*, *17*(5), 514–518.

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA*, *18*(5), 552–556.

Verspoor, K., Yepes, A.J., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., & Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database*.

Wu, S., Kaggal, V., Dligach, D., Masanz, J., Chen, P., Becker, L., et al. (2013). A common type system for clinical natural language processing. *Journal of Biomedical Semantics*, *4*(1), 1.

Zweigenbaum, P., Baud, R. H., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., et al. (2005). A unified medical lexicon for French. *International Journal of Medical Informatics*, *74*(2–4), 119–124.