The *Quaero* French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization

Aurélie Névéol[1] **Cyril Grouin**[1] **Jeremy Leixa**[2]
Sophie Rosset[1] **Pierre Zweigenbaum**[1]
[1]CNRS, UPR 3251, LIMSI [2]ELDA
91403 Orsay, France 75013 Paris, France
{firstname.lastname}@limsi.fr leixa@elda.org

Abstract

A vast amount of information in the biomedical domain is available as natural language free text. An increasing number of documents in the field are written in languages other than English. Therefore, it is essential to develop resources, methods and tools that address Natural Language Processing in the variety of languages used by the biomedical community. In this paper, we report on the development of an extensive corpus of biomedical documents in French annotated at the entity and concept level. Three text genres are covered, comprising a total of 103,056 words. Ten entity categories corresponding to UMLS Semantic Groups were annotated, using automatic pre-annotations validated by trained human annotators. The pre-annotation method was found helful for entities and achieved above 0.83 precision for all text genres. Overall, a total of 26,409 entity annotations were mapped to 5,797 unique UMLS concepts.

Keywords: Corpus annotation; Named Entity Recognition; Entity Normalization

1. Introduction

A vast amount of information in the biomedical domain is available as natural language free text. Over the past decades, substantial research efforts have addressed the development of methods and tools to automatically process biomedical free-text and provide organized and structured representations of domain knowledge contained in the literature and in clinical documents. Much of this work relied on manually annotated corpora of biomedical texts written in English, such as GENETAG (Tanabe et al., 2005), GENIA (Kim et al., 2003) or corpora available through shared tasks such as i2b2 (Savova et al., 2011), BioCreative (Arighi et al., 2011) or BioNLP (Bossy et al., 2013). However, an increasing number of documents in the field are written in languages other than English. Therefore, it is essential to develop resources, methods and tools that address Natural Language Processing in the variety of languages used by the biomedical community.

While it is generally agreed that domain knowledge is expressed in text through *mentions* or *named entities* that may refer to domain concepts captured in domain terminologies or ontologies, the numerous annotated resources available for the biomedical domain show that there is no consensus on the representation and annotation of entities: what are the entities of interest? Should they be defined through meaning, syntax, both? What level of granularity should be taken into account? In this work, we chose to annotate a range of high-level semantic categories, in line with recent work addressing named entity recognition (Rebholz-Schuhmann et al., 2013) and corpus development (Albright et al., 2013) in the biomedical domain.

We produced corpora and annotation guidelines for named entities which are both hierarchical and compositional[1] (Grouin et al., 2011), and which we used in contrastive studies of news texts in French (Rosset et al., 2012). In this definition, in addition to the hierarchy and compositionality, the inclusion of entities covers antonomasia and metonymy but also takes into account cases of entities that are built upon other entities. This inclusion process seemed particularly interesting. Specifically, in this definition, complex entities can be decomposed into several simpler entities that may be of different categories.

[1]Corpora, guidelines and tools are available through ELRA under references ELRA-S0349 and ELRA-W0073.

2. Material and Methods

2.1. Corpus selection

A selection of text comprising relevant biomedical entities was made using three different types of documents: information on marketed drugs from the European Medicines Agency (EMEA),[2] titles of research articles indexed in the MEDLINE database,[3] and patents registered with the European Patent Office (EPO).[4] All three sources were recently used in an international challenge for cross-lingual medical named entity recognition, CLEF-ER (Rebholz-Schuhmann et al., 2013). We decided to use all three corpora because they cover three different genres of biomedical documents, and the availability of the documents in at least one language other than French (English for MEDLINE, English and German for EPO, several European languages for EMEA) provides an opportunity for a richer corpus.

Our goal in selecting documents was to obtain a sample with a few thousand annotations per semantic category in order to have a good representation of how these types of concepts were referred to in biomedical text. Based on manual annotation of a small sample of each text type, we randomly selected 2,500 MEDLINE titles, 13 EMEA documents and 25 EPO patents. To ensure the relevance of the selected patents to the biomedical domain, patents were selected among those containing at least one of the words "*maladie*" (disease) and "*médicament*" (drug).

[2]http://opus.lingfil.uu.se/EMEA.php
[3]http://www.ncbi.nlm.nih.gov/pubmed/
[4]http://www.epo.org/

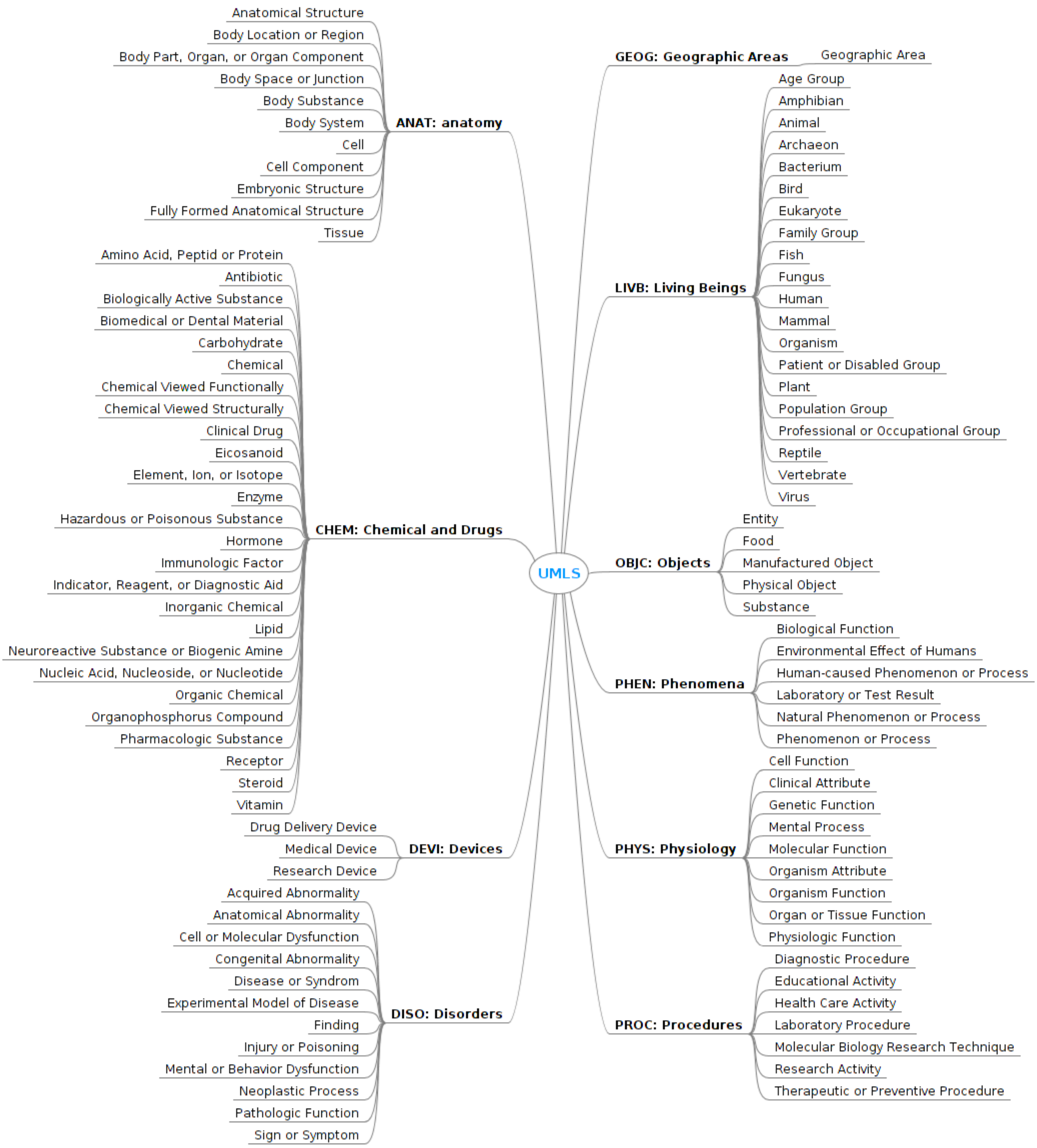

Figure 1: Annotated Entities

2.2. Annotation Guidelines

2.2.1. Entities are defined using the UMLS

The entities are defined based on the UMLS® (Unified Medical Language System®). The Metathesaurus unifies concepts from several dozen terminologies in the biomedical domain, including linked terms and relations. The Semantic Network comprises Semantic Types and Semantic Relations, which are organized hierarchically. The 134 Semantic Types can be clustered into 15 Semantic Groups (McCray et al., 2001). Each concept in the UMLS is assigned a Concept Unique Identifier (CUI), a set of terms and one or more Semantic Types. Semantic Groups were designed so that each concept could be assigned to only one semantic category (Bodenreider and McCray, 2003). Figure 1 shows the subset of entities defined according to the UMLS Semantic Groups and Semantic Types used in this work. The annotation task was carried out at two granularity levels: Semantic Groups and Concepts. The annotators used freely available tools for accessing the UMLS in

French[5] and in English[6] in order to assess whether a mention in text referred to a specific Semantic Group and to identify the specific concept.

2.2.2. Annotations are comprehensive

The goal of the annotation task was to provide a resource that would be as complete and comprehensive as possible. The guidelines applied the principles of *Quaero* annotations in the general domain, which imply that a complex entity can be decomposed into simpler entities that may belong to different categories. The main guidelines were as follows:

- If a mention can refer to more than one Semantic Group, all the relevant Semantic Groups should be annotated. For instance, the mention "*récidive*" (recurrence) in the phrase "*prévention des récidives*" (recurrence prevention) should be annotated with the category "DISORDER" (CUI C2825055) and the category "PHENOMENON" (CUI C0034897);

- If a mention can refer to more than one UMLS concept within the same Semantic Group, all the relevant concepts should be annotated. For instance, the mention "*maniaques*" (obsessive) in the phrase "*patients maniaques*" (obsessive patients) should be annotated with CUIs C0564408 and C0338831 (category "DISORDER");

- Entities which span overlaps with that of another entity should still be annotated. For instance, in the phrase "*infarctus du myocarde*" (myocardial infarction), the mention "*myocarde*" (myocardium) should be annotated with category "ANATOMY" (CUI C0027061) and the mention "*infarctus du myocarde*" should be annotated with category "DISORDER" (CUI C0027051);

- Discontinuous entities should be annotated separately. For instance, in the phrase "*maladies rares et chroniques*" (rare and chronic diseases) the entity "*maladies rares*" (rare diseases) should be annotated with the category "DISORDER" (CUI C0678236) and the entity "*maladies chroniques*" (chronic diseases) should be annotated with the category "DISORDER" (CUI C0008679).

2.3. Annotation Development

This section describes the annotation process performed by the human annotators. Two people were involved in the task over the course of 4 weeks: the project manager (JL) and one annotator specifically recruited for the project. Both annotators are native French speakers with a good command of English. When working out the workload distribution, we found that an expert annotator and a beginner could reasonably manage to annotate the medical corpus.
Before working on the medical corpus, the beginner annotator was trained for the annotation process on files taken from another (general) corpus. As part of the training, this annotator started working on the general corpus to practice applying the general annotation principles. Then, switching to the pre-annotated medical corpus was seamless because the annotation guidelines were very similar. The main difference was the use of the UMLS and PTS websites, and their role in the annotation process:

- For each Named Entity in the text (pre-annotated or not), search the corresponding concept in the UMLS database;

- If a suitable concept is found in the UMLS, check whether its category is listed in the annotation manual;

- If the Named Entity's category is of interest, report the associated CUI;

- Create a complete annotation for the Entity.

For the annotation task, the three sub-corpora (EMEA, MEDLINE and EPO) were divided equally between the annotators. In order to get used to the annotation process, the human annotators decided to begin with a few files from the MEDLINE sub-corpus (10 per annotator, out of 2,500), before working on the whole EMEA sub-corpus.
The tools used for the annotation task were Xemacs and dedicated configuration files enabling the processing of embedded annotations files from the medical corpus. As mentioned earlier, the medical corpus was supplied to the human annotators with a set of pre-annotations automatically obtained based on the method described in Bodnari et al. (2013). Table 1 shows a sample file with the original text content, the pre-annotated content supplied to the human annotators with embedded annotations and the annotations they finally produced. Note that due to time constraints, the documents were supplied to the human annotators without prior tokenization. This had an impact on the annotation task: because the Xemacs tool is only able to create annotations at the token level, in many cases, punctuations were included in the annotations done by the annotators, as seen in the "*cancer intestinaux.*" (intestinal cancer) example in Table 1, where the final full stop is part of the annotation.
A small portion (about 5%) of the corpus was annotated independently by both annotators, in order to calculate Inter Annotator Agreement (IAA). The first files from the MEDLINE sub-corpus were annotated in collaboration between the annotators, in order to discuss any annotation issues early on. The remaining files were annotated independently, but annotators still met once a week to discuss their issues and share comments on the annotation experience.

2.4. Corpus quality assurance and formatting

In order to ensure high-quality annotations, several quality assurance steps were included in the annotation process.
About 5% of the corpus was selected from the EPO and EMEA sub-corpora to be annotated by both the human annotators, and Inter Annotator Agreement (IAA) was computed (in terms of Kappa and F-measure) at the entity level for this sample. Additionally, two samples of 100 MEDLINE titles were selected randomly and an annotator from

[5] The Portail Terminologique de Santé (PTS) developed by the Rouen city hospital (`http://pts.chu-rouen.fr`)

[6] The National Library of Medicine Metathesaurus Browser (`https://uts.nlm.nih.gov/metathesaurus.html`)

Plain document	Facteurs de croissance et cancers intestinaux.
English translation	Growth factors and intestinal cancers.
Pre-annotated document	Facteurs de croissance et <DISO CUI="C0346627"> cancers intestinaux. </DISO>
Annotated document	<CHEM CUI="C0018284"> Facteurs de <PHYS CUI="C18270"> croissance </PHYS> </CHEM> et <DISO CUI="C0346627"> <DISO CUI="C0027651"> cancers </DISO> <ANAT CUI="C0021853"> intestinaux. </ANAT> </DISO>

Table 1: A sample MEDLINE document (PMID 1421706) with the automatic pre-annotations supplied to the human annotators, and the annotations they produced

the curator team who contributed to writing the guidelines (AN) revised the annotations supplied by the human annotators for these samples. IAA was computed between the human annotations version and curator-revised version of the samples.

A list including the most frequently encountered CUIs (such as most common body parts and diseases, age groups) was compiled so that both annotators would tag the most common Named Entities the same way.

At the end of the annotation process, the most common entities (such as "*syndrome*" (syndrome) and "*maladie*" (disease)) were pulled and the consensus category and CUI for each entity was enforced using Search and Replace method, when no ambiguity was possible. The consistency of CUI-category assignment for each entity was also checked automatically based on UMLS data: cases where, according to the UMLS, the CUI assigned to an entity did not belong to a Semantic Type corresponding to the assigned category (Semantic Group) were automatically flagged for correction. For example, when processing the entity "*<GEOG CUI="C0331677"> Paris </GEOG>*", the annotation of "*Paris*" with CUI C0331677 was flagged because CUI C0331677 (Genus Paris) is associated with the semantic group LIVB in the UMLS. This annotation could then be revised to the correct CUI C0030561 because the text referred to the city of Paris. The excerpt shown in Table 1, shows a case where the CUI C18270 was flagged because of a typo error. This annotation could then be revised to the correct CUI C0018270.

To improve the quality and accessibility of the annotations, the annotated corpus was post-processed for tokenization.

3. Results

3.1. Corpus statistics

Table 2 presents general corpus statistics including the number of tokens, the number of annotated entities, and the number of unique CUIs in the corpus.

Table 3 presents our evaluation of the performance of entity pre-annoation. We computed inter-annotator agreement scores between the pre-annotated (automatic) and final (revised by humans) versions of each corpus. We computed the Kappa coefficient (lower bound) and F-measure (higher bound), as defined in Grouin et al. (2011).

	EMEA	**MEDLINE**	**EPO**	**All**
Tokens	58,874	23,647	20,537	103,057
Pre-annotations				
Entities (all)	7,280	1,692	1,662	10,634
Entities (unique)	1,672	1,194	305	3,009
CUIs (all)	12,098	3,207	2,211	17,516
CUIs (unique)	1,653	1,879	378	3,325
Final corpus				
Entities (all)	12,761	8,781	4,865	26,407
Entities (unique)	2,839	5,600	960	8,460
CUIs (all)	12,647	8,767	4,867	26,281
CUIs (unique)	1,807	4,156	759	5,796

Table 2: Overview of the *Quaero* Medical corpus

Metric	EMEA	MEDLINE	EPO
Kappa	0.361	0.142	0.187
F-measure	0.595	0.303	0.428
Precision	0.831	0.937	0.841
Recall	0.463	0.181	0.287
Correct #	5,906	1,585	1,398
Insert #	6,261	7,123	3,394
Delete #	610	38	192
Substitution #	588	69	72

Table 3: Inter-annotator agreement scores and annotation comparisons between pre-annotated and final versions of each corpus

Table 4 presents the specific distribution of entity annotations over the ten Semantic Groups in each sub-corpus, in the pre-annotated version supplied to the annotators and the final version revised by annotators.

3.2. Inter-Annotator Agreement

Inter-annotator agreement was computed between the annotators and a domain expert (one of the guidelines writers) on three random samples of 100 MEDLINE titles. For the first sample, both the annotators and the expert worked from the pre-annotated documents. The analysis of annotation differences was used to consolidate the annotation guidelines. Due to time constraints, for the other two sam-

	EMEA				MEDLINE				EPO			
	Pre-annotated		*Final*		*Pre-annotated*		*Final*		*Pre-annotated*		*Final*	
	Total	%	Total	%	Total	%	Total	%	Total	%	Total	%
ANAT	461	6.33	1,031	8.08	150	8.87	1,464	16.67	197	11.85	323	6.64
CHEM	2,330	32.01	4,696	36.80	286	16.90	1,028	11.71	771	46.39	2,445	50.26
DEVI	84	1.15	340	2.66	3	0.18	126	1.43	16	0.96	256	5.26
DISO	2,608	35.82	2,698	21.14	838	49.53	2,825	32.17	171	10.29	381	7.83
GEOG	37	0.51	98	0.77	90	5.32	113	1.29	0	0.00	0	0.00
LIVB	991	13.61	1,370	10.74	283	16.73	899	10.24	183	11.01	315	6.47
OBJC	355	4.88	415	3.25	10	0.59	97	1.10	225	13.54	310	6.37
PHEN	21	0.29	142	1.11	3	0.18	160	1.82	30	1.81	248	5.10
PHYS	293	4.02	604	4.73	29	1.71	438	4.99	69	4.15	325	6.68
PROC	100	1.37	1,367	10.71	0	0.00	1,631	18.57	0	0.00	262	5.39
All	7,280	100.00	12,761	100.00	1,692	100.00	8,781	100.00	1,662	100.00	4,865	100.00

Table 4: Number of annotations for each category in both the pre-annotated and the final corpora

ples considered, the expert revised the annotators' final annotations.

IAA for entities. We consider agreement on entities at the mention and category level. This measure evaluates whether annotators selected the same text span and assigned the same category, among the ten Semantic Groups listed in the guidelines. Due to the difficulty of evaluating the number of markable entities according to Hripcsak and Rothschild (2005), Grouin et al. (2011), and Fort et al. (2012), IAA was assessed using F-measure computed on strict boundary and type match. For the first set, agreement on entities was 62% Kappa and 77% F-mesure. For the other two sets, the agreement was 92% and 90% F-measure.

IAA for CUIs. The IAA for CUIs was computed using F-measure at the sentence level. Each sentence was indexed using the set of CUIs assigned by the annotators to entities in the sentence. IAA was assessed using F-measure computed on strict CUI match. For the first set, agreement on CUIs was 66% F-mesure. For the other two sets, the agreement was 91% F-measure.

4. Discussion

4.1. Quality of the annotations

4.1.1. Influence of pre-annotation

The annotators felt that the pre-annotation was useful at the entity level. They felt that little revision was needed to the terms selected by the automatic system. Changes mostly consisted in adding entities missed by the system. However, the annotators did not like the CUI suggestions provided automatically. For many entities, the CUI suggestion consisted in a very long list and it was often felt that looking for a CUI independently from the suggestion was a faster method for doing the entity normalization.

This perception of the pre-annotation performance is reflected in Table 3, which shows good precision and poor recall for entity pre-annotation, and Table 2, which shows that the number of CUI per entity is much higher in the pre-annotations vs. final annotations.

Both Table 2 and 3 show significant differences between the pre-annotations and the final annotations, indicating that substantial work was required to prepare the corpus.

4.1.2. Annotation challenges

We believe that the annotation task addressed in this work was particularly difficult due to the combination of large entity coverage (ten types of entities), large target vocabulary for normalization (the entire UMLS), and the specialized nature of the texts in the corpus. We elaborate on more specific difficult points below.

Annotation time. Annotation time was about 0.85 annotation per minute for a domain expert revising the pre-annotated documents, as estimated based on one sample of 100 MEDLINE titles. Annotation time reached an estimated 2.2 annotations per minute for a domain expert revising annotators' work (based on two samples of 100 MEDLINE titles). This reflects on the difficulty of the task: annotation time was overall quite long, though it was faster to revise annotators' work, compared to the automatic pre-annotations. The manual annotations were of better quality; a smaller number of changes and CUI checks were needed. Annotation time was significantly longer than for broadcast news; in Rosset et al. (2013), we estimated annotation time ranged between 35 annotations per minute (for expert annotators) and 13 annotations per minute (for novice annotators). Note that our task covered both entity annotation and entity normalization, while our previous study (Rosset et al., 2013) only covered entity annotation.

Language barrier. While the PTS is a precious source of biomedical terminology in French, many concepts are only linked to terms in English. Because the annotation instructions stated that an entity should be annotated only if a corresponding concept is found in the UMLS, the annotators may have failed to find a relevant concept in the Metathesaurus browser because they did not know how to express the concept in English. In spite of these difficulties, one third of the unique CUIs assigned to the entities in the corpus (1,939 out of 5,796) do not have a French term associated with them, including the most frequent CUI in the corpus C0087111 "Therapeutic procedure" (*Traitement*).

Complex entities. The annotation of complex entities was a source of inter-annotator disagreement, as one annotator sometimes omitted to annotate either one component or the complex entity itself. For example, the phrase "*syndrome de Moschcowitz*" (Moschowitz's syndrome) was annotated with category "DISORDER" (CUI C0034155) by two annotators, however only one annotator annotated the phrase "*syndrome*" (syndrome) with category "DISORDER" (CUI C0039082). In another case, in the phrase "*Anévrisme de l'aorte thoracique*" (thoracic aorta aneurysm), both annotators annotated "*Anévrisme*" (aneurysm) with the category "DISORDER" (CUI C0002940) and "*aorte thoracique*" (thoracic aorta) with the category "ANATOMY" (CUI C1281571), but only one of them annotated the whole phrase with the category "DISORDER" (CUI C0162872). In both cases, all the annotations were relevant.

Completeness. Annotating all the relevant entities was a challenge. One reason for this difficulty is the lack of prior knowledge of biomedical terminologies for some of the annotators. Another reason is that all concepts are not covered in the terminologies; for instance many Geographical Locations or Living Being concepts are not covered so annotators may have been tempted to assume that a concept did not exist. For example, "*Chicago*" should be annotated as "GEOGRAPHIC" entity (CUI C0008044). However, there is no concept in the UMLS for "*Detroit*", which should not be annotated.

4.2. Contributions of this work

This work reports on the on-going development of a substantial high-quality ressource to the community in a language other than English, viz. French. It provides a unique annotated corpus covering three types of biomedical text and ten Semantic Groups with discontinuous annotations and overlapping annotations. Interestingly, about one third of the concepts assigned to entities do not currently have a French term linked to them through the current versions of the terminologies. This shows the potential contribution of this corpus for terminology development. Furthermore, the corpus provides a semantic characterization of three types of biomedical text, which has adds to previous sub-domain studies (Mihğailğa et al., 2012). It can be seen from Tables 2 and 4 that the profiles of each sub-corpus are quite different in terms of semantic types represented, variety and redundancy of concepts represented.

4.3. Comparison to other work

The choices made in the design of this annotation work were guided by the goal of developing a comprehensive annotated corpus. In doing so, we built on the experience of previous annotation efforts both within and outside of the biomedical domain.

Overall annotation methodology. Entity annotation and concept mapping were performed together (Similar to CRAFT (Bada et al., 2012) but unlike SHARP (Savova et al., 2012) and NCBI disease corpus (Doğan et al., 2014) where entity annotation was a separate task, performed prior to normalization). The annotation guidelines explicitly stated that if an entity that belonged to the annotated categories was found but could not be linked to a UMLS concept, no annotation was to be created.

Normalization method. The entities annotated were to be normalized using concepts in the entire UMLS Metathesaurus (more than one million concepts). Unlike SHARP and NCBI corpora, we did not limit normalization to one or two vocabularies such as SNOMED, MeSH or OMIM (several thousand concepts). When no suitable concept was found in the UMLS to normalize a given entity, similarly to CRAFT guidelines, we chose to not create an entity annotation at all, instead of assigning a "CUI-less" concept (SHARP) or assigning a close concept or combination of concepts (NCBI). The Bacteria Biotope Corpus from BioNLP 2013 normalized habitat entities to the OntoBiotope ontology (several thousand concepts). When an entity corresponded to a concept that was not in the OntoBiotope ontology, it was linked to the closest general concept.

Semantic coverage. The *Quaero* medical corpus covers entities that belong to 10 UMLS Semantic Types. This makes it the largest annotated corpus in a language other than English for the biomedical domain. The MANTRA initiative also used the same three text types, but it currently offers silver standard (i.e., automatically obtained) annotations and only covers 9 Semantic Groups. Other corpus in English provide annotations for some of the Semantic Groups covered in the *Quaero* medical corpus. For instance, the SHARP/CLEF e-Health corpus and NCBI disease corpus cover the "DISORDER" group.

Nested entities. The i2b2 2011 (Uzuner et al., 2012; Savova et al., 2011) and BioNLP 2013 (Bossy et al., 2013) corpora feature nested entities. However, for i2b2, nesting was limited to two entity types (anatomy and medical problems); for BioNLP, there were only 3 entity global categories (bacteria, geographical, habitat). Entity nesting was permitted, including for same-type entities. In our corpus many nested entities covered the entire span of a larger entity while in the BioNLP corpus, a nested entity often covered only a small portion of the span of a larger entity.

5. Conclusion and perspectives

We presented the development of a large annotated corpus for biomedical texts in French. The annotation effort was guided by the desire to provide the community with a comprehensive entity resource. We relied on the UMLS for category definition and entity normalization.

Our experience shows that annotating a large corpus with a somewhat complex scheme is a hard task, both technically and cognitively. Recent reviews of annotation tools (Neves and Leser, 2012) and annotation frameworks (Comeau et al., 2013) are testimony to the technical issues. Annotation quality was assessed throughout the project by computing inter-annotator agreement on randomly selected corpus samples.

Further quality control is being performed on the final corpus in order to resolve inconsistencies and formatting issues. We are also planning to convert the corpus to the BioC format (Comeau et al., 2013) in order increase its accessibility and usability. The corpus will be used for organizing a challenge/shared task on entity annotation and normalization, and released to the community.

Acknowledgments

This work has been partially funded by OSEO under the Quaero program and by the French ANR Cabernet project (ANR-13-JS02-0009).

6. References

Albright, D., Lanfranchi, A., Fredriksen, A., 4th Styler, W., C, C. W., Hwang, J., Choi, J., Dligach, D., Nielsen, R., Martin, J., Ward, W., Palmer, M., and Savova, G. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, 20(5):922–930.

Arighi, C., Lu, Z., Krallinger, M., Cohen, K., Wilbur, W., Valencia, A., Hirschman, L., and Wu, C. (2011). Overview of the BioCreative III workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(161).

Bodenreider, O. and McCray, A. T. (2003). Exploring semantic groups through visual approaches. *J Biomed Inform*, 36(6):414–32.

Bodnari, A., Névéol, A., Uzuner, O., Zweigenbaum, P., and Szolovits, P. (2013). Multilingual named-entity recognition from parallel corpora. In *Proc of the CLEF 2013 Evaluation Labs and Workshop*.

Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., and Nédellec, C. (2013). BioNLP shared task 2013 – an overview of the bacteria biotope task. In *Proc of BioNLP Shared Task 2013*, pages 161–169, Sofia, Bulgaria. Association for Computational Linguistics.

Comeau, D. C., Doğan, R. I., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T. C., Wu, C. H., and Wilbur, W. J. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, page bat064.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *J Biomed Inform.*, 47:1–10.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012). Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proc of LREC*, Istanbul, Turkey. European Language Resources Association (ELRA).

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc of Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Hripcsak, G. and Rothschild, A. S. (2005). Technical brief: Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*, 12(3):296–298.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for biotextmining. *BMC Bioinformatics*, 19(Suppl 1):i180–2.

McCray, A. T., Burgun, A., and Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In *Proc of MedInfo*, volume 10, pages 216–20.

Mihğailğa, C., Batista-Navarro, R. T., and Ananiadou, S. (2012). Analysing entity type variation across biomedical subdomains. In *Proc of Building and Evaluating Resources for Biomedical Text Mining Workshop*.

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

Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., and Kors, J. A. (2013). Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In *Proc of the CLEF 2013 Evaluation Labs and Workshop*.

Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., and Zweigenbaum, P. (2012). Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proc of Linguistic Annotation Workshop (LAW-VI)*, pages 40–48, Jeju, Republic of Korea. Association for Computational Linguistics.

Rosset, S., Grouin, C., Lavergne, T., Ben Jannet, M., Leixa, J., Galibert, O., and Zweigenbaum, P. (2013). Automatic named entity pre-annotation for out-of-domain human annotation. In *Proc of Linguistic Annotation Workshop and Interoperability in Discourse (LAW-VII & ID)*, pages 168–177, Sofia, Bulgaria. Association for Computational Linguistics.

Savova, G., Chapman, W., Zheng, J., and Crowley, R. (2011). Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–65.

Savova, G., Styler, W., Albright, D., Palmer, M., Harris, D., Zaramba, G., Haug, P., Clark, C., Wu, S., and Ihrke, D. (2012). SHARP template annotations: Guidelines. Technical report, Mayo Clinic.

Tanabe, L., Xie, N., Thom, L., Matten, W., and Wilbur, W. J. (2005). Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (2012). Evaluating thestate of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, 19(5):786–91.