CAS: French Corpus with Clinical Cases

Natalia Grabar, Vincent Claveau, Clément Dalloux

CAS: French Corpus with Clinical Cases

Natalia Grabar
UMR CNRS 8163 – STL
F-59000 Lille, France
natalia.grabar@univ-lille.fr

Vincent Claveau, Clément Dalloux
CNRS, IRISA, Rennes, France
vincent.claveau@irisa.fr
clement.dalloux@irisa.fr

Abstract

Textual corpora are extremely important for various NLP applications as they provide information necessary for creating, setting and testing these applications and the corresponding tools. They are also crucial for designing reliable methods and reproducible results. Yet, in some areas, such as the medical area, due to confidentiality or to ethical reasons, it is complicated and even impossible to access textual data representative of those produced in these areas. We propose the CAS corpus built with clinical cases, such as they are reported in the published scientific literature in French. We describe this corpus, currently containing over 397,000 word occurrences, and the existing linguistic and semantic annotations.

1 Introduction

Textual corpora are extremely important for various NLP applications as they provide information necessary for creating, setting and testing these applications and the corresponding tools. Yet, in some areas, due to confidentiality or to ethical reasons, it is complicated and even impossible to access representative textual data. Medical and legal areas correspond to such examples: in the legal area, information on lawsuits and trials remain confidential, while in the medical area, the medical secret must be respected. In both situations, personal data cannot be used. For several years now, anonymization and de-identification methods and tools have been made available and provide competitive and reliable results (Ruch et al., 2000; Sibanda and Uzuner, 2006; Uzuner et al., 2007; Grouin and Zweigenbaum, 2013) reaching up to 90% precision and recall. But even de-identified data may be difficult to be freely accessed and used for the research purpose because there is a risk of re-identification of people, and more particularly of patients (Meystre et al., 2014; Grouin et al., 2015) because several medical histories are unique, or because of other reasons. Hence, the application of the de-identification tools on personal data often does not permit to make these data freely available and usable within the research context.

Yet, there is a real need for the development of methods and tools for several applications suited for such restricted areas. For instance, in the medical area, it is important to have suitable tools for information retrieval and extraction, for the recruiting of patients for clinical trials, and for performing several other important tasks such as indexing, study of temporality, negation, etc. (Embi et al., 2005; Hamon and Grabar, 2010; Uzuner et al., 2011; Fletcher et al., 2012; Sun et al., 2013; Campillo-Gimenez et al., 2015; Kang et al., 2017). Another important issue is related to the reliability of tools and to the reproducibility of study results across similar data from different sources. The scientific research and clinical community are indeed increasingly coming under criticism for the lack of reproducibility in the biomedical area (Chapman et al., 2011; Collins and Tabak, 2014; Cohen et al., 2016), as well as in other areas. First step towards the reproducibility of results is the availability of freely usable tools and corpora. In our work, we are mainly concerned by building freely available corpora from the medical area.

The purpose of our work is to introduce the CAS corpus with French medical data, containing clinical cases such as those published in scientific literature or used for the education and training of medical students. In what follows, we first present some works on creation of medical corpora stressing more particularly on corpora freely available for the research (Section 2). We then introduce and describe the CAS corpus in French (Section 3) and its current annotations. We conclude with some directions for the future work (Section 4).

2 Freely available clinical corpora

Within the medical area, we can distinguish two main types of medical corpora: scientific and clinical. *Scientific corpora* are issued from scientific publications and reporting. Such corpora are becoming increasingly available for the research thanks to the recent and less recent initiatives dedicated to the open publication, such as those promoted by the NLM (National Library of Medicine) through the PUBMED portal[1] and specifically dedicated to the biomedical area, and by the HAL[2] and ISTEX[3] initiatives, which provide generic portals for accessing scientific publications from various areas, including medicine. Such corpora describe the research works, their motivation, methods, results and issues on precise research questions. Other portals may also provide access to scientific literature following specific purposes, like indexing of reliable literature, such as proposed by HON (Boyer et al., 1997), CISMEF (Darmoni et al., 1999), and other similar initiatives (Risk and Dzenowagis, 2001). Thanks to some research works, there are also scientific corpora which provide precise annotations and categorizations. These are mainly built for the purposes of challenges (Kelly et al., 2013; Goeuriot et al., 2014) but may also be provided from works of researchers, such as POS-tag (Tsuruoka et al., 2005) and negation (Szarvas et al., 2008) annotated corpora. As for *clinical corpora*, they are related to hospital and clinical events of patients. Such corpora typically describe medical history of patients and the medical care they are undergoing. It is complicated to obtain free access to this kind of medical data and, for this reason, there are very few clinical corpora freely available for the research. In our work, we are mainly interested in clinical corpora: the proposed literature review of the existing work is aimed at clinical corpora which are freely available for the research. We present here the main existing clinical corpora.

MIMIC (Medical Information Mart for Intensive Care) corpora, now in their version III, provide the largest available set of structured and unstructured clinical data in English. MIMIC III is a single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. These data include vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. The database supports applications including academic and industrial research, quality improvement initiatives, and higher education coursework (Johnson et al., 2016). These data are widely used by researchers, for instance for the prediction of mortality (Anand et al., 2018; Feng et al., 2018), for the diagnosis identification and coding (Perotte et al., 2014; Li et al., 2018), for the study of temporality (Che et al., 2018) or for the identification of similar clinical notes (Gabriel et al., 2018) to cite just a few of such works. Data from these corpora are also used in challenges, such as I2B2, N2C2 and CLEF-eHEALTH.

I2B2 (Informatics for Integrating Biology and the Bedside)[4] is an NIH-funded initiative promoting the development and test of NLP tools for healthcare improvement. In order to enhance the ability of NLP tools to process fine grained information from clinical records, I2B2 challenges provide sets of fully deidentified clinical notes enriched with specific annotations (Uzuner, 2008; Uzuner et al., 2011; Sun et al., 2013), such as: deidentification, smoking status, medication-related information, semantic relations between entities, or temporality. The clinical corpora and their annotations built for the I2B2 NLP challenges are available now for the general research purposes.

N2C2 (National NLP Clinical Challenges)[5], held for the first time in 2018, is dedicated to the inclusion of patients in clinical trials and the detection of adverse-drug events.

CLEF-eHEALTH challenges[6] held in 2013 and 2014 provide annotations for the detection of disorders and normalization of abbreviations, in 2016 the focus was done on structuring of Australian free-text nurse notes, and in 2016 and 2017 death reports in French, provided by the CépiDc[7], have been processed for the extraction of death causes.

Finally, medical data, close to those handled in the clinical context, can be found in the clinical trials protocols. One example is the corpus of clinical trials annotated with information on numerical

[1] https://www.ncbi.nlm.nih.gov/pubmed
[2] https://hal.archives-ouvertes.fr/
[3] https://www.istex.fr/
[4] https://www.i2b2.org/NLP/DataSets/Main.php
[5] https://n2c2.dbmi.hms.harvard.edu/
[6] https://sites.google.com/site/shareclefehealth/
[7] http://www.cepidc.inserm.fr/

*A term female infant was born by vaginal delivery with normal birth weight, body length and APGAR score, from a 42-year-old mother with 13 previous pregnancies resulting in 3 miscarriages and 10 live births. The mother had no history of antenatal medical illness nor of exposure to smoking, drinking and other drugs. At birth, general and systemic examination revealed a round face, single palmar crease, left precordial systolic murmur. Two hours after birth a deterioration of the general condition occurred, with generalized hypotonia, cyanosis, poor feeding. The blood count revealed white blood cell count of 35.6*10 /μL with 20.6*10 /μL, 57.9% monocytes, normal neutrophils, lymphocytes and eosinophils count, hemoglobin levels of 19.1 g/dl and 27*10 /μL platelets count. The acute phase reactants were negative. Because she maintained the altered general condition and the platelets ranged between 17-18 *10 /μL, on the 8th day after birth she was referred to our unit for proper diagnosis and treatment. Physical examination showed a phenotype suggestive for Down syndrome, later confirmed by karyotyping (47, XX + 21). She was lethargic, tachypneic and a systolic heart murmur was observed. The liver was 2 cm below the right costal margin, along with a slight enlargement of the spleen. The laboratory tests on the first day of admission in our unit revealed a white blood count of 15.8*10 /μL, with an abnormal monocyte count (increased absolute and percentile count: 5.66*10 /μL, respectively 35.5%), normal absolute neutrophil count (5.53*10 /μL), a hemoglobin level of 15.9 g/dl and severe thrombocytopenia (15*10 /μL). The biochemical parameters including electrolytes, uric acid, creatinine, bilirubin, liver enzymes were normal. The serum lactate dehydrogenase was raised. The bacterial culture work-up and titers of antibodies against toxoplasmosis, cytomegalovirus, Epstein Barr virus, hepatitis C, HIV were negative. The peripheral blood smear presented atypical cells. The bone marrow aspiration showed hemodiluted aspirate with blast cells. Immunophenotyping revealed 23% blast cells, positive for megakaryocytic markers (CD42b, CD41, CD61), myeloid markers (CD33), progenitor cell markers (CD117, CD34) and T cell marker - CD7 positive. MPO and HLA/DR were negative. The mutational status of AMLETO, PML-RARα, FLT3 and NPM1 fusion genes came out absent. The positive diagnosis was acute megakaryoblastic leukemia (AMKL).*
The echocardiography found a patent foramen ovale. The infant underwent chemotherapy according to the Down syndrome-specific AML chemotherapy protocol, consisting in four cycles of treatment: the first two cycles (induction phase) included combinations of cytarabine and liposomal daunorubicin and the last two cycles (consolidation phase): etoposide, cytarabine and mitoxantrone. Our patient aquired clinical and hematological remission without serious adverse events.

Figure 1: Example of clinical case

values in English (Claveau et al., 2017), and on negation in French and Brazilian Portuguese (Dalloux et al., 2018).

3 The CAS corpus

3.1 Content of the corpus

We present the CAS corpus in French. It contains clinical cases such as published in scientific literature and training material. Cases from these different sources are included in the corpus. Usually, the source data are available as pdf files. Their conversion in the text format is automatic but then needs to be fully checked out in order to correct potential segmentation errors (remove the paratext specific to a given journal, verify the conversion of columns, of end of lines and pages, etc.).

Similarly to clinical documents, the content of clinical cases depends on the clinical situations which are illustrated and on the disorders, but also on the purpose of the presented cases (description of diagnoses, treatments or procedures, expected audience, etc.).

Figure 1 presents an example of clinical case in English. Such data are de-identified by the auhors and their publication is done with the written permission of patients. The case reports can be related to any medical situation (diagnosis, treatment, procedure, follow-up...) and to any disorder. Publication of clinical cases usually has didactic purposes: train medical students, report on unusual or new clinical situations, present novel treatment or imaging issue... A typical structure of publications with clinical cases starts with the introduction to the clinical situation, then one or more clinical cases are presented to support the situation. Schemas, imaging, examination results,

word	PoS	lemma	uncert. cue	uncert. scope	CUI	neg cue	neg scope
L'	B-determiner	le	O	O	O	O	O
adolescent	B-common_noun	adolescent	O	O	B-C0205653	O	O
parait	B-present_verb_form	paraître	B-u-1	O	O	O	O
triste	B-adjective	triste	O	B-u-1	O	O	O
et	B-coordination_conjunction	et	O	O	O	O	O
ne	B-adverb	ne	O	O	O	B-n-1	O
parle	B-present_verb_form	parler	O	O	O	O	B_n-1
pas	B-adverb	pas	O	O	O	I-n-1	O
.	B-ending_punctuation_mark	.	O	O	O	O	O

Table 1: Example of the annotated sentence from the corpus (B-u-x stands for the beginning of the uncertainty cue or scope number x, B-n-y for the negation cue or scope number y)

patient history, lab results, clinical evolution, treatment, etc. can also be provided for the illustration of clinical cases. Finally, these clinical cases are discussed. Hence, such cases may present an extensive description of medical problems. Such publications gather medical information related to clinical discourse (clinical cases) and to scientific discourse (introduction and discussion). Related scientific literature is also provided.

As we can see from Figure 1, the clinical part of publications on clinical cases may be very similar to clinical documents: it describes patients, and proposes their diagnosis based on examination, imaging, and biological and genetic information. Besides, numerical values and abbreviations are also present. Misspellings, which are quite frequent in clinical documents, may be missing in publications on clinical cases.

3.2 Annotation of the corpus

Currently, the corpus contains linguistic and semantic annotations.

At the linguistic level, the corpus is PoS-tagged and lemmatized with a tool developed in-house and available as a web-service at `https://anonymized_url`. Then, several layers of semantic annotation are performed automatically:

- *Concept Unique Identifiers (CUI)* corresponding to French terms from the UMLS (Lindberg et al., 1993) for single or multi-word terms. For multi-word terms, the annotations exploits the IOB (Inside-Outside-Begin) format. For instance, the two-word term *vitamine B12* is encoded as follows:

  ```
  ...         O
  vitamine    B-C0042845
  B12         I-C0042845
  ...         O
  ```

  In the current version of the corpus, in case of several concurrent CUIs, only the longest, and supposedly more precise, CUIs are kept. For instance, *carence en vitamine B12 (deficiency in B12 vitamin)* (C0042847) will be preferred to *vitamine B12* (C0042845);

- *Negation*. Negation indicates whether a given disorder, procedure or treatment are present or not in the medical history and care of a given patient. For this reason, its annotation and detection are important. We adopt the approach proposed by Fancellu et al. (2016) and adapted for French by Dalloux et al. (2018) based on Machine Learning techniques trained on annotated data. This follows a two-step process: (1) the negation markers are detected with a specifically trained CRF; (2) the scope of each detected marker is found with a neural network (Bi-LSTM with a CRF layer). On the French and English data tested, the detection of negation gives up to 0.98 for the cues and 0.86 for their scope;

- *Uncertainty*. Uncertainty is also an integral part of medical discourse and should be taken into account for a more precise computing of the status of disorders, procedures and treatments. A set of markers has been built manually. It contains simple and complex lexical markers like *probablement, certainement (probably, certainly)* and morphological cues like conditional verbs (*indiquerait, proviendrait (should indicate, may be caused by)*). These markers and cues are projected on the corpus and their scope are found by heuristic rules. Detection of uncertainty gives about

type	# annotations
CUI	47,708
uncertainty	4,723
negations	4,620

Table 2: Statistics on annotations

0.90 F-measure for the cues and 0.80 for the scope.

Since there may be several markers of negation and uncertainty in a sentence, they are numbered with their scopes accordingly.

In Table 1, we present an excerpt from the corpus with all the aforementioned linguistic and semantic annotations for the sentence *L'adolescent paraît triste et ne parle pas. (The teenager seems to be sad and doesn't speak.)*

3.3 Annotation statistics

Overall, the corpus currently contains 20,363 sentences and over 397,000 word occurrences excluding punctuation marks. Table 2 indicates the number of units automatically recognized for each category.

4 Conclusion

We presented a new corpus in French which provides medical data close to those produced in the clinical context: description of clinical cases and their discussion. Overall, the corpus currently contains over 397,000 word occurrences excluding punctuation marks. The corpus is currently annotated with several layers of information: linguistic (PoS-tagging, lemmas) and semantic (the UMLS concepts, uncertainty, negation and their scopes). The corpus will be enriched with more clinical cases published. Other annotation layers will be added and their correctness cross-validated by human annotators. The enriched version of the corpus will undergo a more detailed description, such as statistics on age and gender of patients, their diseases, or the sources of publications.

Besides, similar corpora will be built for other languages. For instance, the repository of clinical cases in English is available on a dedicated website *Archive of Clinical Cases*[8] respecting the Creative Commons License.

The very purpose of our work is to make these annotated corpora freely available for research. We expect that this may encourage development of robust NLP tools for medical free-text documents in French and other languages.

[8] http://www.clinicalcases.eu

Acknowledgements

This work was partly funded by the French government support granted to the CominLabs LabEx managed by the ANR in Investing for the Future program under reference ANR-10-LABX-07-01.

The authors would like to thank Cyril Grouin for the discussions on existing medical corpora and the reviewers for their helpful comments.

References

RS Anand, P Stey, S Jain, DR Biron, H Bhatt, K Monteiro, E Feller, Ranney ML, Sarkar IN, and Chen ES. 2018. Predicting mortality in diabetic icu patients using machine learning and severity indices. In *AMIA Jt Summits Transl Sci Proc*, pages 310–319.

Celia Boyer, O Baujard, Vincent Baujard, S Aurel, M Selby, and RD Appel. 1997. Health on the net automated database of health and medical information. *Int J Med Inform*, 47(1-2):27–9.

B Campillo-Gimenez, C Buscail, O Zekri, B Laguerre, E Le Prisé, R De Crevoisier, and M Cuggia. 2015. Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. *Trials*, 16(1):1–15.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, 18(5):540–543.

Z Che, S Purushotham, K Cho, D Sontag, and Y Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Sci Rep*, 8(1):6085.

Vincent Claveau, Lucas Emanuel Silva Oliveira, Guillaume Bouzillé, Marc Cuggia, Claudia Maria Cabral Moro, and Natalia Grabar. 2017. Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation. In *AIME (Artifical Intelligence in Medicine in Europe)*.

K. Bretonnel Cohen, Jingbo Xia, Christophe Roeder, and Lawrence E. Hunter. 2016. Reproducibility in natural language processing: A case study of two r libraries for mining pubmed/medline. In *LREC Int Conf Lang Resour Eval*, pages 6–12.

FS Collins and LA Tabak. 2014. Nih plans to enhance reproducibility. *Nature*, 505:612–613.

Clément Dalloux, Vincent Claveau, Natalia Grabar, and Claudia Moro. 2018. Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien. In *TALN 2018*, pages 1–6.

SJ Darmoni, JP Leroy, F Baudic, M Douyère, J Piot, and B Thirion. 1999. CISMeF: cataloque and index of french speaking health resources. In *Stud Health Technol Inform*, pages 493–6.

PJ Embi, A Jain, J Clark, and CL Harris. 2005. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *Ann Symp Am Med Inform Assoc (AMIA)*, pages 231–35.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *An Meeting of the Ass for Comp Linguistics*, volume 1.

M Feng, JI McSparron, DT Kien, DJ Stone, DH Roberts, RM Schwartzstein, A Vieillard-Baron, and LA Celi. 2018. Transthoracic echocardiography and mortality in sepsis: analysis of the mimic-iii database. *Intensive Care Med*, 44(6):884–892.

B Fletcher, A Gheorghe, D Moore, S Wilson, and S Damery. 2012. Improving the recruitment activity of clinicians in randomised controlled trials: A systematic review. *BMJ Open*, 2(1):1–14.

RA Gabriel, TT Kuo, J McAuley, and CN Hsu. 2018. Identifying and characterizing highly similar notes in big clinical note datasets. *J Biomed Inform*, 82:63–69.

Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth Jones, and Henning Müller. 2014. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *CLEF*, Lecture Notes in Computer Science (LNCS), pages 43–61. Springer.

Cyril Grouin, Nicolas Griffon, and Aurélie Névéol. 2015. Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? In *Proc of LOUHI*, Lisbon, Portugal.

Cyril Grouin and Pierre Zweigenbaum. 2013. Automatic de-identification of french clinical records: Comparison of rule-based and machine-learning approches. In *Stud Health Technol Inform, Proc of MedInfo*, volume 192, pages 476–80, Copenhagen, Denmark.

T Hamon and N Grabar. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, 17(5):549–54.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-iii, a freely accessible critical care database. *Scientific Data*, 3(160035):1–9.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noemie Elhadad, and Chunhua Weng. 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc*, 24(6):1062–1071.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, Guido Zuccon, and Joao Palotti. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *CLEF*, Lecture Notes in Computer Science (LNCS). Springer.

M Li, Z Fei, M Zeng, F Wu, Y Li, Y Pan, and J Wang. 2018. Automated ICD-9 coding via a deep learning approach. In *IEEE/ACM Trans Comput Biol Bioinform*.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods Inf Med*, 32(4):281–291.

Stephane Meystre, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014. Can physicians recognize their own patients in de-identified notes? In *Stud Health Technol Inform 205*, pages 778–82.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *J Am Med Inform Assoc*, 21:231–237.

Ahmad Risk and J Dzenowagis. 2001. Review of internet information quality initiatives. *Journal of Medical Internet Research*, 3(4).

Patrick Ruch, Robert H. Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. In *Ann Symp Am Med Inform Assoc (AMIA)*, pages 729–733, Los Angeles, CA.

T Sibanda and O Uzuner. 2006. Role of local context in de-identification of ungrammatical, fragmented test. In *NAACL-HLT 2006*, New York, USA.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, 20(5):806–813.

G Szarvas, V Vincze, R Farkas, and J Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *BIONLP*, pages 38–45.

Y Tsuruoka, Y Tateishi, JD Kim, T Ohta, J McNaught, S Ananiadou, and J Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *LNCS*, 3746:382–392.

O Uzuner. 2008. Second i2b2 workshop on natural language processing challenges for clinical records. In *Ann Symp Am Med Inform Assoc (AMIA)*, pages 1252–3.

O Uzuner, Y Luo, and P Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14:550–563.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556.