An Overview of Terminological Resources available for French bioNLP

Aurélie Névéol[1], Dietrich Rebholz-Schuhmann[2,3], and Pierre Zweigenbaum[1]

[1] LIMSI-CNRS UPR 3251, Rue John von Neumann, 91400 Orsay, France
[2] ICL, University of Zurich, 8050 Zurich, Switzerland
[3] EBI, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

Abstract. Terminological resources in languages other than English are playing a critical role in all kinds of medical administrative solutions in countries where English is not an official language, or not the only official language. French benefits from being among the top five most covered languages in the UMLS in number of concepts. Herein, we review the extent of this coverage as well as the sources and methods for additional material available for French.

Keywords: France; Natural Language Processing; Translating; Terminology as topic/methods

1 Introduction

While the bulk of the biomedical literature is available in English, many other important documents in the biomedical domain are written in other languages, such as electronic health records. Furthermore, there is a large audience of non-English speakers who can benefit from accessing health information in their native language. For these reasons, it is crucial to ensure that the work being done in Natural Language Processing in the biomedical domain (bioNLP) is not limited to English. Herein, we provide an overview of the terminological and linguistic resources that are currently available for French bioNLP. We describe the contributions of institutional organizations, independent research groups and individuals. Finally, we analyze the results of these contributions to provide some insight into successful efforts, open needs, and conclude with some recommendations for the community to address them in the near future.

2 Material and methods

2.1 Currently available resources

The UMLS® (Unified Medical Language System®; Bodenreider 2004) is a major terminological resource for many languages, including French. The UMLS Metathesaurus® is the structured union of existing biomedical vocabularies, some of which include terms in languages other than English. The second row of

Table 1 provides an overview of the various sources in the UMLS Metathesaurus (version 2012AA) with an indication of size using the number of unique strings and Concept Unique Identifiers (CUIs) covered[4]. As far as French is concerned, the MeSH® thesaurus (Medical Subject Headings) and the Medical Dictionary for Regulatory Activities (MedDRA) make the bulk of the French terms available in the UMLS.

Rows 3 to 5 in Table 1 show some of the other terminologies available through the UMLS for English that are also available in French but not included in the UMLS release. It should be noted that some light processing is required to map the French terms (strings) from these resources to UMLS CUIs as the data is provided natively for each terminology, i.e. each term is mapped to a terminology unique identifier which is also present in the UMLS Metathesaurus. Using the English version of the UMLS, the mappings between these identifiers and UMLS CUIs can be recovered: for instance, the French term "bras robotique" in SNOMED v3.5 has code A-01100 which is found in the Metathesaurus with CUI C0336542 and preferred English term "Robotic arm". Therefore, it can be inferred that the French term "bras robotique" should be assigned to CUI C0336542.

Table 1. Overview of terminological resources in French with links to the UMLS (2012AA) that are publicly available

Source vocabulary	Number of Strings	Number of CUIs
UMLS 2012AA	171,764	85,685
MSHFRE	105,758	48,005
MDRFRE	65,071	41,229
WHOFRE	3,631	3,091
MTHMSTFRE	1,833	1,636
ICPCFRE	702	722
ICD10[1]	26,337	12,143
FMA[2]	4,564	4,452
SNOMED[3]	139,792	93,632
All	336,264	169,123

[1] `http://www.icd10.ch/index.asp`
[2] `http://sig.biostr.washington.edu/projects/fma/release/index.html`
[3] `http://esante.gouv.fr/snomed/snomed`

Other resources in French have not been mapped to the UMLS. The European Health Terminology Ontology Portal (EHTOP[5]) provides an integrated access to four of them: CCAM, ATC, Orphanet, and CISMeF. CCAM (Classification

[4] This number was obtained using the following SQL query: `select LAT, count(distinct STR collate utf8_bin) from MRCONSO where LAT="FRE";` and similarly for the CUIs.
[5] `http://www.ehtop.eu/`

Commune des Actes Médicaux) is a French counterpart of CPT (Current Procedural Terminology). The Anatomical Therapeutic Chemical (ATC) classification system is a drug classification system managed by the World Health Organization (WHO); its counterpart in the UMLS is RXNORM, which is developed by the US National Library of Medicine. Orphanet is a portal of information on rare diseases which has driven the development of a terminology which, in turn, led to the OntoOrpha ontology. CISMeF (Catalogue et Index des Sites Médicaux Francophone; Darmoni et al. 2000) is an online portal providing health information in French relying on MeSH indexing of online documents. The so-called CISMeF terminology comprises MeSH as well as additional resources in French: additional Entry Terms, additional resources types, and metaterms, which are broad categories corresponding to medical specialties. CISMeF metaterms comprise links to relevant MeSH terms, allowing for specialty-oriented document searches and specialty profiling of corpora. Finally, we can also mention the case of the Foundational Model of Anatomy (FMA). The English version of this ontology is part of the UMLS and is also included in ETHOP. In addition to English, a handful of other languages are partially covered, including French: 4,452 concepts out of 82,083 (5.4%) have at least one term in French.

Then, the issue of string normalization and pre-processing for use in Natural Language Processing applications remains to be addressed. Many of the strategies used for English are also applicable for French [9, 14].

Over the past decade, there has been a collaborative effort to develop a French "Specialist Lexicon" comprising linguistic information about inflected forms and lemmas found in biomedical terms [15, 3]. Table 2 shows details of the UMLF (Unified Medical Language for French) contents.

Table 2. Overview of biomedical linguistics resources available in the UMLF

Type	**Number of Strings**	**Sample entry**
Adjective	41,972	diabétiques|diabétique|Afpmp
Noun	34,984	insuline|insuline|Ncfs
Proper Name	6,522	Marfan|Marfan|Np
Latin	3,357	abducens|abducens|LAT
Adverb	1,586	initialement|initialement|Rgp
Function word	410	dans|dans|PREP
Present participle	12	recombinant|recombiner|Vmpp—
All	**88,847**	

This contributes a first level of normalization based on lemmatization. The UMLF lexicon is freely available for research purposes upon request from the authors.

2.2 Contributors

Contributions to the resources described above came from many researchers throughout Canada, France, and Switzerland. These efforts came from institutional initiatives or from the concerted work of individual researchers.

INSERM has been a partner of the National Library of Medicine since 1969. They have been producing a version of MeSH translated into French every year since 1986. While all main headings were available in French, only selected Entry Terms are also translated. To address the need for additional Entry Terms in French, INIST has helped INSERM with the manual translation of several thousand terms. Meanwhile, the CISMeF team has been enhancing the MeSH resources available in French by adding synonyms [7] and researching methods for automatically translating Entry terms from English into French ([11, 13, 6]. Other efforts towards building an automatic tool for MeSH indexing also resulted in the development of MeSH terminological resources [12].

The French version of SNOMED is the results of the efforts of Dr. R.A. Côté [4] in the development of SMOMED v3.5 (also known as SNOMED International). It has been used together with other resources to explore methods for obtaining a French version of SNOMED CT [1]. This work was part of a larger effort to develop a portal integrating medical terminologies in French and other European languages [8]: the above mentioned EHTOP portal.

Several nationally-funded projects contributed to the extension of these resources. Beyond the already cited UMLF project, let us mention VUMeF [5] and then InterSTIS (`http://www.interstis.org/`) which helped in the French translation of MeSH and SNOMED.

3 Discussion

3.1 Needs

We identify three important needs for enhancing the existing resources for French, in order of increasing difficulty:

1. **integrate available resources into UMLS** in order to facilitate their access and use: for instance, French versions of ICD10, FMA, SNOMED;
2. **enhance the content of resources currently integrated in the UMLS** with the result of work that has been on-going for the past decade, e.g. UMLF, MeSH terminological resources for indexing;
3. have **the UMLS cover terminologies that are used in clinical practice** in countries where French is an official language used in hospitals (France, Canada, Belgium, Switzerland, French-speaking African countries, etc.). For instance, as mentioned above, terminologies such as CCAM are not part of the UMLS. [10, 2] recently proposed semi-automated methods to map CCAM terms to UMLS CUIs. However, there is a need for terminology developers to officially validate the results of the automatic mappings and promote/authorize the distribution of the final resource. Besides, the

composed nature of CCAM terms makes them very specific, which entails a limited exact correspondence to procedures defined in other medical vocabularies and a need to create new concepts in the UMLS Metathesaurus.

3.2 Barriers

We consider that the main barriers to further progress are the following:

1. **lack of a concerted effort**: while the NLM is prioritizing the development of resources in English, some resources for other languages are developed for a particular application and the means (time, funds) to produce a distributable version are not available. A roadmap for further developments is needed, listing aims and objectives, and methods to obtain appropriate support from relevant teams and organizations. A project such as MANTRA could provide a venue for preparing such a concerted effort. Additionally, societies such as IMIA could provide a longer-term framework for such an initiative. For instance, its recently created Francophone SIG intends to address the promotion of work on French-language corpora and controlled vocabularies. Coordinated action in different languages could build on language-specific actions.
2. **licensing issues**: the access to some terminologies is restricted (SNOMED CT, ATC, CISMeF, ...); terminologies underpin the development of science, standards and information systems and should be made widely available at no cost.

4 Conclusion

This overview of French resources for the biomedical domain shows that there is a real interest from the community for using these resources, as well as efforts to create quality resources. However, there is an on-going need for encouraging the sharing of the end-results under the umbrella of a standard framework such as the UMLS. Funded projects are a possible means for this type of activity (see, e.g., UMLF, VUMeF, InterSTIS, EHTOP, MANTRA); societies (e.g., IMIA) are another framework where it could be coordinated.

Acknowledgements

The authors acknowledge travel funding from the EU STREP project grant 296410 ("Mantra") under the 7th EU Framework Programme within Theme "Intelligent Content and Semantics", "Challenge 4: Technologies for Digital Content and Languages" (ICT 2011.4.1).

References

1. Abdoune, H., Merabti, T., Darmoni, S.J., Joubert, M.: Assisting the translation of the CORE subset of SNOMED CT into French. In: Moen, A., Andersen, S.K., Aarts, J., Hurlen, P. (eds.) MIE. Studies in Health Technology and Informatics, vol. 169, pp. 819–823. IOS Press (2011)
2. Bousquet, C., Souvignet, J., Merabti, T., Sadou, E., Trombert, B., Rodrigues, J.M.: Method for mapping the French CCAM terminology to the UMLS metathesaurus. In: Mantas, J., Andersen, S.K., Mazzoleni, M.C., Blobel, B., Quaglini, S., Moen, A. (eds.) Stud Health Technol Inform. Studies in Health Technology and Informatics, vol. 180, pp. 164–168. IOS Press (2012)
3. Cartoni, B., Zweigenbaum, P.: Extension of a specialised lexicon using specific terminological data: the Unified Medical Lexicon for French (UMLF). In: Proc. EURALEX (2010)
4. Côté, R.A., Brochu, L., Cabana, L.: SNOMED Internationale – Répertoire d'anatomie pathologique. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec (1997)
5. Darmoni, S.J., Jarrousse, E., Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., Joubert, M., Vallée, H., Côté, R.A., Buemi, A., Bourigault, D., Recourcé, G., Jeanneau, S., Rodrigues, J.M.: VUMeF: Extending the French involvement in the UMLS Metathesaurus. In: Proc. AMIA Symp. p. 824 (2003), (poster)
6. Deléger, L., Merabti, T., Lecrocq, T., Joubert, M., Zweigenbaum, P., Darmoni, S.J.: A twofold strategy for translating a medical terminology into French. In: Proc AMIA Symp. pp. 152–156 (2010)
7. Douyère, M., Soualmia, L.F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J., Thirion, B., Darmoni, S.J.: Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Information and Libraries Journal (2004)
8. Grosjean, J., Merabti, T., Griffon, N., Dahamna, B., Darmoni, S.J.: Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. In: Proc. 9th international conference on Terminology and Artifical Intelligence. pp. 119–122 (2011)
9. Hettne, K.M., van Mulligen, E.M., Schuemie, M.J., Schijvenaars, B.J.A., Kors, J.A.: Rewriting and suppressing UMLS terms for improved biomedical term identification. J. Biomedical Semantics 1, 5 (2010)
10. Merabti, T., Massari, P., Joubert, M., Sadou, E., Lecroq, T., Abdoune, H., Rodrigues, J.M., Darmoni, S.J.: An automated approach to map a French terminology to UMLS. Stud Health Technol Inform 1040–4(Pt 2), 5 (2010)
11. Névéol, A., Ozdowska, S.: Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In: Pinson, S., Vincent, N. (eds.) EGC. Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-3, pp. 655–666. Cépaduès-Éditions (2005)
12. Névéol, A., Douyère, M., Rogozan, A., Darmoni, S.J.: Construction de ressources terminologiques en santé pour un système d'indexation automatique. In: Proc. Journées INTEX/NOOJ. Tours, France (2004)
13. Ozdowska, S., Névéol, A., Thirion, B.: Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. In: Proc. 6th international conference on Terminology and Artifical Intelligence. pp. 83–84 (2005)
14. Wu, S.T.I., Liu, H., Li, D., Tao, C., Musen, M.A., Chute, C.G., Shah, N.H.: Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. JAMIA 19(e1) (2012)

15. Zweigenbaum, P., Baud, R.H., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Le Duff, F., Forget, J.F., Douyère, M., Darmoni, S.: A unified medical lexicon for French. International Journal of Medical Informatics 74(2–4), 119–124 (Mar 2005), http://dx.doi.org/10.1016/j.ijmedinf.2004.03.010