

Language Resources for French in the Biomedical Domain

Aurélie Névéal¹, Julien Grosjean², Stéfan J. Darmoni^{2,3}, Pierre Zweigenbaum¹

¹LIMSI-CNRS UPR 3251
91400 Orsay, France
{neveol,pz}@limsi.fr

²CISMef-TIBS-LITIS EA 4108, Rouen University Hospital
76031 Rouen, France
{julien.grosjean,stefan.darmoni}@chu-rouen.fr

³LIMICS-Inserm UMR 1142
75006 Paris, France

Abstract

The biomedical domain offers a wealth of linguistic resources for Natural Language Processing, including terminologies and corpora. While many of these resources are prominently available for English, other languages including French benefit from substantial coverage thanks to the contribution of an active community over the past decades. However, access to terminological resources in languages other than English may not be as straight-forward as access to their English counterparts. Herein, we review the extent of resource coverage for French and give pointers to access French-language resources. We also discuss the sources and methods for making additional material available for French.

Keywords: BioNLP; Corpora; French; Specialized Domain; Terminology

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 (EHR), medical school lecture notes, or clinical practice guidelines. Terminological resources in languages other than English are playing a critical role in many medical administrative solutions in countries where English is not an official language, or not the only official language. 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. In practice, other languages, including French, benefit from substantial coverage thanks to the contribution of an active community over the past decades. However, accessing terminological resources in languages other than English may not be as straight-forward as accessing their English counterparts. For instance, some English-language resources that are directly available from any UMLS[®] (see section 2) distribution are only available directly from the individual terminology providers for French. In addition to creating a technical barrier for access, this results in frequent underestimation of the amount of resources available. For languages other than English, it is difficult to know whether a resource that is not present in the UMLS is either not available or not distributed through this channel. The latter is a common assumption.

One major contribution of this paper is to clearly identify linguistic resources available for French and where to find them. This survey extends (Névéal et al., 2013) by provid-

ing information not only on biomedical terminologies, but also on text corpora, and giving more detailed information on lexical and terminological resources. We describe the contributions of institutional organizations, independent research groups and individuals to the development of these resources. Finally, we analyze the results of a variety of 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. Survey of Resources and Tools

In this section, we review the resources and tools that are relevant to natural language processing for French in the biomedical domain, including terminological resources, linguistic resources and corpora. We also review tools that exploit these resources or enhance them.

2.1. Terminologies and Ontologies

The UMLS (Unified Medical Language System[®]; Bodenreider 2004) is a major terminological resource in the biomedical domain 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) covered¹. The importance of increasing the coverage of French in the UMLS has

¹This 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.

been recognized early on with a roadmap for improving the French versions of MeSH[®] (Medical Subject Headings) and SNOMED v3.5 (Darmoni et al., 2003). As a result, the MeSH thesaurus and the Medical Dictionary for Regulatory Activities (MedDRA) make the bulk of the French terms available in the UMLS. The French version of MeSH has been produced by INSERM (the French Institute for Health and Medical Research) through a partnership with the National Library of Medicine. While all main headings have been available in French since 1986, only selected Entry Terms were initially translated. To address the need for additional Entry Terms in French, INIST (the French Institute for Scientific and Technical Information) has been helping INSERM with the manual translation of several thousand terms. Meanwhile, several efforts contributed to enhancing the MeSH resources available in French by adding synonyms (Douyère et al., 2004) and researching methods for automatically translating Entry terms from English into French ((Névéal and Ozdowska, 2005; Ozdowska et al., 2005; Névéal and Ozdowska, 2006; Deléger et al., 2010). Other efforts towards building an automatic tool for MeSH indexing also resulted in the development of MeSH terminological resources (Névéal et al., 2004).

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 2 shows the results of on-going translation efforts coordinated by the CISMeF team. The corresponding terms are displayed to users in HeTOP.

The French version of SNOMED was part of SNOMED v3.5, also known as SNOMED International (Côté et al., 1997). It has been used together with other resources to explore methods for obtaining a French version of SNOMED CT (Abdoune et al., 2011).

There is currently no equivalent to the UMLS for terminologies that are available in French only. These resources are developed by different individuals and organizations that control the format and releases. The resources may be available freely in formats that are challenging for computer processing. For instance, the ADICAP² Thesaurus is distributed under a creative common license as a pdf document intended for human perusal. It can also be noted that some of the French-language resources are intended for an

²Association pour le Développement de l’Informatique en Cytologie et Anato-mo-Pathologie, a French association supporting informatics for cytology and anatomopathology applications. <http://www.adicap.asso.fr>

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	41,229
MDRFRE	65,071	48,005
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
ICNP ⁴	2,801	1,158
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>

⁴ <http://www.icn.ch/fr/piliers-et-programmes/icnpr-translations/>

Table 2: Overview of terminological resources in French with links to the UMLS that can be browsed in HeTOP

Source vocabulary	Number of French Strings	Number of English Strings
NCIt ¹	31,334	181,070
FMA	31,080	128,033
RadLex ²	11,543	42,411
HPO ³	10,867	10,206
OMIM ⁴	6,627	22,905
MedlinePlus ⁵	878	849

¹ NCI Thesaurus, <http://ncit.nci.nih.gov>

² <http://www.rsna.org/RadLex.aspx>

³ Human Phenotype Ontology, <http://www.human-phenotype-ontology.org/>

⁴ Online Mendelian Inheritance in Men, <http://www.omim.org>

⁵ <http://www.nlm.nih.gov/medlineplus/>

administrative use within French health care structures and focus on concepts defined by the French health department regulations. For instance, NABM³ provides a list of laboratory tests covered by the French national health insurance. Table 3 provides a list of French-only terminologies. Information on these resources (as well as links to the providers) is available from the “Description” tab of the Health Terminology/Ontology Portal (HeTOP), which we describe further in section 2.3.

2.2. Linguistic Resources

To make a full use of the terminology and ontology resources described above, the issues 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 (Hettne et al., 2010; Wu et al., 2012); however, language-specific

³Nomenclature des Actes de Biologie Médicale

Table 3: Overview of French-only terminological resources

Source vocabulary	Number of (French) Strings	Number of Native Concepts	Number of CUIs mapped	Availability
BNPC ¹	91,750	103,280	30,564	proprietary
CCAM	9,666	18,314	7	browsing in HeTOP
ADICAP	9,189	9,189	143	browsing in HeTOP
Cladimed ²	4,546	4,548	191	proprietary
LPP ³	4,546	4,546	0	browsing in HeTOP
DRC ⁴	3,313	3,324	520	browsing in HeTOP
BHN ⁵	2,534	2,544	0	browsing in HeTOP
NABM	1,084	1,084	0	browsing in HeTOP
BNCI ⁶	786	802	351	proprietary

¹ Base Nationale des Produits et des Compositions

² CLAssification des DIpositifs MÉDicaux, <http://www.cladimed.com/>

³ Liste des Produits et des Prestations, http://www.codage.ext.cnamts.fr/codif/tips/index_presentation.php?p_site=AMELI

⁴ Dictionnaire des Résultats de Consultation, http://www.sfm.org/demarche_medicale/demarche_diagnostique/dictionnaire_des_resultats_de_consultation/

⁵ Biologie Hors Nomenclature, http://www.chu-montpellier.fr/publication/inter_pub/R300/rubrique.jsp

⁶ Base Nationale des Cas d'Intoxications

Table 4: Overview of biomedical lexical resources available in the UMLF

Type	# 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	codant coder Vmpp
All	88,847	

resources are also needed.

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 (Zweigenbaum et al., 2005; Cartoni and Zweigenbaum, 2010). Table 4 shows details of the UMLF (Unified Medical Language for French) contents. This contributes a first level of normalization based on lemmatization.

The UMLF lexicon is freely available for research purposes upon request from the authors. The UMLF has previously been used as a core component in a morphosemantic linguistic-based parser called DériF (Namer and Baud, 2007), which provides an automatic analysis of French biomedical derived and compound words. The results of this analysis include definitions of complex medical terms.

Further work addressed the extraction of relationships between lemmas, including orthographic variants and derivational relations. Specifically, as shown in table 5, a number of derivational relations involving adjectives have been extracted, relying on morphological principles.

2.3. Visualization and Search Tools

One of the first tools available to visualize Medical Terminology in French was developed by Inserm to navigate in the French version of the MeSH thesaurus <http://mesh.inserm.fr/mesh/>.

CISMeF (Catalogue et Index des Sites Médicaux Francophone; (Darmoni et al., 2000)) is an online portal providing access to 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 (Douyère et al., 2004). CISMeF metaterms provide links to relevant MeSH terms, allowing for specialty-oriented document searches and specialty profiling of corpora. In addition to a document search interface, CISMeF features a Health Terminology/Ontology Portal (HeTOP⁴) that supports queries for French medical terms within MeSH and other terminologies. For each medical concept, HeTOP displays a description that includes links to outside terminology sources when relevant (e.g. Inserm and NLM’s MeSH browsers, NCBO BioPortal). HeTOP also provides direct access to medical search engines by creating relevant queries using the medical concept browsed.

Recently, HeTOP became an even more global effort to develop a portal integrating medical terminologies in French and up to 35 other languages (Grosjean et al., 2011). Among other services, it provides integrated access to several terminologies in French, including four that have not been mapped to the UMLS: CCAM, ATC, Orphanet, and CISMeF. CCAM (Classification Commune des Actes Médicaux) is a French counterpart of CPT (Current Procedural Terminology). The Anatomical Therapeutic Chemi-

⁴<http://www.hetop.org/>

Table 5: Overview of biomedical derivational resources available in the UMLF

Type	Number of relations	Sample entry
Adjective/Noun (deverbal)	161	occlusif occlusion
Adjective/Noun (other)	2,126	adipeux graisse
Adjective compounds	314	gastropleural plèvre estomac
Adjectival constructs	1,089	transgénique gène across
Orthographic variants (hyphen)	2,499	péri-opérateur périopérateur
Orthographic variants (oe)	89	œsophagien oesophagien
All	6,278	

cal (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.

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 HeTOP. 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. Recently, Merabti et al. (Merabti et al., 2011) assessed a variety of automatic approaches to help with translating the remainder of FMA into French.

2.4. Corpora

Corpora are important resources that can be used for many Natural Language Processing tasks in the biomedical domain, including terminology enrichment. Interestingly, many of the relevant corpora are in fact parallel corpora that are available in a number of languages, including French. In Table 6 we provide a list of these corpora, along with the number of sentences and total number of (French) words in each corpus. We also provide information on the cost and availability of each corpus.

The corpora from the European Medicines Agency (EMA), MEDLINE, and the European Patent Office (EPO, <http://www.epo.org/>) were recently used in an international challenge for cross-lingual medical named entity recognition, CLEF-ER (Rebholz-Schuhmann et al., 2013), with the larger goal of using the Multilingual Annotation of Named Entities for Terminology Resources Acquisition⁵. It is to be noted that the version EPO corpus pointed to here comprises additional comparable patent corpora in English, French and German, not restricted to the biomedical domain. A strictly parallel subset of the EPO corpus restricted to the biomedical domain is available upon request from the organizers of the 2013 CLEF-ER challenge.

The Corpus Médical du Centre de Recherche en Terminologie et Traduction (CMCR TT) is the only publicly available monolingual French corpus for the biomedical domain. It was processed with an automatic part-of-speech tagger and is available as plain or PoS tagged text (Maniez, 2009).

Also of interest even though it is not a medical corpus, the

Hansard (available for a fee from <http://www.isi.edu/natural-language/download/hansard/>) has been used in experiments to enrich medical terminologies (Névéal and Ozdowska, 2006).

3. Creating and Sharing Language Resources Efficiently

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. We believe that these efforts could benefit the community to an even greater extent through continued and strengthened collaboration.

3.1. Terminology developers play a key role in resource availability

To enhance the existing terminology resources for French, collaboration with terminology developers is essential. There is a need to integrate available resources into the UMLS in order to facilitate their access and use: for instance, French versions of ICD10, FMA, and SNOMED international (3.5) are available from the individual resource providers, or through HeTOP but not from the UMLS. Because of the aggregated nature of the UMLS, the process of having additional (French) terms directly available in the metathesaurus must organically come from the original terminology developers who supply the National Library of Medicine with the most up-to-date version of the terminology to be integrated into the UMLS.

The content of resources currently integrated in the UMLS would also be greatly enhanced with the result of work that has been on-going for the past decade, e.g. contents of the UMLF, MeSH terminological resources for indexing. As pointed out previously, this effort needs to be carried out in collaboration with terminology developers.

It would also be very useful to 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. (Merabti et al., 2010; Bousquet et al., 2012) 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

⁵MANTRA, <http://www.mantra-project.eu/>

Table 6: Overview of corpus resources in French and their availability

Corpus	Number of Sentences	Number of Words	Availability
EMEA	373 K	5,515 K	Free ¹
MEDLINE	572 K	6,024 K	Free ²
EPO	120 K	6,690 K	Free ³
Cochrane	18 K	427 K	With permission ⁴
Santé Canada	255 K	9,000 K	With fee ⁵
CMCRTT	1,526 K	26,379K	Free ⁶

¹ <http://opus.lingfil.uu.se/EMEA.php>

² titles from <http://www.ncbi.nlm.nih.gov/pubmed/>; additionally (Yepes et al., 2013) provide free software for obtaining abstract text in French

³ <http://www.ifs.tuwien.ac.at/imp/marec.shtml>

⁴ <http://www.cochrane.fr/>

⁵ Corpus CESART; http://catalog.elra.info/product_info.php?products_id=993

⁶ http://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm

to procedures defined in other medical vocabularies and a need to create new concepts in the UMLS Metathesaurus.

3.2. An international concerted action is needed

We consider that the main barrier to further progress is the lack of an international concerted effort to produce and make resources available. 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. Societies such as IMIA (International Medical Informatics Association) or ELRA (European Language Resources Association) could provide a longer-term framework for such an initiative. For instance, the recently created IMIA 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.

Furthermore, licensing issues restrict the access to some terminologies (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 paper identified a large number of French language resources for the biomedical domain, including many that are only known as English-language resources due to a divergence in the distribution of English vs. non-English versions. We also show that there is a real interest from the community for French resources, as well as efforts to create quality resources. By listing and describing French Language resources, this paper contributes to raising the awareness of the community and promoting the use of lesser-known resources. Nonetheless, 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, MANTRA); societies (e.g., IMIA) are another framework where it could be coordinated.

5. Acknowledgements

This work was supported by the French National Agency for Research under grant CABeRneT⁶ ANR-13-JS02-0009-01

6. References

- Hocine Abdoune, Tayeb Merabti, Stéfan Jacques Darmoni, and Michel Joubert. 2011. Assisting the translation of the CORE subset of SNOMED CT into French. In Anne Moen, Stig Kjær Andersen, Jos Aarts, and Petter Hurlen, editors, *MIE*, volume 169 of *Studies in Health Technology and Informatics*, pages 819–823. IOS Press.
- Cédric Bousquet, Julien Souvignet, Tayeb Merabti, Eric Sadou, Béatrice Trombert, and Jean Marie Rodrigues. 2012. Method for mapping the French CCAM terminology to the UMLS metathesaurus. In John Mantas, Stig Kjær Andersen, Maria Christina Mazzoleni, Bernd Blobel, Silvana Quaglini, and Anne Moen, editors, *Stud Health Technol Inform*, volume 180 of *Studies in Health Technology and Informatics*, pages 164–168. IOS Press.
- Bruno Cartoni and Pierre Zweigenbaum. 2010. Extension of a specialised lexicon using specific terminological data: the Unified Medical Lexicon for French (UMLF). In *Proc. EURALEX*.
- Roger A. Côté, Louise Brochu, and Lyne Cabana, 1997. *SNOMED Internationale – Répertoire d’anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.
- Stéfan J. Darmoni, J.-P. Leroy, Benoît Thirion, F. Baudic, Magali Douyere, and J. Piot. 2000. CISMef: a structured health resource guide. *Meth Inf Med*, 39(1):30–35.

⁶CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

- SJ Darmoni, E Jarrousse, P Zweigenbaum, P Le Beux, F Namer, R Baud, M Joubert, H Vallée, RA Côté, A Buemi, D Bourigault, G Recource, S Jeanneau, and JM Rodrigues. 2003. VUMeF: extending the French involvement in the UMLS Metathesaurus. In *AMIA Annu Symp Proc*, page 824.
- Louise Deléger, Tayeb Merabti, Thierry Lecroq, Michel Joubert, Pierre Zweigenbaum, and Stéfan Jacques Darmoni. 2010. A twofold strategy for translating a medical terminology into French. In *Proc AMIA Symp*, pages 152–156.
- Magaly Douyère, Lina F. Soualmia, Aurélie Névéol, Alexandra Rogozan, Badisse Dahamna, JP Leroy, Benoît Thirion, and Stéfan J. Darmoni. 2004. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Information and Libraries Journal*.
- Julien Grosjean, Tayeb Merabti, Nicolas Griffon, Badisse Dahamna, and Stéfan Jacques Darmoni. 2011. Multi-terminology cross-lingual model to create the European Health Terminology/Ontology Portal. In *Proc. 9th international conference on Terminology and Artificial Intelligence*, pages 119–122.
- Kristina M. Hettne, Erik M. van Mulligen, Martijn J. Schuemie, Bob J. A. Schijvenaars, and Jan A. Kors. 2010. Rewriting and suppressing UMLS terms for improved biomedical term identification. *J. Biomedical Semantics*, 1:5.
- François Maniez. 2009. L'adjectif dénominal en langue de spécialité : étude du domaine de la médecine. *Revue Française de Linguistique Appliquée*, XIV-2(La terminologie : orientations actuelles (coord. J. Humbley)):117–30.
- Tayeb Merabti, Philippe Massari, Michel Joubert, Eric Sadou, Thierry Lecroq, Hocine Abdoune, Jean-Marie Rodrigues, and Stéfan Jacques Darmoni. 2010. An automated approach to map a French terminology to UMLS. *Stud Health Technol Inform*, 1040–4(Pt 2):5.
- Tayeb Merabti, Lina F Soualmia, Julien Grosjean, O Palombi, JM Müller, and Stéfan Jacques Darmoni. 2011. Translating the Foundational Model of Anatomy into French using knowledge-based and lexical methods. *BMC Med Inform Decis Mak*, 11:65, Oct 26.
- Fiammetta Namer and Robert Baud. 2007. Defining and relating biomedical terms: towards a cross-language morphosemantics-based system. *Int J Med Inform*, 76(2-3):226–33, Feb-Mar.
- Aurélien Névéol and Sylwia Ozdowska. 2005. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In Suzanne Pinson and Nicole Vincent, editors, *EGC*, volume RNTI-E-3 of *Revue des Nouvelles Technologies de l'Information*, pages 655–666. Cépaduès-Éditions.
- Aurélien Névéol and Sylwia Ozdowska. 2006. Terminologie médicale bilingue anglais/français: usages clinique et législatif. *Glottopol*, (8):5–21, Jul.
- Aurélien Névéol, Magaly Douyère, Alexandra Rogozan, and Stéfan Jacques Darmoni. 2004. Construction de ressources terminologiques en santé pour un système d'indexation automatique. In *Proc. Journées IN-TEX/NOOJ*, Tours, France.
- Aurélien Névéol, Dietrich Rebholz-Schuhmann, and Pierre Zweigenbaum. 2013. An overview of terminological resources available for French bioNLP. In *Working notes of CLEF 2013*, Valencia, Spain.
- Sylwia Ozdowska, Aurélien Névéol, and Benoît Thirion. 2005. Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. In *Proc. 6th international conference on Terminology and Artificial Intelligence*, pages 83–84.
- Dietrich Rebholz-Schuhmann, Simon Clemenatide, Fabio Rinaldi, Senay Kafkas, Erik M. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and Jan A. Kors. 2013. Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In *CLEF 2013 Evaluation Labs and Workshop Online Working Notes*, pages 1–10.
- Stephen Tze-Inn Wu, Hongfang Liu, Dingcheng Li, Cui Tao, Mark A. Musen, Christopher G. Chute, and Nigam H. Shah. 2012. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *JAMIA*, 19(e1).
- Antonio Jimeno Yepes, Elise Prieur-Gaston, and Aurélien Névéol. 2013. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 30(14):146, Apr.
- Pierre Zweigenbaum, Robert H. Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère, and Stéfan Darmoni. 2005. A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2–4):119–124, March.