Towards the enrichment of terminological resources by scientific corpora analysis

Izabella Thomas, Iana Atanassova

Research Centre in Linguistics and in Natural Language Processing Lucien Tesnière,
University of Franche-Comté, Besançon 25030, France
E-mail: izabella.thomas@univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Abstract

The research presented in this paper explores the possibility of enriching terminological databases through the analysis of recent scientific publications. Our main concern is to evaluate how useful automatic term extraction can be to a human expert. To carry out our experiment, we constructed two corpora of recent scientific papers in two different sub-domains of the bio-medical sciences. Then we proceeded with three steps: automatic term extraction and ranking from a set of corpora of scientific papers; evaluation of the overlap of the candidate terms (CTs) extracted from the corpora and those present in the multidisciplinary terminology portal TermSciences; and evaluation by domain experts of the three sets of the top 200 CTs extracted from the different corpora. To extract terms we used the Sensunique Platform, a web based platform for building terminological resources. Our results show that only about 10% of the extracted CTs are present in the TermSciences resource, which means that many of the extracted CTs, if validated, could potentially be used to enrich the terminological database. Furthermore, the expert evaluation of the top 200 terms for each sub-corpus shows clearly that about 75% of these CTs are correct terms in the respective domains. This validates our ranking algorithm.

Keywords: terminology; term acquisition; term extraction; term recognition; scientific papers

1. Introduction

The research presented in this paper aims to explore the possibility of enriching terminological databases through the analysis of recent scientific publications. The analysis is intended to be representative of a typical situation of a terminologist at work; therefore, it is constrained by the size of the corpora and the number of candidate terms (CTs) to be managed by an analyst. One can imagine two applicative scenarios: enriching an existing resource or building a new terminological resource from scratch, as can be the case for some institutions. Our main concern is to evaluate the usefulness of automatic term extraction for human experts, i.e. the relevance of automatically constructed lists of CTs compared to the given terminological resource. More precisely, we investigate the improvement of the strategy of filtering of CTs proposed by automatic term extractors in order to organize better the work of domain experts by ordering the list of CTs according to their termhood probability.

An interest in automatic term acquisition from corpora has been developing since the

1990s (Jacquemin & Bourigault, 2003). The task consists of the automatic recognition and extraction of terminological units from different domain-specific text collections. Resulting CTs can be used in more complex applications such as Information Extraction and Retrieval, ontology construction, document indexing etc. Building and enriching domain-specific vocabularies by the analysis of corpora constitutes one of the major applications in this domain. Its objective is to help domain experts find the best term candidates from corpora, taking into consideration the type of resource to be constructed (Bourigault & Jacquemin, 2000; Bourigault et al., 2004). Since the 1990s, numerous automatic tools, mostly term extractors, have been developed based essentially on two types of approaches: statistic or linguistic, or a hybrid of these two methods[1]. Some of these tools have been developed, or can be used, for the French language, for example ANA (Enguehard & Pantera, 1995), Acabit (Daille, 1995), Lexter (Bourigault, 1993), TermoStat (Drouin, 2002), YaTeA (Aubin et al., 2006). The term extractors are considered mature technology nowadays (Cerbah et al., 2006), but this affirmation depends on the objective of the terminology acquisition: Information Retrieval or terminological mono- or multilingual resource building, requires higher quality results. In this context, the main problems concerning term extractors are the distinction between terms and non-terms, the quantity of noise in the results and the omission of relevant terms (silence). To improve the quality of the results, the task of term extraction is completed by CT scoring and ranking with the aim of classifying the extracted CTs according to their termhood probability, i.e. an evaluation of how likely it is that a particular CT is a term.

Scientific papers are used to construct domain specific corpora, sometimes along with other types of texts, such as technical documents, instruction manuals, web pages, sometimes as the only sort of documents included in the corpus (for example Kim et al., 2003; QasemiZadeh, 2014). Often, scientific corpora are used to study the inter-disciplinary scientific language or the structure of scientific discourse (Bertin et al., 2015). For the terminological purpose, the construction of the corpus depends generally on the objective of the terminological task and varies in several parameters, among which: domain and degree of specialization, reliability of sources, type of sources, and type of resources to be constructed (Cabré, 2007). We choose scientific publications to construct our corpus because they are considered good sources of terminology, and they reflect the up-to-date state of scientific terminology. We work with peer-reviewed open access journals, to guarantee the quality and validity of the text as well as its accessibility. By comparing the specialist vocabularies that are actually used in texts with existing terminological dictionaries, we can identify novel terms that are commonly used among specialists but have not yet appeared in the online terminological databases.

The originality of our work lies in the choice to investigate the specific, human expert-oriented terminological task. First, we query relatively small corpora. Even if

[1] For a synthesis of the methods see for example Cabré et al. (2001) and Drouin (2002).

nowadays the tendency is to use large corpora, we are interested in small text collections (about 20,000 words).The reason for this is that an expert has to build a new corpus for any new terminological project and this is not a trivial task. The small size of the corpora requires an accurate estimation of their degree of specialization: they should not concern too large a domain, but rather pertain to specific sub-domains. Even if the concepts of domain and sub-domain are rather naive and not formally defined, they are useful considerations for terminologists (Kageura, 1999). The other problem with large corpora is the number of CTs proposed by automatic extractors. For example, for the corpus of European patents concerning pharmacology, which comprises 2,500,000 words, 303,648 CTs were proposed (Mondary et al., 2013). Any new term added to a terminological database should be necessarily validated by a human expert. It is hardly imaginable (and not necessary) to humanly manage hundreds of thousands of CTs extracted from large text collections in a specific domain. Therefore, automatic strategies of filtering are necessary.

Our previous experience with a public French institution (Etablissement Français du Sang [National Blood Bank Organization], Bourgogne/Franche-Comté, France) revealed that some organizations do not hold large text collections (Plaisantin Alecu et al., 2012). This is confirmed by Drouin (2002), who used corpora, of sizes comparable to ours provided by a private company and described as representative of their terminological work, to test his term extractor. The disadvantages of using small corpora could be the lower efficiency of statistical measures and frequencies in automatic extraction of CTs, which could influence the quality of the extracted CTs.

We investigate the overlap of the CT sets extracted from scientific corpora with existing terminological databases, in particular with the objective of identifying novel terms for the enrichment of these resources. It is commonly admitted that there is a gap between the terminology used in texts and that used in existing terminological resources. This can be explained by the fact that terminological activity has been defined by what is called the general theory of terminology, established by Wüster and the Vienna Circle. This theory prescribes the onomasiological top-down approach to terminology: from concept to term. Therefore, the real usage of terms in context has been neglected in the process of establishing terminological dictionaries.

The overlap between terminological resources and specialized vocabularies extracted from corpora can serve different objectives; for example, evaluating the results of the term extractors. Other studies evaluate the relationship between a corpus and a terminological resource in terms of 'lexical coverage', a sort of adequacy between a corpus and a resource in order to match the most relevant resource to a given corpus (Ninova et al., 2005). Our approach is slightly different: for a given corpus and a given resource, we want to propose the most relevant terms from the texts that do not exist in the resource.

2. Methods

To extract terms from the corpora, we use three previously mentioned term extractors that are part of the Sensunique Platform[2] (Thomas et al., 2014): YaTeA (Aubin et al., 2006), Termostat (Drouin, 2002) and Acabit (Daille, 1995). The Sensunique Platform compiles the results proposed by each extractor into a unique list of CTs. The Platform is also linked to web services from an external resource: TermSciences[3], a multidisciplinary terminology portal developed by CNRS-INIST (France). This allows us to check automatically which of the extracted CTs exist in this resource.

In the Platform, the termhood probability score is obtained by a weight assignment algorithm which takes into account two features: the number of extractors that propose the same term (which we call 'multi-extraction' and which is a sort of a 'voting system' for extractors) and whether or not a CT is present in the TermSciences (see more details in section 2.2). We hypothesize that the weighted sum of these features can provide an efficient ranking criterion for the extracted CTs in terms of their termhood probability.

This methodology has already been used for the task of establishing the lexicon of a Controlled Language (Thomas et al., 2015): the Sensunique Platform was developed towards this particular objective. One of the aims of the current research is to verify its suitability to more classical terminological tasks. It is important to know that the platform is analyst-oriented, i.e. it includes a CT management interface with numerous functionalities facilitating the analysis and validation of the extracted CTs (visualization of CTs in their corpus of origin, search and filters of the list of CTs, advanced concordancer for searching in the corpus of origin etc.).

2.1 Protocol

Our main study questions are a) whether scientific papers can be used to enrich the existing terminological databases, and b) how the ranking of automatically acquired lists of CTs could facilitate the task of term validation for a human analyst. More precisely, we want to estimate how many of the best ranked CTs will be validated as terms by a human expert. To answer these questions, we proceed with three steps:

1) automatic term extraction and ranking from a set of corpora of scientific papers using Sensunique Platform;

2) evaluation of the overlap of the CTs extracted from the corpora and those present in the TermSciences resource;

[2] Station Sensunique, http://www.station-sensunique.fr/
[3] TermSciences, http://www.termsciences.fr/

3) evaluation of the top 200 CTs proposed by the platform for different corpora by domain experts.

To complete this research we also evaluate how the variability of corpora influences the automatic extraction results. Some additional results (performance of each extractor, distribution of termhood probability scores) are provided to facilitate discussion of the relevance of the features that are used to rank CTs.

2.2 Corpora and resources

To carry out our experiment, we constructed two corpora in two different sub-domains of the bio-medical sciences: *Mesenchymal stem cells* (C1) and *Vaccination* (C2). Each corpus consists of recent scientific papers taken from the chosen thematic issues (respectively 2011 and 2007) of the French specialized online medical revue Médecine/Sciences[4]. This journal is peer-reviewed and available in open access. The fact that the issues are thematic guarantees the homogeneity of the corpora. All the articles are written in French.

Each of the two initial corpora was used to obtain three different sub-corpora in the following way: for each sub-corpus one third of the papers were replaced by other papers from the same sub-domain. As a result, each pair of sub-corpora contains two thirds of common papers and one third of papers which are specific to each sub-corpus. This allows us to study the stability of the extracted CT sets with respect to variations in the corpus.

All the sub-corpora have similar sizes. The number of words in each of the six resulting sub-corpora is given in the Table 1.

Corpus	**C1 Mesenchymal stem cells**			**C2 Vaccination**		
Sub-corpus	C1a	C1b	C1c	C2a	C2b	C2c
Total number of words	17,213	17,839	17,266	21,042	21,244	21,075

Table 1: Corpus size

TermSciences is a multi-lingual and multi-purpose terminological database assembling vocabularies produced by major French research institutions (Khayari et al., 2006). Currently, it contains 650,000 terms related to 190,000 concepts. TermSciences includes three biomedical terminology resources: the French translation by the Institut National de la Santé et de la Recherche Médicale (INSERM) of the MeSH thesaurus from the US National Library of Medicine, the public health thesaurus of the Banque de Données de Santé Publique (BDSP) and the dictionary of human and mammal

[4] http://www.medecinesciences.org

reproduction biotechnology of the Institut National de la Recherche Agronomique (INRA). It is difficult to know the number of terms that each of these resources contains, since such detailed information is not available on the website of TermSciences. According to the INSERM website[5], the French version of MesH 2014 contains 83,399 terms distributed into 16 themes. The public health thesaurus of the Banque de Données de Santé Publique (BDSP) version 4 contains 12,825 terms[6] and the paper version of the dictionary of human and mammal reproduction biotechnology of the Institut National de la Recherche Agronomique (INRA) contains over 200 terms (Bouroche-Lacomb, 2011).

The choice of the TermSciences terminological database was motivated by several factors: it has a large coverage of different subjects in bio-medicine, it combines several other terminology resources and it is the biggest multi-domain resource in France. For these reasons, we expect that terms from the two specific sub-domains of our corpora, *Mesenchymal stem cells* and *Vaccination*, are present in the TermSciences database.

2.3 Termhood probability scoring

Terms extracted from each corpus were ranked using the same weight assignment algorithm. For the needs of our experimentation, we used the following two criteria:

1. the number of extractors proposing a CT: the highest score is attributed to the CTs extracted simultaneously by the three extractors, then to those extracted by two of them, and finally to those extracted by only one extractor; this procedure, called multi-extraction (Plaisantin Alecu et al., 2012), has proved to give better results than using only one term extractor (21% higher recall and 9% higher precision values compared to the use of only one extractor). The results of the multi-extraction (on much bigger corpora and with a larger number of extractors) are also judged relevant by Mondary et al. (2013).

2. the presence of a CT in the external resource (TermSciences): the Platform verifies if a CT is already present in TermSciences; for the composed CTs, three types of attestations are looked for (with decreasing score attributed): a) the whole composed CT, b) its head **and** modifier separately, i.e. occurring in two different entries in TermSciences c) its head **or** modifier separately, i.e. either the head or the modifier occurring in TermSciences. For example, for the CT *cellules souches (stem cells)*, if the whole CT is not present in TermSciences, the Platform will look for its head (*cellules*) and/or its modifier (*souches*) separately. This procedure is motivated by the hypothesis that a composed CT containing an already attested terminological element is more likely to be a term than a CT without any terminological constituent.

[5] Accessed at: http://mesh.inserm.fr/mesh/presentation.htm (20/05/2015).
[6] Accessed at: http://asp.bdsp.ehesp.fr/Thesaurus (20/05/2015).

The combination of these different criteria results in a termhood probability score, ranked as shown in Table 2. The best termhood probability score (rank 1) is obtained by the CTs proposed simultaneously by three extractors and attested as a whole term in TermSciences. The second best score (rank 2) is given to the CTs proposed by two extractors and attested in TermSciences etc. The lowest termhood probability score (rank 12) is attributed to the CTs proposed by only one extractor without any attestation in TermSciences.

TERMHOOD PROBABILITY RANK	CRITERIA					
	Number of extractors			Attestation in TermSciences		
	1	2	3	whole CT	head and modifier	head or modifier
1			x	x		
2		x		x		
3	x			x		
4			x		x	
5			x			x
6			x			
7		x			x	
8		x				x
9	x				x	
10		x				
11	x					x
12	x					

Table 2: Termhood probability score

2.4 Evaluation

To evaluate the quality of the extracted CTs for each sub-corpus we proceeded as follows. We considered the terms which are present in TermSciences as valid terms and therefore we did not need to evaluate them by human experts. We can directly observe the number of these terms for each sub-corpus. For the rest of the terms, which have been extracted by the Sensunique Platform but are not present as a whole term in TermSciences (and therefore have termhood probability ranks below 3), we considered the top 200 terms. Two highly qualified human experts in the domain (professors of immunology) were consulted for the evaluation. Each expert was presented with a list of extracted terms and asked whether the CT corresponds to a term in the domain. The possible answers were: *yes, no* and *possibly* (for the cases that need deeper analysis or additional information).

Additionally, we measured the overlaps between the sets of CTs extracted from each sub-corpus. This gives us an indication of the stability of the extracted lists of CTs depending on modifications of the corpus within the same domain.

3. Results and Discussion

3.1 General results

Tables 3 and 4 present the general results of the analysis of each sub-corpus in terms of the number of CTs proposed per extractor and the number of CTs attested in TermSciences (any type of attestation).

	C1a	% total CTs extracted	C1b	% total CTs extracted	C1c	% total CTs extracted
Total words	17,213		17,839		17,266	
Total CTs extracted	5,173		5,072		5,242	
YaTeA	3,390	65.53%	3,379	66.62%	3,434	65.51%
Acabit	2,204	42.61%	2,146	42.31%	2,261	43.13%
TermoStat	1,489	28.78%	1,445	28.49%	1,481	28.25%
Total CTs present in TermSciences	*4,022*	*77.75%*	*3,935*	*77.58%*	*4,001*	*76.33%*

Table 3: General results for C1

	C2a	% total CTs extracted	C2b	% total CTs extracted	C2c	% total CTs extracted
Total words	21,042		21,244		21,075	
Total CTs extracted	5,894		5,655		5,586	
YaTeA	3,784	64.20%	3,592	63.52%	3,675	65.79%
Acabit	2,586	43.88%	2,516	44.49%	2,370	42.43%
TermoStat	1,583	26.86%	1,458	25.78%	1,535	27.48%
Total CTs present in TermSciences	*4,365*	*74.06%*	*4,215*	*74.54%*	*4,100*	*73.40%*

Table 4: General results for C2

The sum of the CTs extracted by the extractors is not equal to 100% of all the CTs extracted, because some CTs are extracted by several extractors; in these statistics they are counted separately for each extractor.

In general, the number of CTs extracted from each sub-corpus remains relatively stable, which means that this number varies little with small changes of the papers in the corpus. The percentage of CTs proposed by each extractor is also stable across the sub-corpora and moreover across the different corpora. YaTeA is the most prolific term extractor: it extracts between 63.52% and 66.62% of all extracted CTs; the results of TermoStat vary between 25.78% and 28.78% of all extracted CTs.

The number of the CTs present in TermSciences is stable across the sub-corpora and seems rather high (more than 73% for each sub-corpus). However, this result is to be handled with care, since all types of attestations are taken into consideration, even if only a part of a CT is found. Consequently, not all of the CTs attested will be finally validated as terms.

3.2 Distribution of termhood probability score and ratio of CTs attested in TermSciences

Tables 5 and 6 present for each corpus the ratio of the CTs extracted per specific termhood probability (TP) rank.

TP rank	C1a	% total CTs extracted	C1b	% total CTs extracted	C1c	% total CTs extracted
1	54	1.04%	52	1.03%	54	1.03%
2	165	3.19%	141	2.78%	153	2.92%
3	320	6.19%	308	6.07%	295	5.63%
Total of CTs present in TermSciences as terms	*539*	*10.42%*	*501*	*9.88%*	*502*	*9.58%*
4	105	2.03%	99	1.95%	108	2.06%
5	243	4.70%	232	4.57%	226	4.31%
6	12	0.23%	11	0.22%	12	0.23%
7	114	2.20%	124	2.44%	116	2.21%
8	719	13.90%	747	14.73%	775	14.78%
9	152	2.94%	172	3.39%	154	2.94%
10	84	1.62%	98	1.93%	90	1.72%
11	2,150	41.56%	2,060	40.62%	2,120	40.44%
12	1,055	20.39%	1,028	20.27%	1,139	21.73%
Total CTs extracted	***5,173***	***100.00%***	***5,072***	***100.00%***	***5,242***	***100.00%***

Table 5: Detailed results for C1 ratio of CTs per TP

TP rank	C2a	% total CTs extracted	C2b	% total CTs extracted	C2c	% total CTs extracted
1	55	0.93%	44	0.78%	57	1.02%
2	140	2.38%	147	2.60%	143	2.56%
3	306	5.19%	313	5.53%	309	5.53%
Total CTs present in TermSciences as terms	*501*	*8.50%*	*504*	*8.91%*	*509*	*9.11%*
4	111	1.88%	108	1.91%	118	2.11%
5	257	4.36%	230	4.07%	246	4.40%
6	13	0.22%	10	0.18%	15	0.27%
7	124	2.10%	118	2.09%	117	2.09%
8	803	13.62%	761	13.46%	745	13.34%
9	191	3.24%	186	3.29%	169	3.03%
10	120	2.04%	101	1.79%	117	2.09%
11	2,378	40.35%	2,308	40.81%	2,196	39.31%
12	1,396	23.69%	1,329	23.50%	1,354	24.24%
Total CTs extracted	***5,894***	***100.00%***	***5,655***	***100.00%***	***5,586***	***100.00%***

Table 6: Detailed results for C2: ratio of CTs per TP

It is also worth noting that for the two corpora, over 60% of the CTs extracted have the two lowest TP scores, i.e. they are rank 11 (extracted by one extractor and having a head or a modifier attested in TermSciences) and rank 12 (extracted by one extractor). This means that for the majority of CTs there is no agreement between different extractors as to what should be considered a term. To exemplify this fact, Table 7 presents the number of CTs extracted by two or three extractors and the number of CTs extracted by only one extractor, for C1.

Corpus	**C1a**		**C1b**		**C1c**	
Extractors	**Number of CTs**	**% total CTs extracted**	**Number of CTs**	**% total CTs extracted**	**Number of CTs**	**%total CTs extracted**
Acabit and YaTeA and TermoStat	414	8.00%	394	7.77%	400	7.63%
Acabit and YaTeA	392	7.58%	410	8.08%	426	8.13%
Acabit and TermoStat	95	1.84%	111	2.19%	116	2.21%
YaTeA and TermoStat	595	11.50%	589	11.61%	592	11.29%
Acabit	1,303	25.19%	1,231	24.27%	1,319	25.16%
YaTeA	1,989	38.45%	1,986	39.16%	2,016	38.46%
TermoStat	385	7.44%	351	6.92%	373	7.12%
Total CTs extracted	***5,173***	***100.00%***	***5,072***	***100.00%***	***5,242***	***100.00%***

Table 7: Multi-extraction for C1

The fact that the majority of CTs is extracted by only one extractor can be explained by the differences in the methods used by each extractor. Consequently, the number of CTs proposed by each extractor is different, as can be seen in Tables 3 and 4. Nevertheless, we make the hypothesis that being proposed by several extractors is a good indicator for a CT to be a term (see section 3.4 Expert evaluation).

The total of CTs attested as terms in TermSciences (ranks 1, 2 and 3) varies by 0.84% for C1 (from 9.58% to 10.42%, Table 5). This ratio is similar for C2 (from 9.66% to 9.90%, Table 6). We can therefore assume that the average ratio of attested terms in different corpora is around 9.50% of all extracted CTs.

3.3 Analysis of the performance of the extractors

To obtain a first evaluation of the performance of extractors, we tested the results against the terms present in the terminological database, i.e. the CTs attested in TermSciences as whole terms and extracted at least by one extractor. For each sub-corpus and each extractor, we calculated the precision (P) relative to the TermSciences terminological database, i.e., the ratio of the extracted CTs and attested in TermSciences as whole terms divided by the total number of the extracted CTs.

Tables 8 and 9 present the results of this evaluation for the two corpora. The first column (T) gives the number of CTs attested as a whole term in TermSciences for each extractor.

Extractor	C1a sub-corpus		C1b sub-corpus		C1c sub-corpus	
	T	**P**	**T**	**P**	**T**	**P**
Acabit	128	5.81%	118	5.50%	123	5.44%
YaTeA	466	13.75%	430	12.73%	429	12.49%
TermoStat	218	14.64%	198	13.70%	211	14.25%

Table 8: Evaluation of the extractors for C1

Extractor	C2a sub-corpus		C2b sub-corpus		C2c sub-corpus	
	T	**P**	**T**	**P**	**T**	**P**
Acabit	129	4.99%	130	5.17%	132	5.57%
YaTeA	444	11.73%	443	12.33%	451	12.27%
TermoStat	178	11.24%	166	11.39%	183	11.92%

Table 9: Evaluation of the extractors for C2

The results are constant between the sub-corpora and corpora. Acabit is the worst scored term extractor; its precision is significantly lower than that of the other extractors. Yatea and TermoStat receive similar precisions, but TermoStat performs slightly better for C1 and YaTeA for C2.

This first evaluation shows that each separate extractor proposes a high number of CTs, most of which are not present in the terminological database. These CTs can be potentially good term candidates to enrich the terminological database, but they have to be validated by human experts. It means that, for example, 83.25% of the CTs proposed by YaTeA (100%-13.75%, Table 8, C1a sub-corpus), namely 2840 CTs, have to be validated manually. In a previous study on similar corpora using the multi-extraction method (Plasaintin-Alecu, 2012), we demonstrated that when considering the whole set of CTs extracted by two or more extractors, the best precision is around 37%. Consequently, we can roughly estimate that about 60% would not be valid terms if we consider the entire list of extracted CTs. For this reason, it is useful to propose a ranking algorithm which assigns weights to the CTs and puts the best candidates at the top of the list. In order to validate the ranking algorithm that we propose, in the next section we present the results of the evaluation by human experts of the top 200 CTs, ranked by our algorithm (see section 3 Termhood probability scoring).

3.4 Expert evaluation

For each sub-corpus, we created a list of the top 200 best scored CTs which are not present as whole terms in TermSciences. These CTs correspond to rank 4 (proposed by three extractors and whose head **and** modifier are attested in TermSciences) and rank 5 (proposed by three extractors and whose head **or** modifier are attested in TermSciences). They were submitted to the experts for evaluation. Table 10 shows the distribution of these CTs per rank for each corpus.

TP rank	C1a	C1b	C1c	C2a	C2b	C2c
4	105	99	108	111	108	118
5	95	101	92	89	92	82
Total CTs	***200***	***200***	***200***	***200***	***200***	***200***

Table 10: Top 200 CTs for C1 et C2 not present in TermSciences as whole terms

The six different sets of 200 terms overlap, and as a result, a total of 595 unique CTs had to be evaluated by the experts: 332 unique terms in C1 and 269 unique terms in C2. To evaluate the stability of the extracted CTs depending on the choice of the papers in the corpus, we observe the overlap between the sets of CTs extracted from each pair of sub-corpuses. Table 11 presents these results.

Corpus		Number of extracted CTs	% (of the total 595)
C1:	C1a, C1b & C1c	14	2.35%
	C1a & C1b	80	13.45%
	C1a & C1c	82	13.78%
	C1b & C1c	78	13.11%
	Only C1a	24	4.03%
	Only C1b	28	4.71%
	Only C1c	26	4.37%
C1 (any sub-corpus)		332	55.80%
C2:	C2a, C2b & C2c	84	14.12%
	C2a & C2b	51	8.57%
	C2a & C2c	62	10.42%
	C2b & C2c	50	8.40%
	Only C2a	3	0.50%
	Only C2b	15	2.52%
	Only C2c	4	0.67%
C2 (any sub-corpus)		269	45.21%
C1 & C2		6	0.01%

Table 11: Overlap between the sets of extracted CTs for the top 200 of CTs extracted from each sub-corpus

We observe that in C1 there is relatively little overlap between the three sub-corpora: only 14 CTs were extracted in total, while for C2 this number is 84. This means that the papers in the C2 corpus seem to be more homogeneous and replacing one third of the corpus has a very low impact on the sets of extracted terms. For the C1 corpus, the majority of CTs are shared between two sub-corpora, and each sub-corpus contributes with around 26 CTs (from 24 to 28).

Another important observation is the number of CTs that were extracted from both C1 and C2. These terms are only six in number and we can hypothesize that this is due to the fact that the two corpora contain articles on two different subjects (*Mesenchymal stem cells* and *Vaccination*) that use different terminologies. We can therefore suppose that the majority of extracted CTs are closely related to the subjects

of the corpora. Table 12 presents the six CTs extracted from both C1 and C2.

CTs extracted from both C1 and C2 (in French)	**English translation**
cellules dendritiques	*dendritic cells*
diabète de type	*diabetes type*
efficacité clinique	*clinical effectiveness*
mécanismes régulateurs	*regulating mechanisms*
passages successifs	*succesive passages*
réponse immunitaire	*immune response*

Table 12: CTs extracted from both C1 and C2

Each CT was evaluated by one expert, who was asked whether they consider this CT as a valid term in the domain. The experts had a choice of three possible answers: *yes, no* and *possibly*. Five of the six terms from Table 12 were positively evaluated by the experts (with the answer *yes*), and the candidate term *diabète de type* was evaluated with the answer *no*. Table 13 presents the results for all sets extracted from the corpora.

Answer	**C1a**	**C1b**	**C1c**	**Total C1**	**C2a**	**C2b**	**C2c**	**Total C2**
yes	154	136	148	240	154	152	151	203
possibly	15	26	16	34	18	21	23	29
no	31	38	36	58	28	27	26	37
Total CTs	***200***	***200***	***200***	***332***	***200***	***200***	***200***	***269***

Table 13: Expert evaluation of the top 200 extracted CTs not present in TermSciences

This table shows that a large majority of extracted CTs were positively evaluated by the experts. Using these results we calculate the precision among the top 200 extracted CTs ranked by the Sensunique platform in two ways:

1. Strict evaluation: only the CTs evaluated with *yes* considered as correct;
2. Loose evaluation: CTs evaluated with either *yes* or *possibly* considered as correct.

	C1a	**C1b**	**C1c**	**Total C1**	**C2a**	**C2b**	**C2c**	**Total C2**
Strict evaluation	77.00%	68.00%	74.00%	72.29%	77.00%	76.00%	75.50%	75.46%
Loose evaluation	84.50%	81.00%	82.00%	82.53%	86.00%	86.50%	87.00%	86.25%

Table 14: Precision for the top 200 extracted CTs for each corpus

Table 14 presents the precision values for this evaluation. These results are very promising. In fact, we can see from Table 14 that for all sub-corpora the precision for the strict evaluation is above 68%, and for five out of six sub-corpora it exceeds 74% and an average of about 75% of the CTs were evaluated as correct. Furthermore, the precision is above 81% for the loose evaluation. This means that the criteria that we have considered allow us to perform ranking with little noise among the top results. At

the same time, as shown in Tables 8 and 9, the results of the three extractors have little overlap with the TermSciences database. This means that the extraction from scientific corpora is an adequate approach for the enrichment of terminological databases.

We work only with the top 200 extracted CTs which are not present in TermSciences, and thus this evaluation concerns only the criteria corresponding to ranks 4 and 5, as the CTs with higher ranks feature much further down the list. The evaluation of all ranks can be carried out but it is very expensive because of the large number of extracted CTs.

4. Conclusions

Using the multi-extraction method implemented in the Sensunique platform, we have carried out the extraction of terms working with relatively small corpora of about 20,000 words. The number of candidate terms extracted from each corpus is very large, about 6,000 (single word terms or multiword terms) which makes the results difficult to use by the experts. The reason for this high number of CTs is that the multi-extraction method combines the results of three different extractors. In this context it is important to consider ranking algorithms that order the lists of extracted CTs by relevance. In our study we considered two major ranking criteria based on an external terminological resource and on votes by several extractors.

The main objective of our study was to propose new strategies for the enrichment of existing terminological resources using scientific corpora. In general, language evolves quickly and there is little overlap between terms found in terminological databases and terms actually used in scientific writing. For example, our results (Tables 5 and 6) show that only about 10% of the extracted CTs are present in the TermSciences resource, which means that many of the extracted CTs, if validated, could potentially be used to enrich this terminological database. Furthermore, the expert evaluation of the top 200 terms for each sub-corpus shows clearly that the majority of these CTs are correct terms in the respective domains. We can therefore conclude that scientific corpora constitute suitable sources for terminological extractions.

In general, the quality of the results of extractors reduces for smaller sized corpora. For example, working with small corpora we have previously found (Plaisantin Alecu et al., 2012) that the best extractor, YaTeA, reaches 58% of recall and the best precision value for a single extractor, Termostat, to be 28%. For this reason, it is interesting to consider the multi-extraction method as it proposes more relevant results in terms of recall. The disadvantage of the multi-extraction, i.e. a larger number of CTs compared to the results of only one extractor, can be compensated using ranking criteria for the extracted CTs. The ranking algorithm that we propose allows us to obtain high precision among the top results, i.e. 75% of the best ranked CTs can be used to enrich the terminological database. Consequently, we have shown that we can produce good results, even if we work with relatively small corpora.

5. Acknowledgements

The authors thank professors Estelle Seillès (Etablissement Français du Sang (National Blood Bank Organization), Bourgogne/Franche-Comté) and Dominique A. Vuitton (Research Federation « Cell and Tissue Biology and Engineering » FED 4234, University of Franche-Comté) for their expert assistance.

6. References

Aubin, S., & Hamon, T. (2006). Improving Term Extraction with Terminological Resources.In *Advances in Natural Language Processing*, Springer, pp. 380–387.

Bertin, M., Atanassova, I., Larivière, V. & Gingras, Y. (2015). The Invariant Distribution of References in Scientific Papers. *Journal of the Association for Information Science and Technology (JASIST)*, doi: 10.1002/asi.23367.

Bourigault, D., Aussenac-Gilles, N. % Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes. Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle* 18 (1), pp. 87–110.

Bourigault, D. & Jacquemin, C. (2000). Construction de ressources terminologiques. In J.-M. Pierrel (ed.) *Industrie des langues*. Hermès, Paris, 2000, pp. 215-233.

Bourigault, D. (1994). Lexter: Un Logiciel d'EXtraction de TERminologie: Application à l'acquisition des connaissances à partir de textes." EHESS, Paris.

Bouroche-Lacomb, A. (2001), Biotechnologies de la reproduction chez les mammifères et l'homme : vocabulaire français-anglais, INRA Editions.

Cabré, M.T. (2007). Constituer un corpus de textes de spécialité. *Cahier du CIEL*, pp. 37-56.

Cabré, M.T., Estopà, R. & Vivaldi, J. (2001). Automatic term detection: A review of current systems. In D. Bourigault D., C. Jacquemin & M.-C. L'Homme (eds.) *Recent Advances in Computational Terminology*. Amsterdam / Philadelphie, John Benjamins, pp. 53–87.

Cerbah, F., & Daille B.(2006). Une Architecture de Services Pour Mieux Spécialiser Les Processus D'acquisition Terminologique. *Traitement Automatique Des Langues (TAL)* 47(3), pp. 39–61.

Enguehard, Ch., & Pantera L. (1995). Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2 (1), pp. 27–32.

Daille, B. (1995). Repérage et Extraction de Terminologie Par Une Approche Mixte Statistique et Linguistique. *TAL. Traitement Automatique Des Langues* 36 (1-2), pp. 101–118.

Drouin, P. (2002). *Acquisition Automatique Termes: L'utilisation Des Pivots Lexicaux Spécialisés*. PhD thesis, Université de Montréal, 2002.

Drouin, P. (2003). Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology* 9(1), pp. 99–115.

Ibekwe-Sanjuan, F. (2006). Repérage et annotation d'indices de nouveautés dans les écrits scientifiques. In *Indice, index, indexation. Actes du colloque international, Université Lille-3*, pp. 1-11.

Jacquemin, Ch. & Bourigault, D. (2003). Term extraction and Automatic Indexing. In Mitkov R. (ed.) *Oxford Handbook of Computational Linguistics*. Oxford University Press.

Kageura, K. (1999). Theories of terminology: A quest for a framework for the study of term formation ». *Terminology 5(1)*, pp. 21-40.

Khayari, M., Schneider, S., Kramer, I., & Romary, L. (2006). Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative. *arXiv preprint cs/0604027*.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1), pp. 180–182.

Mondary, T, Nazarenko, A., Zargayouna, H., & Barreaux, S. (2013). Aide À L'enrichissement D'un Référentiel Terminologique: Propositions et Expérimentations. In *20e Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN'2013)*, pp. 779–786.

Ninova, G., Nazarenko A., Hamon T. & Szulman S. (2005). Comment Mesurer La Couverture D'une Ressource Terminologique Pour Un Corpus. *TALN 2005*, 2005.

QasemiZadeh, B. & Handschuh, S. (2014). The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *COLING 2014: 4th International Workshop on Computational Terminology*.

Plaisantin Alecu, B., Thomas, I., & Renahy, J. (2012). La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques, Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, ATALA/AFCP, pp. 511-518.

Thomas I., Plaisantin Alecu B., Germain B. & Betbeder M.-L. (2014). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources. In A. Abel et al. (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: EURAC research, Volume: II*, pp. 707-726.

Thomas, I., Laroche, L., Plaisantin-Alecu, B., Betbeder, M.-L., Seillès, E., Renahy, J., Blagoskonov, O. & Vuitton, D.-A. (2015). Computerization of a 'controlled language' to write medical standard operating procedures (SOPs). In *Proceedings of Conference on Health and Social Care Information Systems and Technologies, HCist 2015 October 7-9, 2015*, Procedia Computer Science, Elsevier (to appear).