Automatic extraction of specialized verbal units

A comparative study on Arabic, English and French

Nizar Ghazzawi,[1] Benoît Robichaud,[1] Patrick Drouin[1] and Fatiha Sadat[2]
[1]Université de Montréal / [2]Université du Québec à Montréal

This paper presents a methodology for the automatic extraction of specialized Arabic, English and French verbs of the field of computing. Since nominal terms are predominant in terminology, our interest is to explore to what extent verbs can also be part of a terminological analysis. Hence, our objective is to verify how an existing extraction tool will perform when it comes to specialized verbs in a given specialized domain. Furthermore, we want to investigate any particularities that a language can represent regarding verbal terms from the automatic extraction perspective. Our choice to operate on three different languages reflects our desire to see whether the chosen tool can perform better on one language compared to the others. Moreover, given that Arabic is a morphologically rich and complex language, we consider investigating the results yielded by the extraction tool. The extractor used for our experiment is TermoStat (Drouin 2003). So far, our results show that the extraction of verbs of computing represents certain differences in terms of quality and particularities of these units in this specialized domain between the languages under question.

Keywords: specialized verbs, verbal terminological units, Arabic, French, English, terms extraction, terminology, corpus linguistics

1. Introduction

Since automatic extraction of terms has become one of the basic and essential tasks in any terminological and terminographical project, two types of lexical units, among others, are considered, multi-word terms (MWT) and single-word terms (SWT). In spite of the fact that MWT are the most sought-after units, we are interested in SWT. Our interest in these units stems from the fact that they can be particularly problematic, especially when it comes to automatic extraction.

Terminology 23:2 (2017), **207–237**. DOI 10.1075/term.00002.gha
ISSN 0929–9971 / E-ISSN 1569–9994

Moreover, less attention is paid to such units in specialized dictionaries, as MWTs are predominant. Within the category of SWT, we are particularly interested in verbs. Verbs received little attention in terminology, because terms are mainly considered to be nouns (Guilbert 1973; Rey 1979; Sager 1990), or if they are considered as terms, that is because they are conceived as derived from noun terms.

Automatic extraction of verbal terminological units (VTU) constitutes a challenge for different reasons. In a domain such as computing, certain terms are being exposed to a migration from a specialized language to the general one. This is what is meant by the phenomenon of determinologization (Meyer 2000). So, it might be a difficult task to recognize such units with an automatic extraction system. Furthermore, VTUs (and generally any SWT of any part of speech) may be polysemous units, and they do not differ at the surface-level from other lexical units (Lemay et al. 2005). For example, verbs such as *read* in English, *lire* in French and *qara>a* (قرأ) in Arabic[1] might be tricky enough to be recognized as candidate terms by an automatic term extraction system because a difference should be made between senses in which a person can read (a newspaper, for example) and senses in which a computerized program can read (*program reads data*).

Therefore, in our present research, we discuss the steps leading to automatically extracting specialized verbs from specialized multilingual corpora, and investigate any particularities which a language might have. The steps constituting our methodology are based on an advanced multilingual term extractor called TermoStat (Drouin 2003). This term extractor is based on corpus comparison techniques. It exposes an analysis corpus (AC) to a reference corpus (RC) in order to discriminate lexical units with frequencies that are significantly deviated from a theoretical referential frequency found in the RC. The languages under question are Arabic, English and French and our domain of interest is computing. The extractor is designed in such a way that a user can extract candidate terms of any part of speech desired in these languages. In addition, it gives the possibility to extract MWTs or SWTs independently and it generates concordance lists for these units. The results of the extraction undergo manual filtering and terminological validation to ensure that the units obtained are domain specific. It should be pointed out that Arabic language has not yet been integrated officially to TermoStat. It is still under experiment.

The structure of this paper will be as follows. In Section 2 we present the state of the art of previous work based on comparative method. In Section 3 we describe our methodology and the way the extraction was performed. In Section 4

1. For Arabic terms, we give the transliterated form with the original one (Arabic characters) between brackets. Moreover, the linguistic variant used in our research is the MSA, Modern Standard Arabic.

we present our results, and in Section 5 we discuss our results and we give an evaluation. Finally, in Section 6 we conclude with general remarks.

2. Previous work

As mentioned in the introduction, automatic term extraction is one of the most essential tasks in modern terminography. It has become an important part of computer-aided tools (CAT) and natural language processing (NLP). Generally speaking, term extraction implies two steps: identifying candidate terms and filtering the extraction results to eliminate units that do not belong to the domain in question (Abed et al. 2013). So, since the methodology of extraction we used is based on corpus comparison, our main focus will be on major work and projects that are based on this technique, starting from the earliest ones till the most recent. Moreover, since we are concerned with SWTs, we will chiefly be interested in presenting extraction methods for this type of units.

To extract units specific to a specialized domain using the comparative method, two techniques are followed (Rayson and Garside 2000). Either a specialized corpus, or the analysis corpus (AC), is compared to a general language corpus, or the reference corpus (RC); or two or more different "equal" corpora are compared to each other. In these methods frequencies of the items are being compared between the different corpora. This implies that the more "frequently a unit appears in a corpus, the more likely it is to be significant in this corpus." (Lemay et al. 2005, 232) In other words, the use of a lexical unit in a specialized corpus gives an idea of how specific this unit is to the domain. Moreover, in this method of term extraction a distinction is made between three major classes based on the candidates being extracted: positive specificity, negative specificity and unsurprising forms. These classes have been given different labels by authors.

The first studies and researches did not have a terminological perspective; they were rather based on extraction of specific vocabulary of a specialized discourse. All started with the work of Muller (1967, 1977) who used textual statistics in his research on the theatrical works of French playwright Corneille. His statement that the vocabulary of a text is part of larger lexicon[2] (1967, cited by Monsonego 1969, 108) paved the way for other researchers to use the same method or to come up with other techniques based on his method in their endeavors to discover vocabularies of specialized discourses. Muller used a hypergeometric model that aimed at comparing specific fragments to a whole oeuvre. He introduced such

2. Our translation of "Le vocabulaire d'un texte suppose l'existence d'un lexique dont il n'est qu'un échantillon."

notions as "positive characteristic vocabulary" and "negative characteristic vocabulary" to classify subsets of the vocabulary of these fragments.

Later on, other researchers in computational linguistics developed techniques for the extraction of vocabularies, again with no terminological perspective. For example, Lafon (1980) and Lebart and Salem (1994) worked on discourse analysis using the same hypergeometric model. They isolated the vocabularies belonging to the analyzed discourse depending on the statistical deviation that the forms might represent from the theoretical model. Likewise, Ahmad et al. (1994) employed roughly the same method, but with a different technique that they called "co-efficient of weirdness" to isolate the lexical particularities from a specialized corpus. Church and Hanks (1990) worked on collocations using mutual information (MI). Since then, this comparison method has been also used in order to pinpoint corpus specific vocabularies (Scott 1997; Kilgarriff 2001). Finally, Kilgarriff (2001), using the British National Corpus (BNC), investigated the uses and distribution of male/female differences for contrasting lists of most different words gathered using the Chi-square test with those gathered using the Mann-Whitney test.

Within terminology, corpus comparison method has gained the attention of a number of researchers. For example, Nelson (2000) analyzed the business English vocabulary. His method consisted of comparing three different corpora (the Published Materials Corpus, the Business English Corpus and the British National Corpus). Rayson and Garside (2000) worked on field reports of a series of ethnographic studies from an air traffic control center. They designed the log-likelihood measure that relies on frequency profiling. Chung (2003) based her method on a ratio used as a tool depending on the comparative ranges and frequencies of word forms between a technical corpus and a comparison corpus. Based on the works of Lebart and Salem (1994), Drouin (2003) devised a term extraction technique exploiting the standard normal distribution of lexical units. Units that exhibit a non-standard distribution are used as a starting point to identify candidate terms.

Finally, other methods for terminology extraction exist for SWT recognition. These imply statistical, linguistic and hybrid techniques. For example, Fung (1998) used statistical method for extracting SWTs from a bilingual English-Chinese corpus taken from the Wall Street Journal for English and the Nikkei Financial News for Chinese. Similarly, Rapp (1999) used a statistical method to extract SWTs from a comparable journalistic English-German corpus. Déjean and Gaussier (2002) investigated a medical comparable English-German corpus. Xu et al. (2002) based their method of SWT extraction of financial management and stock market terminology on an unsupervised hybrid text-mining by adopting the TF-IDF classification method. Likewise, Abed et al. (2013) presented a hybrid method based on statistical and linguistic filters in extracting SWTs as well as MWTs from Arabic

Islamic documents. The authors used a simple POS tagger to identify MWT with given syntactic patterns and the TF-IDF to rank SWT candidates.

3. Methodology

3.1 TermoStat: A general overview

The term acquisition method in TermoStat is based on a three-stage process: pattern matching, corpora comparison and candidate term ranking and filtering.

The system uses normalized set of tags so that pattern matching can be made part of speech (POS) tagger independent. This set can be easily modified, expanded and maintained. For the current work, TreeTagger was used as the input POS tagger for English and French. The POS tags are used by the system that extracts terms based on patterns. Such patterns are described using regular expressions that allow matching both at the POS level and the lexical level. For the purpose of the current paper, our extraction was limited to verbs and single word nouns. Once the patterns are applied to corpora, the system creates an initial list of candidate terms.

In the second stage, the software relies on a comparison of the frequencies of the lexical items in two corpora: a RC (corpora of journalistic nature of 8 million words for both English and French) and an AC. The hypothesis is that the AC and the RC can be merged to form a larger virtual corpus (VC) and that the AC is then considered as a subcorpus of the VC. If it is the case, word frequencies should not vary in the subcorpus and terms should be distributed normally (from a statistical point of view). Based on the specificity test proposed by Lebart and Salem (2004), we use a statistical test to measure deviance from the standard normal distribution.

Once the corpus comparison is completed, the list of candidate terms is ranked in decreasing order of specificity and a threshold is used to filter candidates. The specificity score can be mapped to a probability and, by default, a value of 3.09 is used as a cut-off point to eliminate candidates. This threshold basically means that the odds of the high frequency observed in the AC being due to chance are less than 1/1000. Lowering the threshold to 2.33 or to 1.96 would lower the corresponding probability to 1/100 and 5/100. Even though all values are computed, the Web interface of the tool solely presents candidate terms of the highest quality.

3.2 Integrating Arabic language to TermoStat

It is well known that written Semitic languages such as Arabic pose many challenges to NLP. These range from character encoding, rich and complex morphology

(involving concatenative and templatic means, as well as cliticization of numerous function words) to relative free word order in syntax and non-compositional semantic constructs at the lexical level (Habash and Sadat 2006). The example we give here is the word *wasayaktubwnahA* (وسيكتبونها), which means "and they will write it". It is composed of three prefixes (a conjunction, markers of time and gender), the stem (*ktb*, to write) and two suffixes (the inflection and a pronoun).

Within the present study, we had to plan and test a processing chain of steps in order to feed the term extraction tool in a very standard way. In particular, TermoStat needs as input tokenized and lemmatized raw text corpora in order to compare frequencies of the lemmas that appear in an AC to the ones in a RC. This is not an easy task as there are not as many freely available tools for the Arabic language as in English or French. In particular, decent lemmatization is very difficult as written Arabic does not encode short vowels (leading to massive ambiguity of surface forms), and that lemmatization tools rely on lexicons that may include obsolete words from Classical Arabic that are no longer used in Modern Standard Arabic (leading to noisy data that jeopardizes the disambiguation process).

Our first attempt was to install and test the MADA toolkit (Habash et al. 2009). The tool works as follows: after a tokenization phase, it "diacritizes" all words (i.e. it reintroduces the short vowels in surface forms) generating for each all the alternatives corresponding to the different possible morphological analysis. In a second step, based on statistical calculi that are made on the context where the words are found, a "best" morphological analysis and lemmatization is chosen among the different hypotheses. Soon we realized that this tool was not adequate as a great proportion of words were not assigned the appropriate analysis or lemma in context.

As MADA is considered the best "all-in-one" freely available tool, we decided to follow another path. The strategy consisted in splitting the job between two different tools: on the one hand, we would use the Stanford POS Tagger (Toutanova and Manning 2000; Toutanova et al. 2003) to perform the tokenization task and the selection of the parts of speech in context; and on the other hand, we would use AraComLex (Attia et al. 2011, Attia et al. 2011a and 2013) to carry out the morphological analysis and the lemmatization task. In Table 1, we give an example of an output from our Arabic corpus processed with AraComLex and Stanford. The sentence being analyzed is *how to scan files from viruses through the internet and without a software*.

Table 1. Output of AraComLex and Stanford

Form	POS	Lexeme
من/IN الفيروسات/DTNNS عن/IN طريق/NN الإنترنت/DTNN وبدون/NNP برامج/NN كيفية/NN فحص/NN الملفات/DTNNS		
كيفية	NN	كيفية
فحص	NN	HuS~_1
الملفات	NN	ملف
من	IN	من
الفيروسات	NN	فيروس
عن	IN	عن
طريق	NN	طريق
الإنترنت	NN	إنترنت
وبدون	NN	duwn_1
برامج	NN	برنامج

Following this path, one difficulty remained: in some cases, we would have to reconcile the analysis made by the different tools. As we observed, this situation happens in two circumstances. The first one is related to the fact that the AC and the RC (the RC is of a journalistic nature of 16 million words) contain words that are not known within the lexicon of AraComLex (in this case, we would have to supply them to the tool); the second is when the POS Tagger would give a wrong part of speech to certain words in contexts (in this case nothing could be done, as we do not have either the material or the time to train the tool on a new corpus).

3.3 Compiling specialized corpora

Since TermoStat is based on comparing lexical units between an AC and a RC, one has to start by compiling the AC for the three languages. We chose to compile a comparable AC in order to obtain homogenous and significant results. The documents composing the corpora come from two different subdomains of computing, namely programming and PC maintenance. For each language, we chose two programming manuals, one on Linux and the other on Python, and one PC maintenance manual. The size of the corpus for each language[3] is approximately 600,000 words (see Table 2). For English and French, documents are original texts. However, for Arabic, the corpus contains two translated documents and one original.

3. For Arabic, the total number of words exceeds 700,000 because in programming manuals there are many English words which are part of the word count.

Table 2. Size of corpora of analysis

Language	Type of corpus	Number of words	Total
Arabic	تعلم البرمجة مع بايثون 3	248,237	732,695
	الإدارة المتقدمة لجنو\لينكس	264,005	
	صيانة الحاسب	220,453	
English	Linux for dummies	199,848	551,807
	Introduction to Python for beginners	135,884	
	PC for all	216,075	
French	Le cahier de l'administration Debian	249,401	596,629
	Apprendre Python 3	226,452	
	Dépannage PC	120,776	

3.4 Pre-processing

As all the texts of our corpora were PDF files, it was necessary to convert them into plain texts (.txt format) with UTF8 encoding. To make sure that all the files had the same encoding, we had to do some adjustments. For example, with French texts, we verified that all characters had the right encoding. For Arabic, the task was more laborious. First, converting files into plain text format was not easy. With certain documents, the conversion eliminated many characters, or words got mixed. The presence of English words within the Arabic texts, led to problems related to character encoding as well as layout problems. For this, we used a web-based tool called *cloud convert* (cloudconvert.com/). This tool is freely accessible online and it can convert PDF files in a way that the return encoding is determined by the user. Another tool used is a freely available converter called PDF to Text Converter Expert (macdownload.informer.com/pdf-to-text-converter-expert/). We used these tools for French and Arabic documents.

3.5 Managing specificity

Regarding specificity, we mainly investigated units with a frequency higher than that observed in the RC, i.e. the positive specificities (SP+). The hypothesis we are making here is that lexical units with SP+ are probably domain specific. As discussed in Section 3.1, TermoStat uses a default value of 3.09. However, we considered units with a specificity threshold of 1.96 and above. Such a specificity threshold would enlarge the scope of analysis to cover more units (compared with threshold of 3.09). Moreover, this threshold was used by some researchers (Lebart and Salem 1994; Muller 1977), as stated in Drouin (2002).

4. Results

In order to manage the output delivered by TermoStat, the candidate VTUs were sorted by specificity classification, with the VTUs of higher specificity being at the top of the list (Tables 3, 4 and 5).

Table 3. Partial list of extraction results for Arabic

Frequency	Specificity	Lemma	POS	forms
5012	244.8126762	Oan~-i_1	V	لأنه
1340	178.4127348	تم	V	لتتم
2946	165.847954	مكن	V	يمكنهم
1411	150.015444	yatum-u_1	V	يتمكن
3219	147.7453048	إلى	V	إليها
3127	139.144891	كان	V	يكونون
1245	112.8225555	استخدم	V	لاستخدم
833	109.0628264	تكون	V	تكون
2261	96.79184515	قام	V	وقمت
1024	91.65969356	جاب	V	أجاب
1480	87.07648698	عمل	V	لتعمل
900	68.82400389	غير	V	لنغير
629	57.19176867	سمح	V	نسمح
1398	55.04272213	أو	V	أو
493	52.98533266	وضع	V	بالأوضاع
161	48.47620445	وازى	V	لتوز
1131	43.58597797	wahiy-a_1	V	أنه
192	42.60451345	qud~_1	V	فلقد
292	42.18859975	شكل	V	وأشكال

Table 4. Partial list of extraction results for English

Candidate	Frequency	Specificity	Orthographic variants	POS
Use	2551	93.04	use___uses___using	V
Click	387	81.26	click___clicks___clicking	V
Install	521	74.25	install___installs___installed___installing	V
Type	270	65.78	type___types___typing	V
configure	224	61.9	configure___configures___configuring	V
Access	211	57.53	access___accesses___accessing	V

Table 4. (*continued*)

Candidate	Frequency	Specificity	Orthographic variants	POS
Display	305	50.02	display___displays___displayed___displaying	V
connect	303	49.09	connect___connects___connected___connecting	V
Log	162	47.36	log___logs___logged___logging	V
File	278	45.98	file___filed	V
Delete	129	45.2	delete___deletes___deleting	V
Copy	158	42.83	copy___copies	V
Plug	119	41.48	plug___plugs___plugging	V
Specify	178	36.72	specify___specifying	V
Select	240	36.59	select___selects___selected___selecting	V
Run	584	32.82	run___runs___running	V
Boot	64	32.22	boot___boots___booted___booting	V
Choose	334	31.99	choose___chosen___choosing	V
Start	487	31.73	start___starts___started___starting	V

Table 5. Partial list of extraction results for French

Candidate	Frequency	Specificity	Orthographic variants	POS
Cliquer	559	164.16	cliquer___cliquez___cliqueront	V
Utiliser	752	89.73	utiliser___utilisera___utilisa___utiliserait___utilisaient	V
L	385	85.13	l	V
sélectionner	204	71.56	sélectionner___sélectionnera___sélectionnent	V
Tkinter	103	71.22	tkinter	V
exécuter	216	70.62	exécuter___exécutera___exécutent___exécutait	V
désactiver	104	68.54	désactiver___désactivera___désactivent	V
configurer	94	67.26	configurer___configurera___configurent___configurons	V
C	130	67.16	c	V
Text	90	66.52	text	V
X	105	65.66	x	V
I	129	64.76	i	V
N	250	64.04	n	V
contenir	313	62.43	contenir___contenait___contiendra___contiendrait___contiendraient	V

Table 5. (*continued*)

Candidate	Frequency	Specificity	Orthographic variants	POS
Page	80	60.36	page	V
Veuillez	69	55.68	veuillez	V
S	177	53.42	s	V
Activer	140	52.96	activer___activera___activons	V
modifier	277	52.61	modifier___modifiera___modifient___modifieront	V

4.1 Filtering results

After obtaining the results, it was necessary to filter certain units that would not be considered in the final analysis. Filtering the results was done manually and in a systematic way, meaning that our method was composed of two steps that we will discuss in the following sections.

4.1.1 *Tagging errors*

The first step was eliminating any form that was not relevant to the targeted units, i.e. verbs. We consider these forms as belonging to tagging errors. Within tagging errors lie the following types of units: abbreviations and acronyms, other parts of speech (nouns, adjectives and adverbs), function words and any other vague forms (forms that we could not identify). As stated above, filtering out those units was done manually. In the case of other parts of speech, the contexts played a major role in detecting the meaning of certain units (see 4.1.3), especially in cases of homography. For example, we observed that verbs such as *download*, *upload* and *command* came in two different forms, verbs and nouns. Examples of tagging errors are given in Table 6.

Table 6. Eliminated units

Type of tagging error	Arabic	English	French
Vague forms	mA_7	utils	ment
Function words	>w (أو) (or)		une (indefinite article)
Modal verbs	kAn (كان) (to be)		pouvoir (can)
Acronyms		sys	txt
Other POS	$abakap (شبكة) (network)	downloads	téléchargement (download)

4.1.2 *General language units*

The second step was filtering out units that did not have any terminological status. Those units fall within the sought-after units, except that they do not have any

specialized value. Those units belong to the general language. The identification of these units went through two steps. In step one, verbs that belonged to the general language discourse were eliminated. Those verbs are used by authors as part of their style of writing (Lorente 2007, 366). In step two, remaining verbs were analyzed, and accordingly, they were either eliminated or kept for terminological validation. In Table 7, we give examples of these units.

Table 7. Eliminating general language units

Arabic	English	French
tam~a (تم) (to fulfill)	add	indiquer (to indicate)
AEtabara (اعتبر) (to consider)	want	souhaiter (to wish)
qAla (قال) (to say)	do	determiner (to determine)

4.1.3 *Concordance list*

In order to detect anomalies in the extracted lists of verbs and to identify potential terms, we had to refer to the concordances of each unit in order to examine its immediate environment. This environment is the contexts that can be found to the right and to the left of the unit. In Figures 1, 2 and 3, we give examples of concordance lists of three verbs. These lists are generated by TermoStat itself. However, the Arabic example is not from TermoStat, since this language is not yet totally integrated into the extractor. We used WordSmith Tools (Reppen 2001) for the processing of the Arabic verbs.

As shown in the figures above, the verb in question is *to execute* for the three languages. According to the contexts, the verb *to execute* is surrounded by terms such as *operations*, *scripts*, *commands* and *orders*. For us, this constituted a clue to discover the terminological status of the verb. However, context was not the only element to consider. Other things were taken into consideration, such as the relevant meaning of the unit to the overall domain. With the help of concordances, we found out that the verb *to execute* is related to a process within the domain of computing through which a user (or a machine) undertakes a certain function in order to activate a computing component (software).

After applying these filters, we obtained the lists that were ready to be examined for the terminological validation.

4.2 Results of filtering

Once filtering the results was over, we gathered some statistics on the number of units we decided to keep and to eliminate from the lists. In the following, we present the total number extracted, units kept (candidates) and units filtered

Figure 1. Partial concordance list for the Arabic verb *naf~a*a* (نفذ) (to execute)

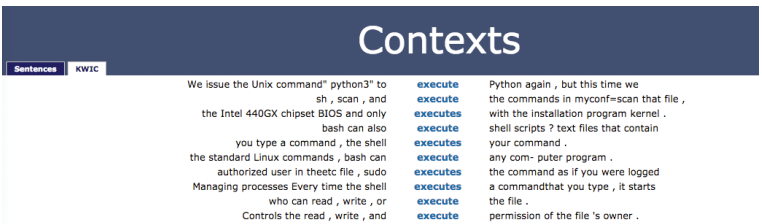

Figure 2. Partial concordance list for the English verb *to execute*

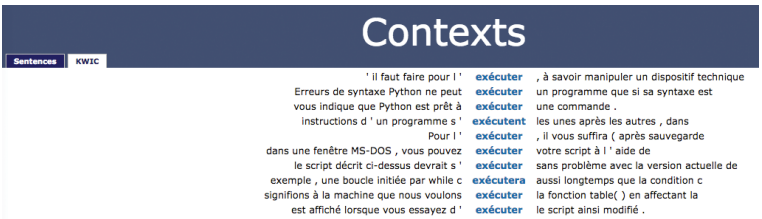

Figure 3. Partial concordance list for the French verb *exécuter*

out. For the latter, we include details on the number of general language units (GLUs) and tagging errors (vague forms, other parts of speech and abbreviations and acronyms).

For Arabic

For the 375 lemmas extracted, we obtained the following results (Table 8).

Table 8. Filtering results of Arabic

Candidates	GLU	Tagging errors	
18	52	305	Vague forms: 68
			Other POS: 237
			Abb. & Acr.: –

As shown in Table 8, the last category of tagging errors (Abb. & Acr.) is empty. We did not find any abbreviations or acronyms in the list produced by the extractor. Moreover, we observed a considerable number of units other than verbs. These units were the result of the conflict between the tagger and the morphological analyzer discussed earlier (see Section 3.2). As for vague forms, we found a number of forms that we could not identify. We believe that this was due to the fact that the POS tagger recognized certain fragments of a word as being whole units.

For English

For the 546 lemmas extracted, we obtained the following results (Table 9).

Table 9. Filtering results of English

Candidates	GLU	Tagging errors	
221	90	235	Vague forms: 89
			Other POS: 107
			Abb. & Acr.: 39

In the case of English, the number of vague forms is high. Moreover, since our corpus had texts dealing with programming, the POS tagger considered certain abbreviations and acronyms as being verbs. As for other POS, for certain units, we realized that the extractor proposed the gerund and the past participle forms as candidates. For example, the verb *to download* exists in two forms in the list: *download* and *downloading*.

For French

For the 840 lemmas extracted, we obtained the following results (Table 10).

Table 10. Filtering results of French

Candidates	GLU	Tagging errors	
211	60	579	Vague forms: 433
			Other POS: 41
			Abb. & Acr.: 102

As shown in Table 10, lots of tagging errors were found for French. Like English, we realized that the extractor proposed a good number of abbreviations and acronyms as candidates. Obviously, those were programming instructions (and other acronyms used in computing) coming originally from English. As for vague forms, like Arabic, the high number of these forms was a problem we noticed with the tagger as not being able to detect certain fragments of words. We believe as well that this might be a problem with the conversion of documents into plain text format.

4.2.1 *Specificity*

After obtaining the final results, we found it important to examine the specificity of the units we selected for validation and those we decided to eliminate. The question of specificity is particularly important, since it helps us to better understand the performance of the extractor and the relatedness of verbs to the specialized domain.

As far as eliminated units are concerned, for English, most of the units filtered out were somehow concentrated towards the bottom of the lists. Units with a specificity of 7.9 and below included several tagging errors. Things got worse when the specificity dropped lower than 5. Units with a specificity of 1.96 were mostly filtered out. For French, as indicated earlier, many tagging errors were found. We could not determine the specificity range for the eliminated units, as they existed almost everywhere in the list (several positions: top, middle, bottom). For example, the following forms were observed as having high specificities: *tkinter* (a module used in Python programming language) 71.22, *vista* (a version of Windows operating system) 50, *veuillez* (*please*) 55 and *txt* (abbreviation of *text*) 33.73. Towards the bottom of the list (below 4.1), general language units were more common. For Arabic, like French, we could not determine the specificity for those units that we decided to eliminate. However, things looked much worse, as the number of tagging errors was much higher.

As for the units that we decided to keep for terminological validation, for the most part, they were situated at the top of the list for English and French. Nonetheless, in certain cases, the specificity might drop to as low as 1.96 (meaning at the very bottom of the list). For example, French units such as *recréer* (to recreate), *renommer* (to rename) and *protéger* (to protect) were all situated at the

bottom of the list with specificities of 2.75, 2.21 and 2.19, respectively. For English, units with low specificity did not have high occurrences. For example, the verb *to reconfigure* has a specificity of 1.97 and with only one occurrence in the corpus. Another example is the verb *to reload*. This verb has a specificity of 2.44 with only 4 occurrences in the corpus.

Regarding Arabic, for units we decided to keep, the selection procedure was not at all an easy task. A meticulous verification of concordance lists was done, given the complex morphology of the language itself. As far as the specificity is concerned, candidate units had to a certain degree a fluctuating specificity. For example, at the top of the list, specificity for candidates ranged between 87.07 for a unit such as *Eamila* (عمل) (to run) to 34.87 and 18.32 for units such as *EaraDa* (عرض) (to display) and *naqara* (نقر) (to click). Units situated at the bottom of the list with low specificity had high occurrences in the corpus. For example, *DabaTa* (ضبط) (to set) (+8.02), *Ha*afa* (حفظ) (to save) (+2.95) and *xaz~ana* (خزن) (to stock) (+3.63) have occurrences of 299, 290 and 182, respectively.

4.2.2 *Certain particularities*

Since we obtained such surprising results for Arabic verbs, we were curious enough to see whether there were any verbs worth considering whose specificity fell under the threshold discussed. So, we decided to take into account all the lemmas that TermoStat yielded, i.e. without a specificity threshold. We reproduce in Table 11 the results for Arabic, but this time lemmas with negative specificity will be included.

Table 11. Units retained with negative specificity for Arabic

Total	**S+ 1.96**	**S=**	**S– retained**
1194	18	5	45

As shown in Table 11, we took into consideration units with positive specificity (S+), units ranked with an unsurprising specificity (S=) and units with negative specificity (S–). We found out that certain verbs of this category (negative specificity) could be considered as terminological ones.

The same procedure was applied to English and French units. By doing so, we realized that, for these two languages, units under the threshold of 1.96 ranged between tagging errors to general language units. Very limited number of specialized units could be observed. Examples of units examined are shown in Table 12.

Table 12. Units under 1.96 for English and French verbs

Unit type	English examples	French examples
General language	design, omit, lack, split, bother, know	garantir (to guarantee), voir (to see), devoir (must), purger (to drain)
Tagging error	value, field, hill, section	police (font), angle (angle), media, site
Candidate units	digitize, unlock, record, retrieve, compose	omettre (to delete), piloter (to manage), inspecter (to scan), composer (to write)

Even though we found it worthy enough to investigate further those units, only the ones with a specificity threshold of 1.96 and above were considered.

In the following section, we validate the chosen VTUs, and we will be looking in depth at those for Arabic.

4.3 VTU validation criterion

As stated in the introduction, since specialized verbs are not often found in specialized dictionaries, we decided to validate the terminological status of our candidate VTUs by applying a list of criteria. The validation process takes into consideration four criteria as proposed by L'Homme (2004). These criteria are:

1. The meaning of the VTU is related to the specialized domain: in this case, a specialized dictionary is needed;
2. The nature of semantic *actants*,[4] if they are admitted as terms: like in *to format* ***a hard disk***;
3. The morphological relationship between the VTU and its derivatives, if they are admitted as terms: the kind of morphological relationship between *to install* and *installation*;
4. The paradigmatic relationships between the VTUs[5]: on the paradigmatic axis, a relationship can be of synonymy, antonymy, etc.

Applying these criteria to the entire list of candidate VTUs gave a good overview as to their terminological status. However, it is worth mentioning that not all of the four criteria can be applied. Certain VTUs could match all of them, others could match only two. For example, some VTUs did not have a definition in the dictionary (criterion 1), and some did not have a corresponding derivative (criterion 3). Verbs such as *to apply*, French *appliquer*, Arabic *Tab~aqa* (طبق), do not have

4. According to the Meaning-Text theory (Mel'čuk et al. 1995), a semantic actant of a lexeme is a lexical meaning which occupies an actantial semantic position associated to this lexeme in the lexicon.

5. The last two criteria are in line with the Meaning-Text theory (Mel'čuk et al. 1995).

any semantic relationship (though they have a morphological one) with the noun form, *application*, French *application*, Arabic *taTbyq* (تطبيق).

In Tables 13, 14 and 15, we give five examples for each language (the first five VTUs from each list). For the first criterion, we verified the relatedness of the English and French VTUs to the domain via a multilingual electronic dictionary available online for computer and Internet terminology. The dictionary is called the DiCoInfo (olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi). For the Arabic language, we consulted an online dictionary called *Almaany* (www.almaany.com). This dictionary is a general one, in the sense that it contains general language units as well as vocabularies of different domains, including computing.

Table 13. Validation criteria for the first five VTUs of the English list

	Specificity	Criterion 1	Criterion 2	Criterion 3	Criterion 4
Click	81.26	Available in DiCoInfo	user, link (icon, settings, etc.)	click (n.)	press (synonym)
Install	74.25	Available in DiCoInfo	program, application	installation (n.)	uninstall (antonym)
Type	65.78	Available in DiCoInfo	text, script	typing (n.)	enter (synonym)
Configure	61.9	Available in DiCoInfo	PC, memory	configuration (n.)	setup (synonym)
Access	57.53	Available in DiCoInfo	Internet, window	access (n.)	enter (synonym)

Table 14. Validation criteria for the first five VTUs of the French list

	Specificity	Criterion 1	Criterion 2	Criterion 3	Criterion 4
cliquer	164.16	Available in DiCoInfo	lien, icône, fenêtre	clic (n.)	appuyer (synonym)
sélectionner	71.56	Available in DiCoInfo	utilisateur, option	selection (n.)	choisir, cocher (synonym)
exécuter	70.62	Available in DiCoInfo	commande, programme	exécution (n.) exécutable (adj.)	arrêter(antonym)
désactiver	68.54	Available in DiCoInfo	function, menu	désactivation (n.)	activer (synonym)
configurer	67.26	Available in DiCoInfo	logiciel, fichier	configuration (n.)	paramétrer (synonym)

Table 15. Validation criteria for the first five VTUs of the Arabic list

	Specificity	Criterion 1	Criterion 2	Criterion 3	Criterion 4
Eamila عمل **(run)**	87.07	Available in Almaany	barnAmaj (برنامج) (program), niZAm (نظام) (system)		A$tagala (اشتغل) (run) (synonym)
daEama دعم **(support)**	36.94	Not available in Almaany	lawHap >um (لوحة أم) (motherboard), muEAlij (معالج) (processor)	daEm (دعم) (n.) (support) madEwm (مدعوم) (adj.) (supported)	
EaraDa عرض **(display)**	34.87	Available in Almaany	mujal~ad (مجلد) (folder), qurS (قرص) (disk), risAlap (رسالة) (massage)	EarD (عرض)(n.) (display)	Axf| (أخفى) (hide) (antonym), >Zhara (أظهر) (display) (synonym)
naqara نقر **(click)**	18.32	Available in Almaany	rAbiT (رابط) (link), >yqwnap (أيقونة) (icon)	naqr (نقر) (n.) (click)	DagaTa (ضغط) (press) (synonym)
DagaTa ضغط **(press)**	16.28	Not available in Almaany	miftAH (مفتاح) (key), zir~ (زر) (button)	DagT (ضغط) (n.) (pressing)	naqr (نقر) (n.) (synonym)

As shown in the three tables above, for English and French, all the examples matched the criteria. For Arabic, certain VTUs simply do not appear in the dictionary, and certain VTUs do not have morphological derivatives or paradigmatic relationships.

4.3.1 *Some particularities regarding validation: English and French*

In some cases, validating the terminological status of VTUs could be problematic. For English, a VTU like *ignore* is used in the corpus in two different ways: sometimes it is used in a general context to denote the action of ignoring something and sometimes it is used with a specialized meaning to denote the action of ignoring a command or a computer peripheral.

> *Ignore what it says on the box: A UPS gives you maybe five minutes of computer power.*
> *The purpose of the On-Line or Select button is to tell your printer whether to ignore the computer.*

Another case for English is the VTU *to corrupt*. Since it is only found in the past participle tense, it is tricky enough to judge whether to consider it.

> *The file system may get corrupted.*

For French, we came across a VTU such as *piloter* in:

> *Il existe un certain nombre de ces langages, avec des variantes pour piloter les différents modèles et marques d'imprimantes du marché.*
> (There is a number of these languages with variants to pilot the different models and printers' types in the market).

At the first glance, this verb seemed to be specialized, but we could not verify its terminological status due to the fact that there are not enough contexts. Another example is the VTU *rafraîchir* (to refresh) which TermoStat ranked with a specificity of 2.77. According to the contexts, this verb represents a terminological meaning (*mettre à jour*, to update), as in:

> *Dans les deux cas, on n'oubliera pas de rafraîchir la base de données.*
> (In both cases, we will not forget to refresh the database).

However, because there were only four contexts for this verb, we were left in doubt as to whether it should be retained or not. As a matter of fact, since our analysis of terms and their behavior in texts is based on a *bottom-up* technique (Teubert 2009), contexts constitute for us the evidence for the use of terms in the specialized domain. For example, in order to make sure that criterion 2 can be applied, maximum participants should be investigated. Consequently, a bigger corpus could always be a more fertile environment, especially when dealing with verbs.

4.3.2 *Some particularities regarding validation: Arabic*

For Arabic, VTUs with negative specificity were highly problematic, and the absence of certain verbs was something that should be considered. To begin with, as mentioned, terms with negative specificity were at first eliminated systematically from the extraction list yielded by TermoStat. According to Drouin (2003), these are forms that are distinguished by their statistically significant behavior in the corpus, but in a negative way. In the case of Arabic, we found out that a good number of these verbs could have a specialized meaning. For example, for the verb *HafiZa* (حفظ) (to save), we found in several contexts that the above-mentioned criteria could be applied to it as explained in Table 16.

Table 16. VTU with an unsurprising specificity

	Specificity	Criteria 1	Criteria 2	Criteria 3	Criteria 4
HafiZa	0.85	Available in Almaany	malaf (ملف) (file), kalimap sir (كلمة سر) (password)	HifZ (حفظ) (n.) (saving)	Ha*afa (حذف) (to delete) (antonym)

Another example with a lower specificity is the verb *>qlaEa* (أقلع) (to boot). Consider Table 17.

Table 17. VTU with a negative specificity

	Specificity	Criteria 1	Criteria 2	Criteria 3	Criteria 4
>aqlaEa	−13.20	Available in Almaany	Haswb (حاسوب) (computer), windwz (ويندوز) (Windows)	>qlAE (إقلاع) (n.) (booting)	A$tagala (اشتغل) (to start) (synonym)

The conclusion we have reached is that the specificity of these terms is "rather semantic and not purely lexical" (Drouin 2004, 81). Moreover, other elements observed in the extraction lists can be added to our conclusion, such as the linguistic phenomenon of homography and polysemy which are highly problematic in the Arabic language of technology. Drouin (2004, 81) concludes:

> Multiple phenomena come into play in this case, including homography, polysemy and de-terminologization (Galisson 1978; Meyer and Mackintosh 2000). In order to be able to identify the lexical items that were missed, one must also look at semantic aspects. Without an additional level of tagging that could take meaning into account, these items cannot be accurately retrieved using a purely statistical approach.

The polysemy that exists in verbs is because in terminological creation, two procedures come into play: neology of form and neology of meaning (Rondeau 1984). The kind of neology we mostly have is the latter. This issue can be easily observed in units such as *Ham~ala* (حمل) (literally meaning *to load*) that has acquired the specialized meaning "to download" found in computing. Yet, this meaning was not distinguished by the morphological analysis.

As far as homography is concerned, Arabic is considered to be of a highly complex morphological system (Habash 2010). The absence of the written short vowels in lexical units creates confusion. For example, the verb *TabaEa* (طبع) (to print) is proposed two times in the extraction list. The first time, it is proposed with its real meaning (to print), but the second time as *Tab~aEa* (to stamp), a meaning which is not expressed in the corpus. The presence of gemination on the "b" (marked by the "~" sign in the example) changes the meaning of the word.

Finally, verbs in technical languages (in the field of computing to say the least) are more often subject to nominalization, due to the fact that in terminology verbs receive less attention compared to nouns (L'Homme 2015). In technical Arabic, the presence of support verbs (and certain verbs that accompany nominal forms) is a way of substituting the verbal form by a nominal one. According to our observations, in the Arabic corpus, constructions such as the following are very common:

qum bitaHmyl Almalaf
(قم بتحميل الملف)
(do the download of the file)

In such a case, the nominal form seems always to be preferred. Examples of these verbs are:

qAma (قام) (do);
samaHa (سمح) (allow);
AstaTAEa (استطاع) (can).

Now that we have the results for the three languages, and before we move on to our last section where we will talk about the precision of the extraction, we would like to examine the first ten units produced by the extractor. Those units represent the ones with the highest specificity, and therefore, the most significant in the corpus. Our aim is to see which ones could be common for the three lists of candidate VTUs. Let us consider Table 18.

Table 18. Comparing extraction results for the three languages

VTUs specificity for Arabic	VTUs specificity for English	VTUs specificity for French
Eamila (عمل) (run) (87.07)	click (81.26)	cliquer (to clic) (164.16)
daEama (دعم) (support) (36.94)	install (74.25)	sélectionner (to choose) (71.56)
EaraDa (عرض) (display) (34.87)	type (65.78)	exécuter (to execute) (70.62)
naqara (نقر) (click) (18.32)	configure (61.9)	désactiver (to disable) (68.54)
DagaTa (ضغط) (press) (16.28)	access (57.53)	configurer (to configure) (67.26)
kataba (كتب) (write) (14.42)	display (50.02)	activer (to activate) (52.96)
Had~ada (حدد) (highlight) (13.43)	connect (49.09)	modifier (to modify) (52.61)
AxtAra (اختار) (select) (13.43)	log (47.36)	démarrer (to boot) (50.66)
Ham~ala (حمل) (download) (9.46)	file (45.98)	connecter (to connect) (46.45)
DabaTa (ضبط) (set) (8.02)	delete (45.2)	mémoriser (to memorize) (45.86)

From what can be observed in Table 18 and according to the first ten units analyzed, the only VTU that is among the top ten units in the three lists is the English VTU *to click*, with the corresponding French *cliquer* and Arabic *naqara* (نقر). For Arabic, it comes at the fifth position, while it is at the first one for English and French. Apart from that, we realized that English and French have only two VTUs in common, English *to configure* and French *configurer*, and English *to connect* and French *connecter*. As for Arabic and French, they have only one VTU in common, Arabic *AxtAra* (اختار) and French *sélectionner* (to select). Finally, we found

that Arabic and English have also two VTUs in common: English *type*, and Arabic *kataba* (كتب) (to write), and Arabic *EaraDa* (عرض) (to display) and English *to display*. Concerning the first couple, both designate the same meaning of *information input*. However, *kataba* (كتب) (to write) is polysemous, it designates the idea of writing a program as well.

5. Evaluation and discussion

In Sections 3 and 4, we presented all the steps constituting our methodology for the extraction of specialized verbs in the domain of computing and we talked about the results. In this section, we will discuss some issues ensuing from the methodology and the results obtained. First, we will talk about the precision of the extractor for the three languages. Second, we will discuss the extraction of nominal units and we will compare them to the verbal ones.

5.1 Precision

Since we are dealing with verbs, examining the precision of the extractor was an important issue to look at. Our evaluation of precision is based on the total number of retained VTUs that are relevant to the domain of computing. We focused on units produced by the extractor before and after filtering (tagging errors). We produce in Tables 19 and 20 what we observed for the three languages.

Table 19. Precision before filtering

	Arabic	English	French
Total	375	546	840
VTUs	18	221	211
Precision	5%	40%	25%

Table 20. Precision after filtering

	Arabic	English	French
Specialized VTU	18	221	211
Non-specialized VTU	25	90	60
Total	43	311	271
Precision	41.8%	71%	77.8%

In Table 19, the extraction for English seems to give the best results, compared with Arabic and French. Note that the precision here is calculated according to the overall units extracted without any filtering. Moreover, in this calculation, we did not take into consideration the units with unsurprising specificity or negative specificity (=S and −S) for Arabic (this applies to Table 20 as well). In Table 20, things seem to change for the three languages. Precision for both English and French looks to be comparable. However, for Arabic, the precision is the lowest.

Considering these precision values and the results obtained from the overall extraction process for Arabic, we decided to see whether the extraction of nominal form of verbs might perform any better in terms of results and specificity. In the following section, we will go through certain details regarding the three languages and we will do some comparisons.

5.2 Comparison between VTUs and NTUs

We start with the assumption that in Arabic technical language the nominal terms are predominant (Ghazzawi 2016). To confirm this assumption, nominal terms were extracted from the very same corpus used for verbs. Then, we analyzed the first fifty nominal terminological units (NTU) (deverbal nouns only) and their corresponding VTUs. In Table 21, we give some examples taken from the top of the list.

Table 21. Specificity of Arabic NTUs

	NTU	**VTU**
	Specificity	**Specificity**
tanfy* تنفيذ (execution)/naf~a*a نفذ (execute)	61.09	−11.42
tavbyt تثبيت (installation)/vab~ata ثبت (install)	48.86	−10.89
TibAEap طباعة (printing)/TabaEa طبع (print)	47.13	−11.18
>t~iSAl إتصال (connection)/it~aSala إتصل (connect)	44.76	−10.77
>rsAl ارسال (sending)/Arsala أرسل (send)	28.42	−8.11

As shown from the examples in Table 21, NTUs are considerably different from VTUs in terms of specificity. The units in the table are taken from the top of the list produced by the extractor. As we went down further in that list, other units showed exactly the same thing. Moreover, certain units (NTUs) did not simply have corresponding verbs (VTUs). The only exception found is the verb *naqara* (نقر) (to click). This verb has a specificity of 18.32 and almost 438 occurrences in the corpus. Its corresponding nominal form, *naqr* (نقر) (click), has a negative

specificity of 9.27. As we analyzed the contexts in which this verb was found, we realized that it came, for the most part, in its imperative form.

To sum up our comparison for both types of units, in Figure 4, we show our conclusion for the fifty units analyzed. The raison for which we chose to analyse 50 units was that we intended to analyse what we believed to be the most used specialized units in computing according to our corpora.

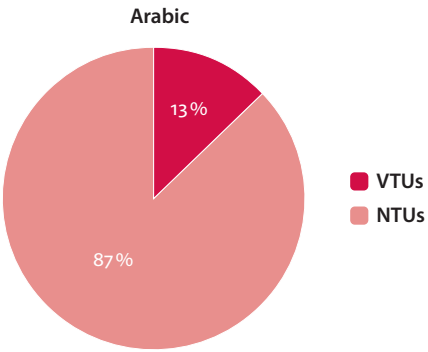

Figure 4. VTU/NTU proportions in Arabic

The same experiment was done on English and French to see whether the same thing could be observed. As stated in the introduction, in the terminological literature, it is a general assumption that terms are mostly nouns. For those two languages, the corpora showed some interesting facts. In Tables 22 and 23, we give examples of proportions for nouns and their corresponding verbs, according to their specificities. For the NTUs, we took the ones with the highest specificity, taken from the top of the list.

Table 22. Specificity of English NTUs

	NTU	VTU
	Specificity	**Specificity**
configuration/configure	63.25	61.9
networking/network	58.85	14.58
connection/connect	54.25	49.09
installation/install	53.1	74.25
partition/partition	32.78	7.18

Table 23. Specificity of French NTUs

	NTU	VTU
	Specificity	**Specificity**
configuration (configuration) /configurer (to configure)	151.36	67.26
connexion (connection)/connecter (to connect)	150.38	46.45
démarrage (startup)/démarrer (to boot)	98.42	50.66
installation (installation)/installer (to install)	66.53	37.33
affichage (display)/afficher (to display)	57.73	38.64

All in all, NTUs seemed to have higher specificity (at least for French). However, for both languages, we discovered some exceptions. Some verbs had higher specificities than their nominal forms. The most significant example we give here is the verb *to click*, French *cliquer*. Just like Arabic, this verb has higher specificity than its nominal form. As a verb, for English, it has a specificity of 81.26 with a frequency of 387, whereas its nominal form has a specificity of 28.47 and a frequency of 77. For French, it has a specificity of 164.16 with a frequency of 387, whereas its nominal form has a specificity of 35.25 and a frequency of 44. Furthermore, French verbs were distinguished from English ones by the fact that most of them had a specialized nominal form, unlike English, where we noticed that certain verbs did not have a nominal form in the corpus. Examples of these verbs are: *to delete* (45.02), *to disable* (14.91), *to load* (20.23), *to rename* (21.1) and *to run* (32.82).

After examining fifty verb/noun couples in the English and French corpora, we obtained the following results, Figures 5 and 6.

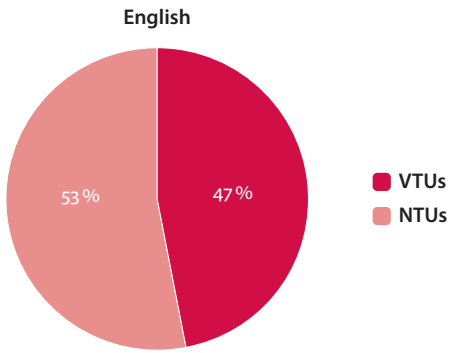

Figure 5. VTU/NTU proportion in English

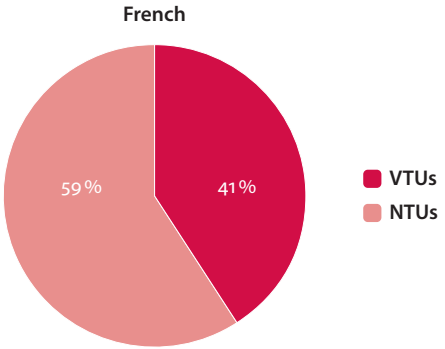

Figure 6. VTU/NTU proportion in French

NTUs represent a higher usage in the corpus compared with VTUs. In English, the difference seems to be less important compared with French.

6. Conclusion

In the present paper, we presented an extraction method based on corpus comparison for discovering specialized verbs, VTUs, in a comparable corpus. The specialized domain is computing and the languages being studied are Arabic, English and French.

For English and French, results yielded by the extractor seemed to be satisfactory in terms of results. As for Arabic, surprising results were obtained. It was found that verbs were not highly significant in the domain of computing. We showed that nominalization of verbs played an important role in Arabic texts. It was shown as well how support verbs participated in systematic rendering of verbs into nouns. For English, it was not exactly the case, since the difference between verbs and nouns extracted was not that big. For French, the difference between these two units was obvious in terms of occurrences and specificity. The terminological validation of the resulting units was based on four criteria. The majority of verbs matched them. However, certain verbs did not match all the four. We concluded that it was not an ultimate condition for a verb, in order to be retained, to match all the criteria. We explained how contexts could be of use when validating the terminological status of a verb. When compiling a specialized corpus, we encountered some difficulties. It was not an easy task converting documents into plain texts. The procedure of converting was, to a certain degree, hard enough and we had to consider certain techniques. The most problematic part was the encoding, as with converting, certain parts of the texts changed into other encodings (the desired encoding being UTF-8). Finally, we believe that much work needs to

be done on Arabic tagging system and morphologic analyzer. Contrary to English and French, Arabic still has a way to go in this regard, as the freely available tools did not fully satisfy our needs. Moreover, more attention has to be paid to specialized Arabic, as most of the researches are carried on the general language.

References

Abed, A. M., S. Tiun, and M. Abared. 2013. "Arabic Term Extraction Using Combined Approach on Islamic Document." *Journal of Theoretical & Applied Information Technology* 58 (3): 601–608.

Ahmad, K., A. Davies, H. Fulford, and M. Rogers. 1994. "What is a term? The Semi-automatic Extraction of Terms from Text." In *Translation Studies: An Interdiscipline*, ed. by M. Snell-Hornby, F. Pöchhacker, and K. Kaindl, 267–278. Amsterdam: John Benjamins. doi: 10.1075/btl.2.33ahm

Almaany. 2017. http://www.almaany.com/. Accessed 30 March 2017.

Attia, M., P. Pecina, A. Toral, L. Tounsi, and J. van Genabith. 2011. "A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer." In *Proceedings Systems and Frameworks for Computational Morphology: Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011*, ed. by M. Cerstin and M. Piotrowski, 98–118. Zurich, Switzerland: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-23138-4_7

Attia, M., P. Pecina, A. Toral, L. Tounsi, and J. van Genabith. 2011a. "An Open-Source Finite State Morphological Transducer for Modern Standard Arabic." In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, 125–133. Blois, France: Association for Computational Linguistics.

Attia, M., P. Pecina, A. Toral, and J. van Genabith. 2013. "A Corpus-Based Finite-State Morphological Toolkit for Contemporary Arabic." *Journal of Logic and Computation* 24 (2): 455–472. doi: 10.1093/logcom/exs070

Chung, T. M. 2003. "A Corpus Comparison Approach for Terminology Extraction." *Terminology* 9 (2): 221–246. doi: 10.1075/term.9.2.05chu

Church, K., and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.

Déjean, H., and E. Gaussier. 2002. "Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables." In *Corpus Linguistics: Critical Concepts in Linguistics*, ed. by W. Teubert and R. Krishnamurthy, 1–22. New York: Routledge.

DiCoInfo. 2017. http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search2.cgi?ui=fr. Accessed 30 March 2017.

Drouin, P. 2002. *Acquisition automatique des termes: l'utilisation des pivots lexicaux spécialisés*. Doctoral thesis. Université de Montréal.

Drouin, P. 2003. "Term Extraction Using Non-technical Corpora as a Point of Leverage." *Terminology* 9 (1): 99–115. doi: 10.1075/term.9.1.06dro

Drouin, P. 2004. "Detection of Domain Specific Terminology Using Corpora Comparison." In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, 79–82. Lisbon, Portugal: ELRA – European Language Resources Association.

Fung, P. 1998. "A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-Parallel Corpora." In *The 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, 1–17. Langhorne, PA, USA: Springer Berlin Heidelberg.

Galisson, R. 1978. *Recherches de lexicologie descriptive : la banalisation lexicale*. Paris: University of Montréal.

Ghazzawi, N. 2016. *Du terme prédicatif au cadre sémanrique: méthodologie de compilation d'une ressource terminologique pour les termes arabes de l'informatique*. Doctoral thesis. University of Montréal.

Guilbert, L. 1973. "La spécificité du terme scientifique et technique." *Langue française* (17): 5–17. doi: 10.3406/lfr.1973.5617

Habash, N., and F. Sadat. 2006. "Arabic Preprocessing Schemes for Statistical Machine Translation." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 49–52. New York, USA: Association for Computational Linguistics. doi: 10.3115/1614049.1614062

Habash, N., O. Rambow, and R. Roth. 2009. "MADA+ TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization." In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, 102–109. Cairo, Egypt.

Habash, N. 2010. "Introduction to Arabic Natural Language Processing." *Synthesis Lectures on Human Language Technologies* 3 (1): 1–187. doi: 10.2200/S00277ED1V01Y201008HLT010

Kilgarriff, A. 2001. "Comparing Corpora." *International Journal of Corpus Linguistics* 6 (1): 97–133. doi: 10.1075/ijcl.6.1.05kil

Lafon, P. 1980. "Sur la variabilité de la fréquence des formes dans un corpus." *Mot* 1 (1): 127–165. doi: 10.3406/mots.1980.1008

Lebart, L., and A. Salem. 1994. *Statistique textuelle*. Paris: Dunod.

Lemay, C., M.-C. L'Homme, and P. Drouin. 2005. "Two Methods for Extracting Specific Single-Word Terms from Specialized Corpora: Experimentation and Evaluation." *International Journal of Corpus Linguistics* 10 (2): 227–255. doi: 10.1075/ijcl.10.2.05lem

L'Homme, M.-C. 2004. *La terminologie: Principes et Techniques*. Montréal, Canada: Les presses de l'université de Montréal.

L'Homme, M.-C. 2015. "Predicative Lexical Units in Terminology." In *Recent Advances in Language Production, Cognition and the Lexicon*, ed. by N. Gala, R. Rappand, and G. Bel-Enguix, 75–93. Switzerland: Springer.

Lorente, M. 2007. "Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació." In *Estudis de lingüístics i de lingüística aplicada en honor de M. Teresa Cabré Catellví. Volum II: De deixebles*, ed. by M. Lorente, R. Estopà, J. Freixa, J. Martí, and C. Tebé, 365–380. Barcelona: Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra.

Mel'čuk, I., A. Clas, and A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Meyer, I. 2000. "Computer Words in Our Everyday Lives: How are They Interesting for Terminography and Lexicography?" In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*, ed. by H. Ulrich, S. Evert, E. Lehmann, and C. Rohrer, 39–58. Stuttgart, Germany: Institut für Maschinelle Sprachverarbeitung.

Meyer, I. and K. Mackintosh. 2000. "When terms move into our everyday lives: An overview of de-terminologization." *Terminology* 6(1), 111–138.

Monsonego, S. 1969. "Ch. Muller: Étude de statistique lexicale. Le vocabulaire du théâtre de P. Corneille." *Langue française* 3 (1): 107–110.

Muller, C. 1967. *Étude de statistique lexicale, le vocabulaire du théâtre de Pierre Corneille*. Paris: Larousse.

Muller, C. 1977. *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Nelson, M. B. 2000. Corpus-based Study of the Lexis of Business English and Business English Teaching Materials. Unpublished Ph.D Thesis, University of Manchester, Manchester.

Rapp, R. 1999. "Automatic Identification of Word Translations from Unrelated English and German Corpora." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ed. by R. Dale and K. Church, 519–526. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: 10.3115/1034678.1034756

Rayson, P., and R. Garside. 2000. "Comparing Corpora Using Frequency Profiling." In *Proceedings of the workshop on Comparing Corpora*, 1–6. Stroudsburg, PA, USA: Association for Computational Linguistics.

Reppen, R. 2001. "Review of MONOCONC PRO and WORDSMITH TOOLS." *Language Learning & Technology* 5 (3): 32–36.

Rey, A. 1979. *La terminologie: noms et notions. Coll. "Que sais-je ?"*. Paris: Presses universitaires de France.

Rondeau, G. 1984. *Introduction à la terminologie*. Chicoutimi, Québec: G. Morin.

Sager, J. C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins. doi: 10.1075/z.44

Scott, M. 1997. "PC Analysis of Key Words – and Key Key Words." *System* 25 (1): 233–345. doi: 10.1016/S0346-251X(97)00011-0

Teubert, W. 2009. "La linguistique de corpus: une alternative." *Semen. Revue de sémio-linguistique des textes et discours* 27: 185–211.

Toutanova, K., and C. Manning. 2000. "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger." In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 63–70. Hong Kong: Association for Computational Linguistics.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer. 2003. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In *Proceedings of HLT-NAACL*, 173–180. Edmonton, Canada: Association for Computational Linguistics.

Xu, F., D. Kurz, J. Piskorski, and S. Schmeier. 2002. "A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, ed. by M. González Rodríguez and C. Paz Suarez Araujo, 134–145. Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA).

Authors' addresses

Nizar Ghazzawi
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7
Canada

nizar.ghazzawi@umontreal.ca

Patrick Drouin
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7
Canada

patrick.drouin@umontreal.ca

Benoît Robichaud
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7
Canada

benoit.robichaud@umontreal.ca

Fatiha Sadat
Université du Québec à Montréal
201 Président Kennedy
Montréal (Québec), H2X 3Y7
Canada

sadat.fatiha@uqam.ca

Biographical notes

Nizar Ghazzawi holds a PhD in translation with a specialization in terminology from the University of Montréal. He worked as a research and teaching assistant at the Translation and Linguistics Department and Literature and Modern Language Department of the University of Montreal. His research interests include terminology, terminography, corpus linguistics, lexical semantics and translation. He is also a professional terminologist, interpreter and translator.

Benoît Robichaud holds a DEA (Diplôme d'Études Approfondies) in theoretical formal and automatic linguistics from the University of Denis-Diderot (Paris 7). He worked in several private companies as a developer researcher in computational linguistics. His projects were on computer-aided tools for correction, automatic translation, and information research. He has been a research agent at the Observatory of Meaning-Text Linguistics of the University of Montreal for almost half a dozen years.

Patrick Drouin is full professor in Translation at the Translation and linguistics Department of the University of Montréal where he teaches terminology and localization. He obtained his Ph.D. in linguistics from the University of Montréal in 2002, working on term extraction using hybrid techniques. His main research interests are computational and corpus linguistics applied to Terminology. Prior to 2002, he worked in the private sector as a senior translation and terminology technology specialist.

Fatiha Sadat is a computer science professor at Université du Quebec à Montreal (UQAM), Canada.