Improving term candidates selection using terminological tokens

Mercè Vàzquez and Antoni Oliver
Universitat Oberta de Catalunya

The identification of reliable terms from domain-specific corpora using computational methods is a task that has to be validated manually by specialists, which is a highly time-consuming activity. To reduce this effort and improve term candidate selection, we implemented the Token Slot Recognition method, a filtering method based on terminological tokens which is used to rank extracted term candidates from domain-specific corpora. This paper presents the implementation of the term candidates filtering method we developed in linguistic and statistical approaches applied for automatic term extraction using several domain-specific corpora in different languages. We observed that the filtering method outperforms term candidate selection by ranking a higher number of terms at the top of the term candidate list than raw frequency, and for statistical term extraction the improvement is between 15% and 25% both in precision and recall. Our analyses further revealed a reduction in the number of term candidates to be validated manually by specialists. In conclusion, the number of term candidates extracted automatically from domain-specific corpora has been reduced significantly using the Token Slot Recognition filtering method, so term candidates can be easily and quickly validated by specialists.

Keywords: automatic term extraction, terminology extraction, domain-specific corpora, terminological tokens, TSR filtering method, TBXTools, term candidates, terminological units

1. Introduction

Since the early 1980s, Automatic Term Extraction (ATE) has been considered a relevant Natural Language Processing task involving terminology and has been used to identify domain-relevant terms by applying computational methods (Oliver et al. 2007; Foo 2012). Automatic Term Extraction has been considered relevant due to its accurate terminology construction that can improve a wide

https://doi.org/10.1075/term.00016.vaz
Terminology 24:1 (2018), pp. 122–147. issn 0929-9971 | e-issn 1569-9994

range of tasks, such as ontology learning, machine translation, computer-assisted translation, thesaurus construction, classification, indexing, information retrieval, as well as text mining and automatic summarisation (Heid and McNaught 1991; Frantzi and Ananiadou 1997; Vu et al. 2008). To this end, different linguistic, statistical and hybrid methods have been implemented to automatically identify domain-relevant terms and to assist in the manual term selection done by specialists (Kageura and Umino 1996; Pazienza et al. 2005; Valaski et al. 2015). However, the main automatic term extraction methods implemented up to now usually involve extracting a large list of term candidates that has to be manually selected by specialists (Bourigault et al. 2001; Vivaldi and Rodríguez 2001). This is a highly time-consuming activity and a repetitive task that poses the risk of being unsystematic, subjective, and very costly in economic terms (Loukachevitch 2012; Conrado et al. 2013; Vasiļjevs et al. 2014). Therefore, it is difficult to select the most reliable terms from the corpora (Nazarenko and Zargayouna 2009; Gornostay 2010).

In order to achieve a more accurate and precise term candidate selection, we implemented the Token Slot Recognition (TSR) method, a filtering method based on terminological tokens which is used to rank extracted term candidates from domain-specific corpora. The TSR filtering method has been implemented in TBXTools, a term extraction tool, and can be used both with statistical and linguistic term extraction (Oliver and Vàzquez 2015). The method is based on the hypothesis that complex terms are formed by tokens that tend to appear in the same position in other complex terms. We observed that terminological tokens tend to appear in the same positions by analysing some of IATE's glossaries. Specifically, we extracted terminological tokens from IATE's Economics and Health glossaries in English, Spanish and French in order to identify the number of tokens in each position (Table 1).

Table 1. Distribution of terminological tokens on IATE's glossaries

Domains and languages	First and second positions	First position	Second position
Economics English	14.36%	58.35%	27.29%
Economics Spanish	3.61%	36.98%	59.41%
Economics French	4.5%	36.07%	59.42%
Health English	8.98%	65.48%	25.53%
Health Spanish	1.93%	35.85%	62.22%
Health French	2.73%	33.7%	63.57%

The primary goal of our study is to determine whether the TSR filtering method applied along with linguistic and statistical term extraction approaches

provides a more accurate and precise term candidate's selection, to better validate terms from candidates and avoid the long post-revision of the output carried out by a specialist. This primary goal is based on two hypotheses: The first hypothesis, leads us to presuppose that a filtering method based on reference terms as well as the frequency of occurrence of a term in a corpus used to rank a list of term candidates, could provide an accurate and precise term candidate selection. The second hypothesis leads us to assume that this filtering method, ranking the most likely term candidates to the top of the list, could improve the manual term selection done by specialists.

In order to address this goal and the baseline hypotheses, this paper describes the TSR filtering method implementation in linguistic and statistical term extraction approaches using TBXTools. The filtering method has been applied in several domain-specific corpora (Economics, Medicine, and Social Services) and different languages (English, French, Spanish, and Catalan).

This paper is structured as follows: in Section 2, the background of automatic term extraction is described. In Section 3, the materials used and the method implemented to extract and automatically filter terms from corpora are presented. The results and discussion are described in detail in Section 4. The paper is concluded with some final remarks and ideas for future research.

2. Background

The task of automatic term extraction uses computational methods to select term candidates from a corpus that can then be processed to carry out a terminology project (Oliver et al. 2007; Foo 2012). Automatic term extraction methods employ different strategies in order to obtain lexical units that are representative of specialised corpora. Then, these methods are briefly described and classified according to the strategy they employ: linguistic, statistical, hybrid, or machine learning (Pazienza et al. 2005; Lossio-Ventura et al. 2014; Valaski et al. 2015).

Linguistic approaches have already been used in early works on terminology extraction. Ananiadou, for example, (1988, 1994a, 1994b), focuses on lexical morphology, and there is also interest in developing methodologies for the recognition of terms that can apply theoretical training related to the formation of terms. In addition, Daille (1997) and Jacquemin (1999) made a thorough syntactic analysis of term candidates, including a morphological analysis and dependency analysis (head-modifier dependency analysis). Linguistic methods seek to identify terms by considering their syntactic patterns (Pazienza et al. 2005), the grammatical form of terminological units (Bourigault 1992), linguistic patterns from which systems can establish correct forms, linguistic knowledge (Basili et al. 1997), or regu-

lar expressions to identify candidate terms (Daille 1994; Justeson and Katz 1995). The linguistic knowledge that these methods use can recognise words based on: lexicographic resources, as in the case of Fastr (Jacquemin 1994); morphological resources, as in the case of Terms (Justeson and Katz 1995); morphosyntactic resources, as in the case of Termino (David and Plante 1990); extraction of the maximum length noun phrases based on a superficial grammatical analysis and an analysis of the maximum length of noun phrases to extract appropriate term candidate units, as in the case of Lexter (Bourigault et al. 1996); grammatical category and word alignment, as in the case of Termight (Dagan and Church 1994); syntactic functions, as in the case of Nodalida (Arppe 1995); and occasionally filtering terms based on non-terms morphology and contextual analysis (Drouin 1997). The current terminology recognition methods incorporate automatic morphosyntactic tagging (part-of-speech tagging), allowing for the categorisation of each word in a lexicon type. They apply regular expressions based on matching patterns and use linguistic filters to refine the results in order to identify the terms in a document.

Statistical approaches use common calculations from other areas such as information retrieval and collocation detection to determine the possibility that a candidate extracted from a corpus is a term. To do so, these calculations take into account the *unithood*, 'the degree of strength or stability of syntagmatic combinations and collocations', and *termhood*, 'the degree that a linguistic unit is related to domain-specific concepts' of a term (Kageura and Umino 1996, 260). Unithood measurements rely on statistical tests to determine the degree of association between the components of a candidate term, ranging from simple frequencies to more complex measures. Lexical association measures extract the candidates that are more likely to be terms by degree of association, such as the log-likelihood ratio, the Pearson chi-squared test, the Odds ratio, the PHI coefficient, the Student's *t* test, the Dice coefficient or mutual information measure (Evert and Krenn 2001, Evert 2005; Wong et al. 2007; Vàzquez and Oliver 2013). Termhood employs measures of relevance such as those used in information retrieval. Frequency is the termhood calculation highlighted to identify relevant textual relationships and multiword units, as indicated in Smadja (1993), Daille (1995), McEnery et al. (1997), Merkel and Andersson (2000), Piao and McEnery (2001) and Pereira et al. (2004). The frequency of occurrence of a term in a corpus is a statistically significant scale in the task of automatic extraction of terms, because the more times the candidate appears in a context, the more evidence there is of it being a term. However, the frequency calculation has to be combined with another filtering method in order to determine whether a unit has a terminological status.

Hybrid approaches combine linguistic and statistical methods and allow the status of a linguistic unit to be confirmed or denied. The profitability of lexical

association measures increases when the statistical knowledge is applied to the list of previously selected candidates from linguistic filters, because the linguistic filters help to select candidates before applying numerical tests, as is the case in Acabit (Daille 1995) and Clarit (Evans and Zhai 1996). One of the first systems using a hybrid approach was Earl (1970). In this work, the noun phrases are first extracted as term candidates and then they are selected according to the frequency of appearance in the corpus. In Daille's (1994) work, the linguistic candidates obtained from syntactic patterns are filtered with different statistical measures such as log-likelihood, frequency, and mutual information. Justeson and Katz (1995) implemented a similar system in which regular expressions are used to extract linguistic candidates from a corpus, which are ranked by frequency. A hybrid method is also used in ANA (Automatic Natural Acquisition of terminology) (Enguehard and Pantera 1995). First, it extracts simple terms from the corpus according to the criterion of frequency. After this, these terms are used in combination with linguistic character heuristics and considerations of frequency to extract other term candidates. Dias (2003) proposed a hybrid system called HELAS (Hybrid Extraction of Lexical Association), which extracts multiword units from a corpus previously morphosyntactically annotated. Unlike conventional methods, it automatically identifies relevant syntactical patterns. It combines linguistic processing using the mutual expectation statistical measure to obtain a level of cohesion that exists in multiword units. A recent study about approaches and strategies applied to relevant term extraction indicates that the hybrid approaches are the most relevant, and the strategies that use noun identification, compound terms and TF-IDF metrics are the most significant (Valaski et al. 2015).

A further step is to deepen linguistic analysis using semantic and contextual information. For this, semantic strategies are used to refine the results with statistical and linguistic extraction methods. There are basically two types of these strategies. The first involves using lexical semantic categories from an external lexical source of the corpus, such as WordNet (Miller 1995), EuroWordNet (Vossen 1998) or AlethDic (Naulleau 1998), which organise lexicons from the meaning of words so they can be integrated into a candidate extraction tool. The second of these strategies involves extracting the semantic categories of the words from the same corpus through contextual elements that refer to the syntactic-semantic combination of words (Fabre 1996). Maynard and Ananiadou (1999) use context-factor to introduce the semantic and contextual information, and Velardi et al. (2001) use domain relevance and domain consensus to achieve a similar purpose. External semantic resources are used in particular for building ontologies in the medical domain to achieve quick access to relevant information by taking advan-

tage of the thesaurus (Bentounsi and Boufaida 2013; Messaoudi et al. 2013; Dramé et al. 2014; Bouslimi et al. 2016).

In recent research, automatic term extraction is done using machine learning tasks (Fedorenko et al. 2013), because they are able to independently learn how to recognise a term, facilitate the use of a large number of measures, and save time extracting terms (Conrado et al. 2013). Supervised machine learning methods are used to extract special domain terms (Liu et al. 2008; Zheng et al. 2009) and are designed for specific entity classes and integrating term extraction and term classification (Lossio-Ventura et al. 2016).

The integration of terminological resources in a term extraction process at the filtering step is implemented successfully in Fastr (Jacquemin 1994). Furthermore, the use of terminology in Yatea (Aubin and Hamon 2006) positively influences the identification of nominal phrases, the linguistic analysis, and finally the extraction from the terms list. In addition, Jiang et al. (2015) uses the title words and the keywords in research papers as the seeding terms to identify similar terms from an open-domain corpus as the candidate terms.

3. Materials and methods

In order to get a more accurate and precise term candidate selection, we implemented the Token Slot Recognition method, a filtering method which uses terminological units to rank extracted term candidates from domain-specific corpora. The algorithm is based on the concept of *terminological token*, that is, a *token* or word of a term (or the whole term in the case of unigram terms) to filter out term candidates. Thus, a unigram term is formed by a token that can be the first token of a term (FT) or the last token of a term (LT) depending on the language, a bigram term is formed by FT LT, a trigram term is formed by FT MT LT (where MT is the middle token of a term), and a tetragram term is formed by FT MT_1 MT_2 LT. In general, an n-gram term is formed by FT MT_1 [..] MT_{n-2} LT. The algorithm reads the terminological tokens from a list of already known terms and stores them in three different stacks:

- First Tokens Stack (FTS): stores terminological tokens appearing in the first position of the reference terms.
- Middle Tokens Stack (MTS): stores terminological tokens appearing in any position other than the first or the last of the reference terms.
- Last Tokens Stack (LTS): stores terminological tokens appearing in the last position of the reference terms.

Unigrams may be considered First Tokens, Last Tokens, or both, depending on the language. For example: for English, a unigram term like *rate* can be considered an LT unit as it can also be part of a bigram term like *interest rate*. However, a term like *interest* can be considered either an LT unit, such as *vested interest*, or an FT, like *interest rate*.

Thus, the TSR method filters term candidates by taking into account their tokens. To do so, two filtering variants are designed: strict and flexible filtering. In *strict TSR filtering*, a term candidate will be kept only if all the tokens are present in the corresponding position. For example, a bigram term W_1 W_2 will be kept only if W_1 is present in the FTS and W_2 is present in the LTS. For trigrams, W_1 W_2 W_3 will be kept only if W_1 is present in the FTS, W_2 in the MTS, and W_3 in the LTS. In *flexible TSR filtering,* a term candidate will be kept only if any of the tokens is present in the corresponding position. For example, a bigram term W_1 W_2 will be kept if either W_1 is present in the FTS or W_2 is present in the LTS stack. For trigrams, W_1 W_2 W_3 will be kept if W_1 is present in the LTS, W_2 in the MTS, or W_3 in the LTS.

The algorithm performs this filtering process recursively, that is, by enlarging the list of terminological elements with the new selected term candidates. In strict mode this is not possible, as all the validated candidates are formed with already known terminological tokens and no new elements are acquired. With flexible filtering, however, it is possible to extract new terminological units, as the candidates are validated if they have a terminological unit in any position, so the rest of the tokens can be added to the stacks as new terminological units. Taking into account that the precision of strict filtering in the higher positions is higher than that of flexible filtering but flexible filtering is able to validate more term candidates, a new filtering strategy has been designed: combined TSR filtering. In *combined TSR filtering,* strict filtering is first used and is then followed by flexible filtering.

It should be noted that in flexible and combined TSR filtering the term candidates are processed in each iteration in descending order of frequency; that is, the most frequent term candidates are filtered first. If a term candidate is not filtered out, this is stored in the output stack following that order. Since the process is recursive in these filtering strategies, the term candidates that have been filtered out in the previous iteration are processed again in descending order of frequency in the following iterations. If they are not filtered out in this iteration, they are stored in that order after the term candidates accepted in the previous iteration. The process is repeated until no new terminological tokens are detected.

The TSR filtering method can be applied to terms longer than trigrams in any of the variants. TSR can also be applied to discover new unigram terms from ter-

minological tokens appearing in complex terms, but cannot be applied to filter new unigram candidates from already known unigram terms.

4. Results and discussion

4.1 Experimental settings

In this study, the TSR filtering method was experimented in seven controlled corpora, including four languages (English, Spanish, French and Catalan) and three domain-specific corpora (Economics, Health and Social Services). The corpora were controlled in the sense that the terms in the corpora were manually selected and validated by specialists. The following corpora were used:

- JRC Economics English corpus
- JRC Economics Spanish corpus
- JRC Economics French corpus
- IULA Economics Spanish corpus
- IULA Health Spanish corpus
- TERMCAT Social Services Spanish corpus
- TERMCAT Social Services Catalan corpus

These corpora and the term list come from three sources. The JRC Economics in English, Spanish and French were compiled by Vàzquez (2014). The IULA corpora (Badia et al. 1998; Vivaldi 2009) were created at the University Institute for Applied Linguistics (IULA), a research and graduate training centre at Pompeu Fabra University in Barcelona (Spain). The Social Services corpora were created by the TERMCAT, the Government of Catalonia and the Institute of Catalan Studies' centre for Catalan language terminology.

The experiments were implemented using bigram terms. To do so, the corpora were divided into two parts: one smaller part, called *training*, and one bigger part, called *test*. Two subsets of the terms present were created: the training set, formed by the already known terms present in the training corpus, and the test set, formed by the terms present in the test corpus. Please note that a term can be in both term sets. The size of the corpora and term sets are shown in Table 2.

The test corpora were used to perform the term extraction processes. Then, the training term sets were used to attain the terminological tokens used by the filter to apply the TSR filtering. The test term sets were used to perform the automatic evaluation, as these terms are the terms present in the sub-corpus used to perform the terminology extraction.

Table 2. Size of the corpora (in words) and the reference term lists

	Corpus			Terms		
	Full	**Training**	**Test**	**Full**	**Training**	**Test**
JRC Economics English	11460	2430	9057	129	74	97
JRC Economics Spanish	13406	4042	13406	126	77	80
JRC Economics French	13594	4095	9498	100	50	76
IULA Economics Spanish	41402	12437	28965	296	108	210
IULA Health Spanish	97406	29349	68056	406	180	306
TERMCAT Social Services Spanish	26830	26830	18765	72	31	61
TERMCAT Social Services Catalan	25732	7730	18002	74	34	62

4.2 Term extraction procedure

To perform the terminology extraction process, a tool developed by our research group known as TBXTools was used (Oliver and Vàzquez 2015). This tool is able to perform both statistical and linguistic term extraction, and the functionality to perform TSR filtering was recently added.

For statistical term extraction, the bigrams of the corpus were extracted and filtered with a stop-word list. Bigrams starting or ending with a word in the list of stop-words were rejected. In Table 3, the number of words in the stop-word list for each language is shown.

Table 3. Number of words in the stop-word list for each language

Language	**Stop-words**
English	399
Spanish	229
French	351
Catalan	407

To perform linguistic terminology extraction, the corpora first need to be POS tagged. TBXTools can use Freeling (Carreras et al. 2004; Padró and Stanilovsky 2012) to POS tag a corpus and a rich formalism is used to represent linguistic patterns. Thus, regular expressions can be used over the word form, lemma or tag.

For English, the following linguistic patterns were used:

#||NN.?	A noun (taking the word form) followed by a noun (taking the
|#|NN.?	lemma)

#||JJ.?	An adjective (taking the word form) followed by a noun (taking the
|#|NN.?	lemma)
#||VBG	A verb, gerund or present participle (taking the word form)
|#|NN.?	followed by a noun (taking the lemma)

For Spanish, French, and Catalan, the following linguistic patterns were used:

|#|NC.*	A common noun (taking its lemma) followed by an adjective
|#|AQ.*	(taking its lemma)
|#|NC.*	A common noun (taking its lemma) followed by a common noun
|#|NC.*	(taking its lemma)

The term extraction program detects such patterns in the tagged corpus and sorts them into descending order of frequency.

4.3 Results and evaluation

In this section, the results of the TSR filtering along with evaluation figures are presented. The detailed results for one of the corpora, JRC Economics English, will first be presented followed by an evaluation of the figures and a discussion of the rest of the corpora.

Results for JRC Economics English

Statistical term extraction

The statistical term extraction was performed calculating all the bigrams of the corpus (resulting in a total of 14,338 bigrams) and filtering them with the stop-word list, for a total of 2,720 term candidates. In Table 4, the evaluation results for the raw statistical extraction sorting of the candidates by frequency can be observed. The results were examined by position. For example, in position 50, only the 50 first candidates with the highest frequency were taken into account: 16 of these candidates are correct, giving a precision of 32% and a recall of 16.49% (as there are 97 terms in the whole test corpus). In all the tables presented the recall refers to the overall number of terms in the test corpus. When the position is lower than the number of terms in the test corpus, the recall and F1 figures in the tables are marked with an asterisk (*) as it is impossible to retrieve the overall number of terms when analysing a smaller number of term candidates.

Then, the TSR filtering was performed. To do this filtering, the train term list formed by 74 terms was used. All of these terms are bigrams, and 48 first-position terminological tokens and 53 last-position terminological tokens are attained.

First, the *strict* variant of the TSR filtering was performed; that is, only terms starting with a first position terminological token and ending with a last position

Table 4. Evaluation results for raw statistical term extraction for JRC Economics English

Position	Correct	Precision	Recall	F1
25	8	32.00	*8.25	*13.11
50	16	32.00	*16.49	*21.77
100	21	21.00	21.65	21.32
200	31	15.5	31.96	20.88
500	45	9.00	46.39	15.08
1000	60	6.00	61.86	10.94

terminological token were retained. As already mentioned, only one iteration is possible with the strict variant, as no new terminological tokens are added to the stack and all candidates are validated or rejected in the first iteration. The results of the evaluation can be observed in Table 5. As this filtering method is very restrictive, only 104 term candidates were obtained, and only evaluation results up to position 100 could be shown. The results show that this filtering method is very productive in terms of precision: for position 100, an increment of 28 points in precision is achieved. However, as the filtering is very restrictive, very few term candidates are obtained (from 2,720 of the raw statistical extraction to only 104 term candidates).

Table 5. Evaluation results for statistical term extraction for JRC Economics English with strict TSR filtering

Position	Correct	Precision	Recall	F1
25	22	88.00	*22.68	*36.07
50	36	72.00	*37.11	*48.98
100	49	49.00	50.52	49.75

It is interesting to note that 60 new term candidates were extracted with the strict TSR filtering, as the remaining 44 were already in the training term list.

If a *flexible* TSR filtering is now performed, several iterations can be done, as new terminological tokens are included in the stacks. Namely, four iterations can be performed: 1,238 new term candidates were detected in the first iteration, 619 in the second, 62 in the third, and 4 in the fourth. In Table 6, the evaluation results for flexible TSR filtering are shown. As can be observed, the improvement in precision is now lower – 16 points for position 50 – but as new terminological tokens can now be added to the stacks and the process can be run recursively in several iterations, many more results are achieved than with the strict TSR filtering. For

position 200 – that is, for the first 200 term candidates – there is an improvement of 7.5 points in precision and 15.46 points in recall.

Table 6. Evaluation results for statistical term extraction for JRC Economics English with flexible TSR filtering

Position	Correct	Precision	Recall	F1
25	17	68.00	*17.53	*27.87
50	24	48.00	*24.74	*32.65
100	38	38.00	39.18	38.58
200	46	23.00	47.42	30.98
500	63	12.60	64.95	21.11
1000	78	7.80	80.41	14.22

A third filtering strategy, called *combined,* can then be developed: a strict TSR filtering is performed in a first iteration and then a flexible filtering is done with the rest of the term candidates. In Table 7, the number of new candidates detected in each iteration is shown.

Table 7. Number of new candidates discovered for each iteration in the combined TSR filtering

Iteration	New Candidates
1	60
2	1243
3	550
4	62
5	8

In Table 8, the evaluation results for combined TSR filtering are shown. As can be observed, the results are exactly the same as with strict TSR filtering up to position 100, since a strict filtering is performed in the first iteration. However, more term candidates can now be found. If the results for lower positions are compared, an improvement with regards to both raw frequency and flexible TSR filtering can be seen. For example, combined filtering has achieved a precision of 14% for position 500, whereas flexible filtering has obtained a precision of 12% and raw frequency a precision of 9%. This improvement in precision is accompanied by an increment in recall (72.16%, 64.95%, and 46.39%, respectively).

Table 8. Evaluation results for statistical term extraction for JRC Economics English with combined TSR filtering

Position	Correct	Precision	Recall	F1
25	22	88.00	*22.68	*36.07
50	36	72.00	*37.11	*48.98
100	49	49.00	50.52	49.75
200	57	28.50	58.76	38.38
500	70	14.00	72.16	23.45
1000	81	8.10	83.51	14.77

Linguistic term extraction

A linguistic terminology extraction was also performed on the same corpus. The same process was followed and all the evaluation results are shown in Table 9. A total of 187 term candidates were obtained with the linguistic term extraction process. When strict TSR filtering was performed, only 29 term candidates were obtained, whereas 120 term candidates were obtained with flexible TSR filtering, the same number as with combined filtering. As can be observed in the table, improvements both in precision and recall with flexible and combined TSR filtering have been obtained for position 100. As the number of terms in the test corpus is 97, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 9.

Table 9. Evaluation results for linguistic term extraction for JRC Economics English

	Raw frequency			Strict TSR Filtering			Flexible TSR Filtering			Combined TSR Filtering		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	40.00	*0.34	*0.68	80.00	*0.69	*0.69	56.00	*0.48	*0.95	80.00	*0.69	*1.36
50	36.00	*0.62	*1.22	–	–	–	36.00	*0.62	*1.22	52.00	*0.89	*1.76
100	23.00	0.79	1.53	–	–	–	27.00	0.93	1.79	27.00	0.93	1.79
200	17.65	1.13	2.13	–	–	–	–	–	–	–	–	–

Results for JRC Economics Spanish

For statistical term extraction, 13,737 bigrams have been extracted and 3,510 term candidates remained after filtering with stop-words. After combined TSR filtering, a total of 2,310 candidates were obtained. In Table 10, the evaluation results can be observed. For the first 200 candidates, a precision of 13% with a recall of 32.5% was obtained with raw frequency. Using combined TSR filtering, the precision was improved by 9 points, reaching 22% with an increment of 22.5 points in recall.

Fewer term candidates were obtained using linguistic extraction, with a total of 284. Only 107 remained after combined TSR filtering. If we observe the evaluation results in Table 9, we can see that combined TSR filtering can provide some improvement up to position 50 (4 points in precision and 2.5 points in recall). As the number of terms in the test corpus is 80, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 10.

Table 10. Evaluation results for JRC Economics Spanish

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	36.00	*11.25	*17.14	84.00	*26.25	*40.00	24.00	*7.50	*11.43	44.00	*13.75	*20.95
50	26.00	*16.25	*20.00	58.00	*36.25	*44.62	20.00	*12.50	*15.38	24.00	*15.00	*18.46
100	18.00	22.50	20.00	41.00	51.25	45.56	17.00	21.25	18.89	14.00	17.5	15.56
200	13.00	32.50	18.57	22.00	55.00	31.43	12.50	31.25	17.86	–	–	–
500	7.20	45.00	12.41	10.00	62.50	17.24	–	–	–	–	–	–
1000	4.10	51.25	7.59	5.90	73.75	10.93	–	–	–	–	–	–

Results for JRC Economics French

A total of 14,256 bigrams were extracted, and 2,641 term candidates were obtained after filtering with the stop-word list. If we observe the evaluation results in Table 11 for the first 200 candidates, 4 points in precision improvement and a recall improvement of 10.52 points are achieved with combined TSR.

For linguistic terminology extraction, a total of 480 term candidates were obtained, and 228 remained after combined TSR filtering. An improvement of 1.5 points in precision and an improvement of 3.95 points in recall were obtained for position 200. As the number of terms in the test corpus is 76, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 11.

Table 11. Evaluation results for JRC Economics French

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	24.00	*7.89	*11.88	72.00	*23.68	*35.64	8.00	*2.63	*3.96	32.00	*10.53	*15.84
50	24.00	*15.79	*19.05	48.00	*31.58	*38.10	8.00	*5.26	*6.35	18.00	*11.84	*14.29
100	17.00	22.37	19.32	26.00	34.21	29.55	8.00	10.53	9.09	10.00	13.16	11.36
200	11.00	28.95	15.94	15.00	39.47	21.74	5.50	14.47	7.97	7.00	18.42	10.14
500	6.60	43.42	11.46	7.00	46.05	12.15	5.42	34.21	9.35	–	–	–
1000	3.80	50.00	7.06	4.30	56.58	7.99	5.42	34.21	9.35	–	–	–

Results for IULA Economics Spanish

For this corpus, a total of 39,696 bigrams were extracted using statistical terminology extraction, obtaining a total of 18,818 term candidates after filtering with stop-words. After applying combined TSR filtering, 8,670 term candidates remained. An improvement of 10.5 points in precision and 9.95 points in recall was obtained for position 200.

For linguistic terminology extraction, a total of 809 term candidates were extracted and 464 remained after combined TSR filtering. Then, improvements up to position 100 were obtained (6 points for precision and 0.22 points for recall), whereas combined TSR filtering achieved worse results than raw frequency for position 200 (−6.5 points for precision and −0.48 points for recall). As the number of terms in the test corpus is 210, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 12.

Table 12. Evaluation results for IULA Economics Spanish

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	28.00	*3.32	*5.93	76.00	*9.00	*16.10	40.00	*0.36	*0.72	48.00	*0.44	*0.87
50	22.00	*5.21	*8.43	50.00	*11.85	*19.16	32.00	*0.58	*1.15	44.00	*0.80	*1.58
100	21.00	*9.95	*13.5	32.00	*15.17	*20.58	27.00	*0.98	*1.90	33.00	*1.20	*2.32
200	18.00	*17.06	*17.52	28.50	*27.01	*27.74	27.00	*1.97	*3.67	20.50	*1.49	*2.79
500	11.40	27.01	16.03	17.00	40.28	23.91	18.40	3.35	5.67	13.58	2.30	3.93
1000	8.40	39.81	13.87	9.60	45.50	15.85	18.05	5.32	8.22	–	–	–

Results for IULA Health Spanish

For this corpus, a total of 90,026 bigrams were extracted using statistical terminology extraction, obtaining a total of 29,882 term candidates after filtering with stop-words. After applying combined TSR filtering, 24,023 term candidates remained. An improvement of 32.5 points in precision and 20.87 points in recall was obtained for position 200.

For linguistic terminology extraction, a total of 2,229 term candidates were extracted and 1,062 remained after combined TSR filtering. Improvements up to position 200 were obtained (2.5 points for precision and 0.33 points for recall), but combined TSR filtering achieved worse results than raw frequency for position 500 (−5.6 points for precision and −1.83 points for recall). As the number of terms in the test corpus is 306, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 13.

Table 13. Evaluation results for IULA Health Spanish

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	4.00	*0.33	*0.60	96.00	*7.82	*14.46	40.00	*0.65	*1.28	68.00	*1.11	*2.18
50	6.00	*0.98	*1.68	78.00	*12.70	*21.85	40.00	*1.30	*2.52	60.00	*1.95	*3.79
100	6.00	*1.95	*2.95	60.00	*19.54	*29.48	28.00	*1.82	*3.43	46.00	*3.00	*5.63
200	7.00	*4.56	*5.52	39.50	*25.73	*31.16	21.50	*2.80	*4.96	24.00	*3.13	*5.53
500	5.20	8.47	6.44	21.20	34.53	26.27	16.60	5.41	8.16	11.00	3.58	5.41
1000	4.30	14.01	6.58	11.80	38.44	18.06	13.90	9.06	10.97	6.80	4.43	5.36

Results for TERMCAT Social Services Spanish

For this corpus, a total of 20,350 bigrams were extracted using statistical terminology extraction, obtaining a total of 5,248 term candidates after filtering with stopwords. After applying combined TSR filtering, 3,420 term candidates remained. For position 200, an improvement of 6 points in precision and 19.67 points in recall were obtained.

For linguistic terminology extraction, a total of 888 term candidates were extracted and 386 remained after combined TSR filtering. Thus, improvements for only the 25 first positions were obtained (12 points in precision and 1.64 points in recall). For position 50, poorer results were obtained with combined TSR filtering than with raw frequency (−2 points in precision and −0.55 points in recall). As the number of terms in the test corpus is 61, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 14.

Table 14. Evaluation results for TERMCAT Social Services Spanish

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	16.00	*6.56	*9.30	64.00	*26.23	*37.21	32.00	*4.37	*7.69	44.00	*6.01	*10.58
50	18.00	*14.75	*16.22	40.00	*32.79	*36.04	30.00	*8.20	*12.88	28.00	*7.65	*12.02
100	10.00	16.39	12.42	24.00	39.34	29.81	22.00	12.02	15.55	17.00	9.29	12.01
200	7.00	22.95	10.73	13.00	42.62	19.92	16.00	17.49	16.71	9.00	9.84	9.40
500	4.60	37.70	8.20	6.20	50.82	11.05	14.40	39.34	21.08	–	–	–
1000	3.00	49.18	5.66	4.00	65.57	7.54	13.18	63.93	21.85	–	–	–

Results for TERMCAT Social Services Catalan

For this corpus, a total of 20,727 bigrams were extracted using statistical terminology extraction, obtaining a total of 4,017 term candidates after filtering with stopwords. After applying combined TSR filtering, 2,484 term candidates remained.

An improvement of 32.5 points in precision and 18.7 points in recall was obtained for position 200.

For linguistic terminology extraction, a total of 870 term candidates were extracted and 388 remained after combined TSR filtering. Then, improvements for the 200 first positions were obtained (2.5 points in precision and 0.33 points in recall). As the number of terms in the test corpus is 62, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 15.

Table 15. Evaluation results for TERMCAT Social Services Catalan

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	4.00	*0.33	*0.60	96.00	*7.82	*14.46	40.00	*0.65	*1.28	68.00	*1.11	*2.18
50	6.00	*0.98	*1.68	78.00	*12.70	*21.85	40.00	*1.30	*2.52	60.00	*1.95	*3.79
100	6.00	1.95	2.95	60.00	19.54	29.48	28.00	1.82	3.43	46.00	3.00	5.63
200	7.00	4.56	5.52	39.50	25.73	31.16	21.50	2.80	4.96	24.00	3.13	5.53
500	5.20	8.47	6.44	21.20	34.53	26.27	16.60	5.41	–	–	–	–
1000	4.30	14.01	6.58	11.80	38.44	18.06	13.90	9.06	–	–	–	–

4.4 Discussion

This study extensively tested TSR filtering with several corpora in several languages. In addition, it was also tested with two terminology extraction methodologies: statistical and linguistic. In the first corpus, JRC Economics English, all the variants of the TSR filtering were tested: strict, flexible and combined. The experimental results have shown that strict filtering is very productive in terms of precision but it has severe problems in recall since it cannot be run iteratively. Flexible TSR filtering obtains clear improvements in recall, but achieves lower precision figures in the first positions. Observing the evaluation results, it is clear that the best TSR variant is combined TSR filtering, as it has the advantages of strict TSR filtering in the first positions and the advantages of flexible TSR filtering regarding recall.

As a general rule, it can be concluded that combined TSR filtering is very productive in the first positions of the results both in terms of precision and recall. In the cases in which a lot of term candidates were obtained (as with statistical term extraction in our experiments), significant improvements are obtained up to position 200 (that is, for the first 200 term candidates with an improvement of 13 points for precision and 26.8 points for recall for JRC Economics English for position 200) and slight improvements are obtained up to position 500 (5 points of precision for position 500 and 2.1 points for position 1000 for the same corpus).

The results are not as good for linguistic term extraction (only 4 points of precision and 0.14 of recall for position 100 for the same corpus), for two main reasons: on the one hand, linguistic term extraction with raw frequency obtains higher precision values than statistical term extraction, so the improvements of TSR filtering are lower; on the other hand, fewer term candidates are obtained with linguistic term extraction.

The TSR filtering method allows for selection of those candidates that are made up of terminological tokens in order to identify terms from specialised corpora. To do so, the method takes into account a list of reference terms to filter term candidates and rank them by frequency. For example, in the JRC Economics English corpus, the method used the reference term *monetary policy* to identify candidates that contain *monetary* and/or *policy* as terminological tokens. Using *policy* as a terminological token in the LTS, the system validated some correct terms, such as *economic policy* (with a frequency of 6), but also some incorrect term candidates, such as *different policy* (with a frequency of 1). Once an incorrect term is validated, its tokens will be added to the stacks, if not already present. In this case *different* would be added to the FTS. This would lead to incorrect validations in the next iteration, for example *different domain* (with a frequency of 1). We must keep in mind, however, that the final list of term candidates is created concatenating the validated candidates of each iteration and ranked by frequency, so all the term candidates validated in the first iteration come higher in the list than the term candidates validated in the second iteration. Thus, in this example, the correct term *economic policy*, validated in the first iteration, is placed in the 6th position, whereas the incorrect term candidate *different policy* is placed in the 158th position (due to its lower frequency). Likewise, the incorrect term candidate validated in the second iteration, *different domain*, is placed in the 931st position. It can therefore be said that those candidates ranked at the top of the list are more likely to be terms than those ranked in lower positions. This approach improves term validation from corpora and helps alleviate the long post-revision of candidates carried out by specialists. To further minimize the effect of incorrect validations, the TSR filtering method can be implemented in a user interface where the terminologist manually validates a given number of term candidates in each iteration, ensuring that the terminological tokens fed into the stacks come from correct terms.

All experiments performed so far have been limited to bigrams. However, to obtain some preliminary results, we have performed an initial experiment for trigrams to verify whether the method will again provide good results. We analysed the positions where tokens from trigram terms are placed. This study was carried out on two domains (Economics and Health) and three languages (English, Spanish and French). For Economics and English, 33.92% of the tokens always appear

in the first position, 20.87% always in the second position and 20.72% always in the third position, whereas only 6.79% appear in all the three positions and 17.70% appear in any other position combination. Results are comparable for other languages and domains, except for position 2 in Spanish and French, where a lower percentage is seen because we have considered trigrams and the second position is frequently occupied by a preposition in these languages. The figures in the Table 16 indicate that at least the first and the last position are relevant for trigrams.

Table 16. Tokens from trigrams terms appearing in different positions

Domains and languages	First, second and third positions	First position	Second position	Third position	Other
Economics English	6.79%	33.92%	20.87%	20.72%	17.70%
Economics Spanish	1.11%	24.05%	9.35%	49.77%	15.73%
Economics French	1.24%	25.82%	7.82%	46.86%	18.26%
Health English	4.96%	35.93%	22.28%	20.49%	16.34%
Health Spanish	0.43%	23.31%	7.63%	58.76%	9.88%
Health French	0.63%	21.38%	8.12%	57.89%	11.98%

To test the algorithm for trigrams we need a larger corpus in order to have enough trigram terms to evaluate the performance. However, to perform an initial experiment based on trigrams, we used the European Central Bank (ECB) corpus and the IATE trigram terms for Economics and Finance. We created a subcorpus with the segments of the ECB corpus containing at least one of the trigram terms. With this process we have obtained a subcorpus with 16,109 segments. We used the 10% of this subcorpus as the corpus for training and the remaining 90% as a test corpus. Two lists of trigram terms were created: the training list, containing the terms appearing in the training corpus; and the test list, containing the terms in the test corpus. With all these data, we performed an experiment and obtain the results shown in Table 17. As the number of terms in the test corpus is 910, the recall and F1 values for positions lower than that figure are marked with an asterisk (*) in Table 17.

As we can see, the TSR filtering methodology also performs well for trigrams in our preliminary test set. For position 100, for example, we get 16 points of improvement for strict and combined TSR filtering and 9 points for flexible TSR filtering, with improvements also in recall and F1 measure. As already mentioned, these results must be seen as preliminary.

Table 17. Preliminary experiments for trigram terms for Economics in English

	Statistical Term Extraction						Linguistic Term Extraction					
	Raw			TSR Combined			Raw			TSR Combined		
Position	P	R	F1	P	R	F1	P	R	F1	P	R	F1
25	16.0	*0.44	*0.86	20.0	*0.55	*1.07	12.0	*0.33	*0.64	20.0	*0.55	*1.07
50	18.0	*0.99	*1.88	32.0	*1.76	*3.33	20.0	*1.1	*2.08	32.0	*1.76	*3.33
100	13.0	*1.43	*2.57	29.0	*3.19	*5.74	22.0	*2.42	*4.36	29.0	*3.19	*5.74
200	11.0	*2.42	*3.96	29.0	*6.37	*10.45	16.5	*3.63	*5.95	29.0	*6.37	*10.45
500	10.4	*5.71	*7.38	29.2	*16.04	*20.71	13.2	*7.25	*9.36	29.2	*16.04	*20.71
1000	9.5	10.44	9.95	26.8	29.45	28.06	11.7	12.86	12.25	26.8	29.45	28.06

5. Conclusions and future work

In this paper, we have presented a novel filtering strategy to select the term candidates from a term extraction process and place them in the first positions of the list to be manually revised. The methodology uses tokens from already known terms to search term candidates containing some of these tokens. As the process is iterative, the list of terminological tokens can be enriched in each iteration, allowing the discovery of completely new terms. In order to determine whether the TSR filtering method provides a more accurate and precise term candidate's selection, the presented methodology was tested in several corpora in four languages: English, Spanish, French and Catalan. The filtering process was tested along with statistical and linguistic term extraction for each corpus. The evaluation results show that combined TSR filtering is more productive both in terms of precision and recall, especially when a long list of term candidates is extracted, than raw frequency. Indeed, TSR filtering provides an accurate and precise term candidate's selection based on reference terms and ranked by frequency, which allows for improvement on the manual term selection done by specialists.

The TSR filtering method with its three variants – strict, flexible and combined – has been implemented in TBXTools, a general tool for terminology extraction, and it is easy to implement in any existing terminology extraction tool.

The TSR filtering method can be used in a fully automatic way or interactively with the user. In this second variant, the user validates the term candidates in each iteration and the new terminological tokens are obtained using only the validated terms. In this way, the terminological token stacks are enriched only with tokens from real terms, avoiding errors in the next iteration.

As a future study, we plan to test the TSR filtering method with larger corpora and in other languages, and also to perform more experiments for trigram and higher n-gram terms for other domains. We also plan to make a comparison of

the results with other filtering strategies and association measures. Once the testing and evaluation of the method is done, we plan to integrate the TSR filtering in a tool with a user interface allowing it to run in interactive mode. When this interface is available, we plan to test it with real users in real situations and perform a usability test and user opinion surveys.

References

Ananiadou, Sofia. 1988. *Towards a Methodology for Automatic Term Recognition*. Dissertation. University of Manchester, Institute of Science and Technology.

Ananiadou, Sophia. 1994a. "A Computational Linguistic Approach to Automatic Term Recognition." In *Proceedings of the 3rd International Society for Knowledge Organization* (ISKO 1994) 4: 134–141. Copenhagen, Denmark: Indeks Verlag.

Ananiadou, Sophia. 1994b. "A Methodology for Automatic Term Recognition." In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)* 2: 1034–1038. Kyoto, Japan. https://doi.org/10.3115/991250.991317

Arppe, Antti. 1995. "Term Extraction from Unrestricted Text." In *Proceedings of the 10th Nordic Conference on Computational Linguistics (NODALIDA 1995)*. Helsinki, Finland: Department of General Linguistics.

Aubin, Sophie, and Thierry Hamon. 2006. "Improving Term Extraction with Terminological Resources." In *Advances in Natural Language Processing. Lecture Notes in Computer Science* 4139. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11816508_39

Badia, Toni, Mercè Pujol, Antoni Tuells, Jorge Vivaldi, Lluis de Yzaguirre, and Teresa Cabré. 1998. "IULA's LSP Multilingual Corpus: Compilation and Processing." In *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada, Spain.

Basili, Roberto, Gianluca De Rossi, and Maria Teresa Pazienza. 1997. "Inducing Terminology for Lexical Acquisition." In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing Conference* (EMNLP 1997). Providence, USA. (http://www.aclweb.org/anthology/W97-0314). Accessed 15 February 2018

Bentounsi, Imene, and Zizette Boufaida. 2013. "Extracting Candidate Terms from Medical Texts." In *International Conference on Computer Systems and Applications (AICCSA)*: 1–4. Fes, Morocco. https://doi.org/10.1109/AICCSA.2013.6616486

Bourigault, Didier. 1992. "Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases." In *Proceedings of the 14th Conference on Computational linguistics* (COLING 1992) 3: 977–981. Nantes, France. https://doi.org/10.3115/992383.992415

Bourigault, Didier, Isabelle Gonzalez-Mullier, and Cécile Gros. 1996. "LEXTER, a Natural Language Processing Tool for Terminology Extraction." In *Proceedings of the 7th European Association for Lexicography International Congress on Lexicography International Congress* (EURALEX 1996): 771–779. Göteborg, Sweden: Göteborg University.

Bourigault, Didier, Christian Jacquemin, and Marie-Claude L'Homme. 2001. "Introduction." *Recent Advances in Computational Terminology* 2, ed. by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, iix–xviii. John Benjamins. https://doi.org/10.1075/nlp.2.01bou

Bouslimi, Riadh, Jalel Akaichi, Mouhamed Gaith Ayadi and Hana Hedhli. 2016. "A Medical Collaboration Network for Medical Image Analysis." *Network Modeling Analysis in Health Informatics and Bioinformatics* 5(1): 1–11.

Carreras, Xavier, Isaac Chao, Lluís Padró and Muntsa Padró. 2004. "FreeLing: An Open-Source Suite of Language Analyzers." In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004). Lisbon, Portugal.

Conrado, Merley S., Thiago A. S. Pardo, and Solange O. Rezende. 2013. "Exploration of a Rich Feature Set for Automatic Term Extraction." *Advances in Artificial Intelligence and Its Applications* 8265: 342–354. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45114-0_28

Dagan, Ido, and Ken Church. 1994. "Termight: Identifying and Translating Technical Terminology." *Proceedings of the 4th Conference on Applied Natural Language Processing*: 34–40. Stuttgart, Germany.

David, Sophie, and Pierre Plante. 1990. "Le progiciel TERMINO : de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes." In *Actes du Colloque international sur les industries de la langue : perspectives des années 1990* 1: 71–88. Montreal, Canada.

Drouin, Patrick. 1997. "Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme." *Meta: Journal des traducteurs* 42(1): 45–54. https://doi.org/10.7202/002593ar

Daille, Béatrice. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Dissertation. Université de Paris 7.

Daille, Béatrice. 1995. *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*. 5. Lancaster, United Kingdom: UCREL Technical Papers.

Daille, Béatrice. 1997. "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology." *The Balancing Act: Combining Symbolic and Statistical Approaches to Language* 1: 49–66. Boston: Massachusetts Institute of Technology.

Dias, Gaël. 2003. "Multiword Unit Hybrid Extraction." In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE 2003) 18: 41–48. Sapporo, Japan.

Dramé, Khadim, Gallo Diallo, Fleur Delva, Jean François Dartigues, Evelyne Mouillet, Roger Salamon and Fleur Mougin. 2014. "Reuse of Termino-ontological Resources and Text Corpora for Building a Multilingual Domain Ontology: an Application to Alzheimer's Disease." *Journal of biomedical informatics* 48: 171–182. https://doi.org/10.1016/j.jbi.2013.12.013

Earl, Lois L. 1970. "Experiments in Automatic Extracting and Indexing." *Information Storage and Retrieval* 6(4): 313–330. https://doi.org/10.1016/0020-0271(70)90025-2

Enguehard, Chantal, and Laurent Pantera. 1995. "Automatic Natural Acquisition of a Terminology." *Journal of Quantitative Linguistics* 2(1): 27–32. https://doi.org/10.1080/09296179508590032

Evans, David A., and Chengxiang Zhai. 1996. "Noun-phrase Analysis in Unrestricted Text for Information Retrieval." In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics* (ACL 1996): 17–24. Santa Cruz, California, USA. https://doi.org/10.3115/981863.981866

Evert, Stefan, and Brigitte Krenn. 2001. "Methods for the Qualitative Evaluation of Lexical Association Measures." In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics:* 188–195.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation. University of Stuttgart.

Fabre, Cécile. 1996. *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations*. Dissertation. Université de Rennes 1.

Fedorenko, Denis G., Nikita Astrakhantsev, and Denis Turdakov. 2013. "Automatic Recognition of Domain-specific Terms: an Experimental Evaluation." In *Proceedings of the Institute for System Programming of the RAS* (ISP RAS) 26(4): 15–23. Russia.

Foo, Jody. 2012. *Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Dissertation. Linköping University.

Frantzi, Katerina T., and Sophia Ananiadou. 1997. "Automatic Term Recognition using Contextual Cues." In *Proceedings of the 3rd DELOS Workshop*: 19–27. Zurich, Suisse.

Gornostay, Tatiana. 2010. "Terminology Management in Real Use." In *Proceedings of the 5th International Conference on Applied Linguistics in Science and Education*: 25–26. Saint Petersburg, Russia.

Heid, Ulrich, and John McNaught. 1991. *EUROTRA-7 Study: Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications*. Final Report. CEC-DG XIII. University of Stuttgart.

Jacquemin, Christian. 1994. "FASTR: A Unification-based Front-end to Automatic Indexing." In *Proceedings of the 4th International Conference on Computer-Assisted Information Retrieval* (Recherche d'information et ses Applications) (RIAO 1994) 2: 34–47. New York, USA: Rockfeller University Press.

Jacquemin, Christian. 1999. "Syntagmatic and Paradigmatic Representations of Term Variation." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL 1999): 341–348. College Park, Maryland, USA.

Jiang, Birong, Endong Xun, and Jianzhong Qi. 2015. "A Domain Independent Approach for Extracting Terms from Research Papers". In *Databases Theory and Applications*. ADC 2015, ed. by Mohamed Sharaf, Muhammad Cheema, and Jianzhong Qi, 155–166. Australia. *Lecture Notes in Computer Science*, vol 9093. Heidelberg, Berlin: Springer.

Justeson, John S., and Slava M. Katz. 1995. "Technical Terminology: some Linguistic Properties and an Algorithm for Identification in Text." *Natural Language Engineering* 1(1): 9–27. https://doi.org/10.1017/S1351324900000048

Kageura, Kyo, and Bin Umino. 1996. "Methods of Automatic Term Recognition: A Review." *Terminology* 3(2): 259–289. https://doi.org/10.1075/term.3.2.03kag

Loukachevitch, Natalia V. 2012. "Automatic Term Recognition Needs Multiple Evidence." In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012): 2401–2407. Istanbul, Turkey.

Liu, Bao, Guiping Zhang, and Dongfeng Cai. 2008. "Technical Term Automatic Extraction Research based on Statistics and Rules [J]." *Computer Engineering and Applications* 44(23): 147–150.

Lossio-Ventura, Juan Antonio, et al. 2014. "Yet Another Ranking Function for Automatic Multiword Term Extraction." In *Advances in Natural Language Processing*. NLP 2014, ed. by Adam Przepiórkowski, and Maciej Ogrodniczuk, 52–64. Poland. *Lecture Notes in Computer Science*, vol 8686. Heidelberg, Berlin: Springer.

Lossio-Ventura, Juan Antonio, et al. 2016. "Biomedical Term Extraction: Overview and a New Methodology." *Information Retrieval Journal* 19(1–2): 59–99.

Maynard, Diana, and Sophia Ananiadou. 1999. "Identifying Contextual Information for Multi-word Term Extraction." In *Proceedings of Terminology and Knowledge Engineering Conference* 99: 212–221. Innsbruck, Austria.

Messaoudi, Abir, Riadh Bouslimi, and Jalel Akaichi. 2013. "Indexing Medical Images based on Collaborative Experts Reports." *International Journal of Computer Applications* 70(5): 1–9. https://doi.org/10.5120/11955-7787

McEnery, Tony, et al. 1997. "The Exploitation of Multilingual Annotated Corpora for Term Extraction." *Corpus Annotation: Linguistic Information from Computer Text Corpora*: 220–230. Boston, MA, USA: Addison Wesley Longman.

Merkel, Magnus, and Mikael Andersson. 2000. "Knowledge-lite Extraction of Multi-word Units with Language Filters and Entropy Thresholds." In *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval* (Recherche d'Information et ses Applications) (RIAO 2000): 737–746. Paris, France.

Miller, George A. 1995. "WordNet: a Lexical Database for English." *Communications of the ACM* 38(11): 39–41. https://doi.org/10.1145/219717.219748

Naulleau, Elie. 1998. *Apprentissage et filtrage syntactico-sémantique de syntagmes nominaux pertinents pour la recherche documentaire*. Dissertation. Université Paris XIII.

Nazarenko, Adeline, and Haifa Zargayouna. 2009. "Evaluating Term Extraction." In *International Conference on Recent Advances in Natural Language Processing* (RANLP 2009): 299–304. Borovets, Bulgaria.

Oliver, Antoni, Salvador Climent, and Joaquim Moré. 2007. *Traducción y tecnologías* 4. Barcelona: Editorial UOC.

Oliver, Antoni, and Mercè Vàzquez. 2015. "TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction." *International Conference on Recent Advances in Natural Language Processing* (RANLP 2015): 473–479. Hissar, Bulgaria.

Padró, Lluís, and Evgeny Stanilovsky. 2012. "FreeLing 3.0: Towards Wider Multilinguality." In *Proceedings of the 8th International Conference on Language Resources and Evaluation Conference* (LREC 2012): 2473–2479. Istanbul, Turkey.

Pazienza, Maria Teresa, Pennacchiotti, Marco, and Zanzotto, Fabio. 2005. "Terminology Extraction: an Analysis of Linguistic and Statistical Approaches." *Knowledge Mining. Studies in Fuzziness and Soft Computing* 185: 255–279. Heidelberg, Berlin: Springer.

Pereira, Rui, Paul Crocker, and Gaël Dias. 2004. "A Parallel Multikey Quicksort Algorithm for Mining Multiword Units." In *Proceedings of the Workshop on Methodologies and Evaluation of Multiword Units in Real-world Application*: 17–23. Lisbon, Portugal.

Piao, Scott S., and McEnery, Tony. 2001. "Multi-word unit Alignment in English-Chinese Parallel Corpora." In *Proceedings of the Corpus Linguistics Conference* 13: 466–475. Lancaster. England.

Smadja, Frank. 1993. "Retrieving Collocations from Text: Xtract". *Computational Linguistics* 19(1): 143–177.

Valaski, Joselaine, Sheila Reinehr, and Andreia Malucelli. 2015. "Approaches and Strategies to Extract Relevant Terms: How are they being applied?" In *Proceedings of the International Conference on Artificial Intelligence* (ICAI 2015): 478–484. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). San Diego, USA.

Vasiljevs, Andrejs, Marcis Pinnis, and Tatiana Gornostay. 2014. "Service Model for Semi-automatic Generation of Multilingual Terminology Resources." In *Proceedings of the Terminology and Knowledge Engineering Conference*: 67–76. Berlin, Germany.

Vàzquez, Mercè, and Antoni Oliver. 2013. "Improving Term Candidate Validation Using Ranking Metrics." In *Proceedings of the 3rd World Conference on Information Technology* (WCIT-2012) 3: 1348–1359. AWERProcedia Information Technology & Computer Science. Barcelona, Spain.

Vàzquez, Mercè. 2014. *Estratègies estadístiques aplicades a l'extracció automàtica de terminologia*. Dissertation. Universitat Pompeu Fabra.

Velardi, Paola, Michele Missikoff, and Roberto Basili. 2001. "Identification of Relevant Terms to Support the Construction of Domain Ontologies." In *Proceedings of the Workshop on Human Language Technology and Knowledge Management – Volume 2001*, 1–8. Association for Computational Linguistics. Morristown, USA.

Vivaldi, Jorge, and Horacio Rodríguez. 2001. "Improving Term Extraction by Combining different Techniques." *Terminology* 7(1): 31–48. https://doi.org/10.1075/term.7.1.04viv

Vivaldi, Jorge. 2009. "Corpus and Exploitation Tool: IULACT and BwanaNet." In *International Conference on Corpus Linguistics* (CICL 2009), *A survey on corpus-based research:* 224–239. Universidad de Murcia, Spain.

Vossen, Piek. 1998. *A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-1491-4

Vu, Thuy, Ai Ti Aw, and Min Zhang. 2008. "Term Extraction through Unithood and Termhood Unification." In *Proceedings of the 3rd International Joint Conference on Natural Language Processing* (IJCNLP 2008) 1: 631–636. Hyderabad, India.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun. 2007. "Tree-traversing Ant Algorithm for Term Clustering based on Featureless Similarities." *Data Mining and Knowledge Discovery* 15(3): 349–381. https://doi.org/10.1007/s10618-007-0073-y

Zheng, Dequan, Tiejun Zhao, and Jing Yang. 2009. "Research on Domain Term Extraction based on Conditional Random Fields." In *International Conference on Computer Processing of Oriental Languages:* 290–296. Heidelberg, Berlin: Springer.

Address for correspondence

Mercè Vàzquez
Universitat Oberta de Catalunya
Faculty of Arts and Humanities
Av. Tibidabo, 39–43
08035 Barcelona
Spain
mvazquezga@uoc.edu

Co-author information

Antoni Oliver
Universitat Oberta de Catalunya
Faculty of Arts and Humanities
aoliverg@uoc.edu