

# Detection of Domain Specific Terminology Using Corpora Comparison

**Patrick Drouin**

Observatoire de linguistique Sens-Texte  
Université de Montréal  
C.P. 6128, succursale Centre-ville  
Montréal (Québec), H3C 3J7  
[patrick.drouin@umontreal.ca](mailto:patrick.drouin@umontreal.ca)

## Abstract

Identifying terms in specialized corpora is a central task in terminological work (compilation of domain-specific dictionaries), but is labour-intensive, especially when the corpora are voluminous which is often the case nowadays. For the past decade, terminologists and specialized lexicographers have been able to rely on term-extraction tools to assist them in the selection of terms. However, most term-extractors focus on the identification of complex terms. Although complex terms (*cellular telephone*) are central to terminology processing, retrieval of uniterms (*telephone*) is still a major challenge. This paper evaluates the usefulness of a corpora comparison approach in order to find pinpoint corpus specific words in order to identify uniterms in the field of telecommunications.

## Introduction

Identifying terms in specialized corpora is a central task in terminological work (compilation of domain-specific dictionaries), but is labour-intensive, especially when the corpora are voluminous which is often the case nowadays. For the past decade, terminologists and specialized lexicographers have been able to rely on term-extraction tools to assist them in the selection of terms. However, most term-extractors focus on the identification of complex terms (Bourigault et al. 2001; Jacquemin 2001). Although complex terms (*cellular telephone*) are central to terminology processing, retrieval of uniterms (*telephone*) is still a major challenge.

From the point of view of most computational linguists, the problem of term extraction is now considered trivial and the problem solved. Current research in the area of computational terminology is mainly aimed towards structuring the output of term extractors so as to access further levels of knowledge (Nazarenko and Harmon 2002). From a user's perspective (terminologist), the problem is far from being solved since all systems lead to a set of results that contains a significant level of noise. The results, although usable to some extent in day-to-day work, need to be refined so as to make it possible for the terminologist to focus on its main task. In that respect, structuring *dirty* data is not best way of helping end-users of the technology. We do not want to underestimate the importance of the research being done on structuring, we consider that it is indeed needed and even critical, but we want to stress the point that precision is still an important issue.

Our main goal is to find a technique that would allow term extractors to discriminate between lexical items, in the form of potential uniterms, that are relevant for terminology work and those which should not be brought to the end-user's attention. In this paper, we describe a technique that relies on corpus comparison in order to identify corpus specific lexicon. We believe

that isolating such a subset of the lexicon of a technical corpus will eventually lead to a precision increase for term extraction of both multiword terms and uniterms. In this study, we will focus on the latter type.

## Previous Work

Quite a bit of work has been done in order to identify words specific to a corpus based on a comparison with a second corpus. Trying to capture the weight of a word (or term in the IR sense) in a collection of documents, Salton (1989) suggested TF.IDF. The test takes into account the frequency of a word and its distribution in various parts of a corpus. Church and Hanks (1990) proposed the Mutual Information (MI) measure aimed at describing co-occurrence phenomena in corpora. Since then, it has been also used in order to pinpoint corpus specific vocabulary (Scott 1997, Kilgariff 2001). Lafon (1980) described a technique, later used by Lebart and Salem (1994) among others, which rely on hypergeometric distribution. Dunning (1993) designed the log-likelihood measure that relies on frequency profiling; the same test was later used by Rayson and Garside (2000). In his search of a method that would make possible comparing corpora, Kilgarriff (2001) experiments with various methods including  $X^2$ , Mann-Whitney rank test,  $t$ -test, MI, log-likelihood, Fisher's exact test and TF.IDF.

Recently, researchers in the field of computational terminology have used corpus comparison to try to capture uniterms or other types of lexical units. Ahmad et al. (1994) and Chung (2003) have used calculations based on normalized frequency; whereas Drouin (2003) resorted to normal distribution as an approximation to the binomial distribution. As underlined by Chung (2003), several differences can be observed in the methods put forward by authors, making them virtually impossible to compare to each other.

## Methodology

This paper is based on the work described in Drouin (2003) where we put forward a two-stage term extraction methodology. The technique uses a subset of the lexicon of a technical corpus, called the analysis corpus (AC), in order to gain access to the terms (uniterms and complex terms) of the same corpus. In order to make this lexical subset stand out, the behaviour of the lexicon is compared to the one of a larger corpus called the reference corpus (RC). With the current paper, we want to evaluate the relevance of the lexical subset identified automatically from the point of view of the terminologist. Since our goal is to eventually use it as the starting point for terminology extraction, we want to make sure this standing ground is solid.

Our technique relies on a statistical comparison of the frequencies observed in corpora that have different properties in order to automatically bring out domain specific terminology. A list of specific words is created by Termostat (Drouin 2003), a piece of software used for term extraction which determines the specificity of words in the AC based on a technique put forward by Lafon (1980). The statistical measure simply takes into account frequencies as observed in both corpora in order to quantify the deviation from a normal distribution. In order to keep only the most significant forms, we selected a threshold that ensures that there is less than one chance out of 1,000 that the frequency observed in the AC is coincidental.

## Corpora

As we previously mentioned, a reference corpus and an analysis corpus are compared. The goal is to identify the subset of the lexicon specific to the AC. So as to evaluate the stability of the approach, we repeated the experiment on a set of three ACs.

### Reference corpus

The reference corpus is composed of 13,746 articles taken from *The Gazette*, a Montreal based newspaper. These articles total up to approximately 7,400,000 tokens which correspond to roughly 82,700 word forms. The size of the RC, although still modest, can guarantee that the articles discuss a wide range of subjects and that their content is, to some extent, heterogeneous. The selected articles were published over a period of three months. We realize that due to the small period of time covered by the RC it is possible that certain topics might be over represented in the RC, but we take for granted that it will not greatly influence the results of our experiments. Although specialized in its nature (journalistic), we consider the RC to be a reflection of the non-technical usage.

### Analysis corpus

For comparison purposes, we will be analyzing three technical corpora referred to as AC<sub>1...3</sub>. The content of these corpora is homogenous and contains documents

pertaining to the field of telecommunications. To be exact, AC<sub>1</sub> discusses the programming interface of fiber optic networks. Its intended readers are programmers. In the case of AC<sub>2</sub> and AC<sub>3</sub>, they talk about the physical properties of the same networks and are written for installers and technicians.

Corpus	Tokens	Word forms
AC <sub>1</sub>	11,947	1,207
AC <sub>2</sub>	28,583	2,066
AC <sub>3</sub>	8,676	1,053

Table 1: Size of the corpora

From information contained in Table 1, the reader will notice that these corpora are rather small. The size of the ACs was determined by the original intent of our research. Since we want to provide a methodology that will help terminologists in their day-to-day work, we decided to use documents that are representative of the ones mined manually by a terminology team. In that regard, all ACs are made up of only one document. During the process, the ACs are dynamically merged with the RC in order to create a global corpus (GC). In other words, the ACs is considered to be a sample of the CG.

All corpora were first tokenized and then tagged with Brill's rule-based part-of-speech tagger (Brill 1992, 1994). The tagging process was done without training and the results of the tagging are used as-is. In that respect, the results we obtain from subsequent modules could only be better if the output of the tagger was corrected and the software trained. Termostat performs root-form analysis on the noun of the corpus in order to work with lemmas. A simple, yet effective set of 8 heuristic rules was used and led to a good analysis in 98.7% of the cases.

## Validation Process

So as to evaluate the quality of the output of the first stage of Termostat (identification of corpus specific words), the second being the term extraction by itself, the results were submitted to a two-step validation process.

### Automatic validation

The first step is an automated validation which consists of a comparison of the identified subset of the lexicon with a list of terms found in a terminology database dedicated to the field of telecommunications. The multilingual (English, French, Spanish, German, Portuguese, Chinese and Japanese) terminology database contains approximately 100,000 terms. During the validation process, we only exploited the English subset of the data which roughly amounts to 61,000 entries.

We are aware that this evaluation procedure has a number of weaknesses, since specialized dictionaries are not always built on observations derived from corpora and their contents rely on editorial decisions made by terminologists. However, we are interested in assessing the value of our comparisons in a terminological setting. We believe that the contents of specialized terminology databases are a reflection, albeit imperfect, of the needs of terminologists.

## Human validation

The list was then submitted to a team of three terminologists (two junior, one senior), specialists of the telecommunications industry. The instructions given to the team were to consider an entry as valid when it met the following two criteria:

- it is representative of the domain,
- it is representative of the main topic of the corpus.

So as to evaluate the adequacy to meet the previous criteria, a Web interface was provided to the team. TermoStat was modified so as to store the results of its first stage in a database system. For this experiment, the open-source MySQL RDBS backend was used while the Web pages were dynamically generated using the PHP scripting language.

Validate	Reject	Candidate	AbsFreq
Validate	Reject	access	2
Validate	Reject	account	1
Validate	Reject	activation	1
Validate	Reject	adm	2
Validate	Reject	alarm	6
Validate	Reject	alarmchoice	6
Validate	Reject	alarmchoice	6
Validate	Reject	alignmate	1
Validate	Reject	ambb	1
Validate	Reject	anel	2
Validate	Reject	api	1
Validate	Reject	appendix	3
Validate	Reject	application	5
Validate	Reject	architecture	2
Validate	Reject	array	3
Validate	Reject	atm	2
Validate	Reject	attribute	5
Validate	Reject	attributeid	19

Figure 1: List of corpus specific words

The interface used by the terminologists is rather simplistic and it allowed them to browse the list of lexical items as well as their frequency (Figure 1) and to have access to the entire list of contexts for a given item on the list (Figure 2).

Context
information on alarm mapping for Internet Protocol ( IP ) access network elements , refer to IP Access device Performance Management
: For the period covered by the service package , access to regular technical support for all issues defined in the

Figure 2: Contexts displayed for the word *access*

The terminologists could either *Validate* or *Reject* the entry based on their knowledge of the corpus and subject area. Each member of the team was given a list of words extracted from one of the ACs to review. After the initial review, a final pass was done by the most senior terminologist for the three corpora.

## Results

Our evaluation of the results focuses solely on precision. We believe that, for term extraction, recall is not a major problem. As far as we are concerned, the key challenge still is to be able to discriminate between relevant and irrelevant entries in a list of candidates. Recall is also hard to evaluate since it requires corpora mined manually to be used as gold standard and such corpora were not available for this study (and are hardly ever available).

	AC <sub>1</sub>	AC <sub>2</sub>	AC <sub>3</sub>
<b>Relevant entries</b>	444	810	273
<b>Irrelevant entries</b>	84	131	101
<b>Total</b>	518	941	374
<b>Precision</b>	84,1%	86,1%	73,0%

Table 2: Precision

Lexical items that did not classify as specific items because they did not reach the probability threshold were not evaluated by the terminologists. Therefore, it is possible that some valid items were not identified by the software.

The validation process showed that 84.1% of the words identified in AC<sub>1</sub> were considered to be relevant; it was also the case for 86.1% of the words extracted from AC<sub>2</sub> and 73.0% from AC<sub>3</sub>. These measures were taken on the list of specific words without a minimal frequency threshold. The precision level is surprising high if we consider that, as pointed out by Dunning (1993) and Labbé and Labbé (2001), the calculation we used does not cope very well with low frequencies and most entries (roughly 55% to 60%) in the list of specific words have a frequency below 5. In other words, from a statistical point of view the data might not be the most reliable source but from a terminological point of view it is still very useful.

As with any lexical based techniques, homography and polysemy can cause problems. During processing of our documents, some domain specific words were ignored by the system. For example, in the case of our analysis corpora, words such as *time*, *process*, *manager*, *exchange*, *state*, *manager*, were left out although they belong to the terminology of telecommunications. One could argue that the specificity of these words is rather semantic and not purely lexical. Obviously, their specificity cannot be observed strictly from a lexical point of view since we were not able to identify them as being domain specific.

Multiple phenomena come into play in this case, including homography, polysemy and de-terminologization (Galissou 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.

## Conclusion and Future Work

We presented a method that compares word frequencies by opposing technical and non-technical corpora. The level of precision obtained indicates that the corpus specific words are useful for day-to-day terminology work. We believe that, in the future, this subset of the lexicon of technical corpora can be used as a starting point to retrieve domain specific terminology and increase precision of term extraction software.

Although we have not yet performed tests to prove that the level of precision obtained can be attributed to the radically different nature of the corpora at hand, we think that it plays a strong part in the quality of the results obtained. Further experiments are thus needed in order to determine if more heterogeneous reference corpora would have an influence on the precision level. We also plan on doing more tests with ACs taken from various domains.

Other statistical methods as the ones described earlier in this paper could probably lead to interesting results. Some of them tend to give better results on lower frequencies while others work better for the other end of the spectrum of frequencies. It would be interesting to compare them with the technique we use and to quantify the impact on the precision level.

Even though we decided not to evaluate recall for this particular project, we realize that it is an issue that needs to be addressed when dealing with techniques that are aimed at increasing precision.

## References

- Ahmad, K., Davies A., Fulford H. and Rogers M. (1994). What's in a Term? The semi-automatic Extraction of Terms from Text. In *Translation Studies. An Interdiscipline*, John Benjamins, 267-278.
- Bourigault, D., Jacquemin C. and M. C. L'Homme (editors) (2001). *Recent Advances in Computational Terminology*. John Benjamins.
- Brill, E. (1994). Some Advances in Transformation-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, 722-727.
- Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied-Natural Language Processing ANLP-1992*, 152-155.
- Chung, T. M. (2003). A Corpus Comparison Approach for Terminology Extraction. In *Terminology*, 9(2), 221-246.
- Church, K. and P. Hanks. (1990). Word Association Norms, Mutual Information, and Lexicography, In *Computational Linguistics*, 16(1), 22-29.
- Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. In *Terminology*, 9(1), 99-115.
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence, In *Computational Linguistics*, 19(1), 61-74
- Galisson, R. (1978). Recherches de lexicologie descriptive : la banalisation lexicale, Nathan.
- Huizong, Y. (1986). A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts, In *Literary and Linguistic Computing*, 1(2), 93-103.
- Jacquemin, C. (2001). Spotting and Discovering Terms through Natural Language Processing Techniques, MIT Press.
- Kilgarriff, A. (2001). Comparing corpora. In *International Journal of Corpus Linguistics*, 6(1), 1-37.
- Labbé, C. and Labbé D. (2001). Que mesure la spécificité du vocabulaire?, In *Lexicometria*, 3, 23 p.
- Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, In *MOTS*, 1, 128-165.
- Lebart L. et Salem A. (1994). Statistique textuelle, Dunod.
- Meyer, I. and K. Mackintosh (2000). When terms move into our everyday lives: An overview of de-terminologization. In *Terminology*, 6(1), 111-138.
- Nazarenko, E and T, Harmon. (editors) (2002). Structuration de la terminologie, In *Traitement automatique des langues*, (43)1.
- Rayson, P. and R. Garside (2000). Comparing corpora using frequency profiling, In *Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, 1-6.
- Salton, G. (1989). Automatic text processing: the transformation, analysis and retrieval of information by computer, Addison Wesley.
- Scott, M. (1997). PC analysis of key words - and key key words. In *Systems*, 25(1), 233-345.