Lexical Profiling of Environmental Corpora

Patrick Drouin, Marie-Claude L'Homme, Benoît Robichaud
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec) H3C 3J7 CANADA
{patrick.drouin, mc.lhomme, benoit.robichaud}@umontreal.ca

Abstract

This paper describes a method for distinguishing lexical layers in environmental corpora (i.e. the general lexicon, the transdisciplinary lexicon and two sets of lexical items related to the domain). More specifically we aim to identify the general environmental lexicon (GEL) and assess the extent to which we can set it apart from the others. The general intuition on which this research is based is that the GEL is both well-distributed in a specialized corpus (criterion 1) and specific to this type of corpora (criterion 2). The corpus used in the current experiment, made of 6 subcorpora that amount to 4.6 tokens, was compiled manually by terminologists for different projects designed to enrich a terminological resource. In order to meet criterion 1, the distribution of the GEL candidates is evaluated using a simple and well-known measure called *inverse document frequency*. As for criterion 2, GEL candidates are extracted using a term extractor, which provides a measure of their specificity relative to a corpus. Our study focuses on single-word lexical items including nouns, verbs and adjectives. The results were validated by a team of 4 annotators who are all familiar with the environmental lexicon and they show that using a high specificity threshold and a low idf threshold constitutes a good starting point to identify the GEL layer in our corpora.

Keywords: terminology, lexical layers, term extraction, corpora, environment

1. Introduction

It is generally recognized that specialized texts comprise three main lexical layers: 1. terminology (the lexicon used to express domain-specific knowledge); 2. general language (the lexicon used by all speakers of a language and that is likely to be found in any kind of texts); and 3. a layer that lies in-between that will be called herein the *transdisciplinary lexicon* (Drouin, 2007; Tutin, 2008; Hatier, 2016)[1]. We believe that in very large domains, such as the environment that encompasses a broad variety of topics (climate change, sustainable development, renewable energy, water pollution, etc.), the terminology (defined above as the "domain-specific lexicon") further divides into two layers. The first layer of the lexicon is topic specific. For instance, terms such as *chlorination* or *marine turbine* are specific to water pollution and renewable energy respectively. The second layer of the lexicon cuts across the entire field of the environment: e.g. *ecosystem*, *sustainable*, *energy*, *development*, etc. We would thus obtain four different lexical layers in specialized texts, as shown in Figure 1.

In given applications (such as terminology resource compilation for which the method proposed in the paper is investigated)[2], identifying items that belong to one layer or the other can be quite difficult. For example, when working with a general environment corpus (such as PANACEA[3] that covers a wide range of topics), some topic specific terminology might be difficult to spot since the corpus will cover several specialized topics related to the overall domain. In contrast, when working with topic specific corpora, some general domain terminology might not be perceived as such since the corpus does not offer broad view of the subject.

For the time being, compilers of resources make decisions based on their intuition, but this can lead to choices that differ from one compiler to another. Furthermore, specialized resources are not necessarily enriched by experts of a domain (in fact, they seldom are). So making fine-grained distinctions between topic specific, general specialized or transdisciplinary lexica can soon become a quite challenging task.

This paper proposes a method for identifying one of the layers mentioned above, i.e. the general environmental lexicon (GEL). In the process, however, we will need to distinguish this lexicon from topic specific lexica, on the one hand, and from the transdisciplinary lexicon, on the other hand.

The general intuition on which this research is based is that the GEL is both well-distributed in a specialized corpus (criterion 1) and specific to this type of corpora (criterion 2). In order to meet criterion 1, distribution of the GEL candidates is evaluated using a simple and well-known measure called inverse document frequency. As for criteria 2, GEL candidates are extracted using a term extractor, which provides a measure of their specificity relative to a corpus.

2. Related Work

Different methods were devised to identify terminology and the transdisciplinary lexicon. Regarding term extraction, methods are now well established and used for different applications (Indurkhya and Damerau, 2010). An efficient method consists of comparing a domain-specific

[1] Other names for this specific layer can be found in the literature: e.g., academic vocabulary (Coxhead, 2000; Paquot, 2014)

[2] There are other applications for which distinguishing lexical layers is important: specialized translation, language teaching, for instance.

[3] `http://catalog.elra.info/productinfo.php?products_id=1184`, ELRA-W0063

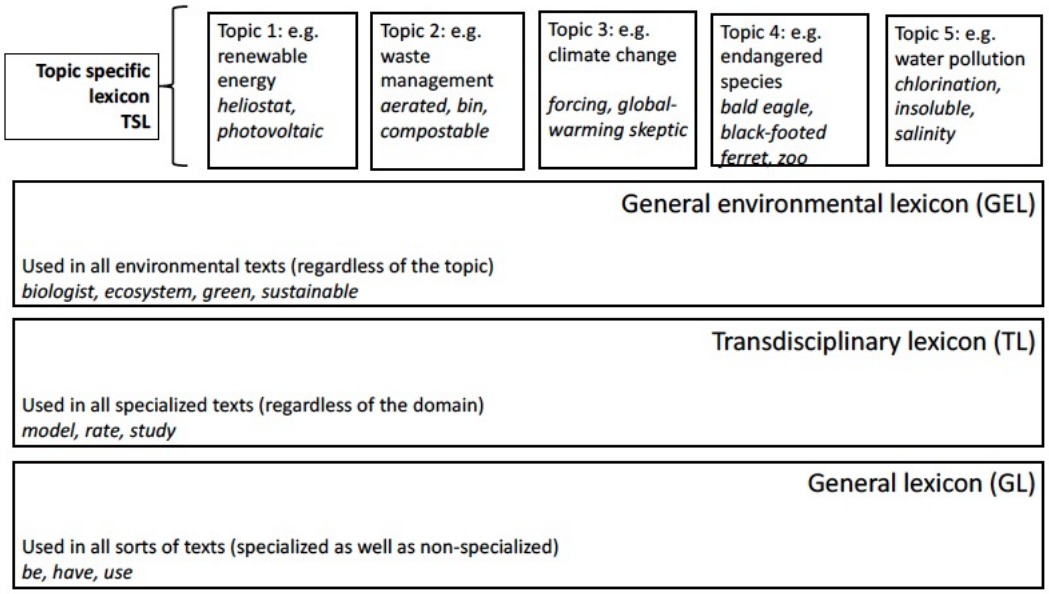

Figure 1: Lexical layers in environmental texts.

corpus to a general one and computing a specificity score[4] of lemmas. For instance, a corpus of environmental texts can be compared to a general balanced corpus such as the British National Corpus. This method was implemented in TermoStat developed by (Drouin, 2003). It was evaluated for the extraction of single-word terms with satisfactory results (Lemay et al., 2005) and supports multiple languages[5]. The concept of "specificity" aims to capture the potential of term candidates to behave like terms (termhood, see (Kageura and Umino, 1996)). In most cases, termhood is linked to a higher than expected frequency in a specialized corpus based on a theoretical frequency computed from a general corpus. Various statistical measures can be used to compute specificity. Such an approach gives us access to the topic specific layer (TSL).

Over the years, methods have also been developed for the identification of the transdisciplinary lexicon (TL) (Drouin, 2007; Tutin, 2008; Hatier, 2016). This second set of lexical items can also be identified with corpus comparison with a general corpus in order to identify this lexical layer. In such a case, however, the corpus that is analyzed should cover various disciplines and be composed of several topic specific corpora such as physics, chemistry and linguistics. Previous work has shown that identifying the transdisciplinary lexicon raises challenges due to different factors such as polysemy of lexical items, interference with other layers.

Since term extraction techniques are targeted at the identification topic specific lexical items solely, they cannot be used as-is and they have to be slightly modified. For our proposed task, the strategy used to identify TL lexical items cannot be considered either as we have a corpus covering one domain, namely the environment. What we need is a technique that can capture that fact that the GEL lexical items are both related to the overall topic of a corpus (thus semantically close to TSL items) and transdisciplinary as far as the overall topic of the corpus is concerned (from this point of view, their behavior bears some similarities with TL items).

Our hypotheses for the current task are that:

1. Lexical items of the TSL should be associated with high specificity measures when compared to a balanced general reference corpus as they are characteristic of the overall subject area. Furthermore, TSL members should have a low distribution across the corpus since topics are addressed in subcorpora.

2. Lexical items that belong to the GEL should also be associated with a high specificity measure when compared to a balanced general reference corpus on the one hand. On the other hand, they should have a large distribution across different subcorpora since they are associated with the environment as an overall domain.

3. Members of the TL should have lower specificity levels as the TLS items since they also occur on a regular basis in a balanced general reference corpus. We expect them, as demonstrated in prior studies, to be highly distributed across the corpus.

4. Common words, or lexical units of the General Lexicon should be both distributed in the corpus and have low specificity values.

3. Method

Our method aims to identify the GEL (2. above). In order to do so, we will apply two criteria designed to model our hypotheses: the first, criterion 1, aims to capture distribution; the second, criterion 2, captures specificity. Distribution is evaluated on the specialized corpus while specificity computation requires that we use two corpora: a general bal-

[4]The concept of *specificity* used in this paper differs from the usage of the nearby concept in the medical context where it is used as a measure of false positive rate.

[5]`http://termostat.ling.umontreal.ca`

anced corpus and a specialized corpus. Figure 2 illustrates the overall process used to reach our goals.

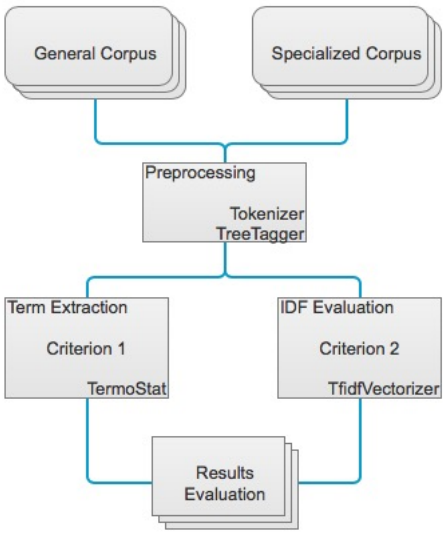

Figure 2: Overview of the method to identify the general environmental lexicon (GEL).

The following sections detail our experimental setup, including measures, tools and the annotation process and scheme.

4. Experimental Setup

4.1. Corpus Data

4.1.1. Specialized Corpora

The specialized corpora used in the current experiment were compiled manually by terminologists for different projects designed to enrich a terminological resource (DiCoEnviro) (L'Homme, 2018). Table 1 gives an overview of the subcorpora combined in order to build our specialized corpus.

Subcorpora	Number of tokens
Climate Change	607,233
Endangered Species	1,276,304
Renewable energy	776,838
Transportation Electrification	747,389
Waste management	626,039
Water pollution	586,849
Total	4,620,652

Table 1: Size of the subcorpora.

4.1.2. General Corpus

The general reference corpus used was built from subsets of two large corpora: the British National Corpus (BNC) (Consortium, 2007) and the American National Corpus (ANC) (Reppen et al., 2005). We extracted 4M tokens from each of these corpora in order to compile our 8M tokens reference corpus.

4.2. Corpus Preprocessing

Basic preprocessing was applied to both the specialized and the reference corpora, which included extracting the text from the XML files that comprise the corpus, replacing non-ASCII characters with ASCII equivalents and tokenizing. The corpora are then tagged and lemmatized using TreeTagger (Schmid, 1994).

4.3. Term Extraction and Specificity Evaluation

Terms[6] were extracted using a modified version of TermoStat (Drouin, 2018) in order to use a general reference corpus designed for this specific experiment. The extraction process was limited to single-word lexical items including nouns, verbs and adjectives.

As mentioned, TermoStat computes a *Specificity* score to represent how far the frequency in the specialized corpus deviates from a theoretical frequency. In order to do so, a measure proposed by Lafon (1980) is used.

	Reference Corpus	Specialized Corpus	Total
Freq. term	a	b	a+b
Freq. of other words	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Table 2: Contingency table of frequencies.

Using values from Table 2, specificity can be calculated as follows:

$$\log P(X=b) = \log (a+b)! + \log (N-(a+b))! + \log (b+d)! + \log (N-(b+d))! - \log N! - \log b! - \log ((a+b)-b)! - \log ((b+d)-b)! - \log (N-(a+b)-(b+d)+b)!$$

This measure has been tested in previous studies (Lemay et al., 2005; Drouin and Langlais, 2006; Drouin, 2006; Drouin and Doll, 2008) and leads to excellent results for both the extraction of single-word terms and multi-word terms. Specificity allows identifying forms that are both over- and under- represented in a corpus. In the case of terminology, a domain- and genre-oriented lexicon, we are solely interested in positive specificities which correspond to forms that over-represented.

Although it is a common practice when dealing with domain-specific units to extract multi-word terms and especially multi-word nouns, we apply criteria that are more compatible with lexicography. Hence, items such as *climate, pollute, green* and *greenhouse effect* are considered as terms; expressions such as *climatic impact* and *renewable energy* are considered as compositional collocations. Since most multi-word expressions are compositional in specialized corpora, it is much more productive for terminologists in our projects to work with lists of single-word lexical items. The drawback of this method is, of course, to

[6]We are using *term* here to describe the output of the term extractor. In fact, this output will encompass both topic-specific lexical items and GEL members.

potentially raise more difficulties when trying to separate the lexical layers to which we refer in the present paper.

Since the specificity scores cannot be represented on a predefined scale, we expressed them on a scale ranging from 0 to 100 where the max specificity score is mapped to 100. This mapping leads to a less granular representation of the scores and a more flexible set of scores to assess.

4.4. Inverse Document Frequency Evaluation

In order to evaluate the distribution of the GEL candidates we used the simple and well-known measure called *inverse document frequency* (Sparck Jones, 1972). This measure returns lower scores for tokens that occur very frequently in a document set, and contrariwise higher scores for tokens that occur rarely. To compute idf, we used its Python implementation (TfidfVectorizer) from the Python scikit-learn library. For our study, default values were used and sentences were considered as documents. As with the previous measure, idf scores were also mapped on a scale of 0 to 100. However, in the case of idf, we reverse the score so that the most "interesting" GEL candidates for our study receive a higher idf. This modification was applied to make the scoring results more intuitive for the team of annotators.

4.5. Annotation of results

4.5.1. Result sampling

Since the volume of GEL candidates identified was too large for our team to proceed to a complete validation, we resorted to a sampling mechanism. In order to do so, we broke down both the idf and the specificity scores in groups of 10 ranging from 0 to 100. The results were then sorted by decreasing order of idf and decreasing specificity scores providing us with a matrix of results of size 10x10. The lower left corner corresponds to a mapped idf score of 0-9 and a mapped specificity score in the same range. At the opposite side, the upper right corner of the matrix contains GEL candidates with mapped idf and specificity scores of 90-100. From each of the cell of the matrix, we sampled a maximum of 15 GEL candidates, which means we could evaluate a theoretical maximum number of 1,500 GEL candidates. In fact, since not all cells contain 15 candidates, our process led to a total of 522 GEL candidates to be evaluated.

4.5.2. Annotation team

A team of 4 annotators who are all familiar with the environmental lexicon were responsible for carrying out the annotation process. They have varying experience in enriching a terminological resource that contains terms related to the different topics mentioned in Table 1.

4.5.3. Annotation guidelines

Since the task given to annotators was to single out the GEL – and thereby distinguish it from the TSL, on the one hand, and from the TL, on the other – annotators held a discussion to agree on a definition for each lexical level. They also defined very broad classes of terms that in their opinion are relevant for characterizing the GEL:

- Related to nature (*ecosystems, species*)
- Related to Earth and to its subdivisions (*ocean, continent, hemisphere*)
- Human impact on nature and human activities (*agriculture, activity, defiorest*)
- Products made by humans; things produced by humans (*chemical, waste*)
- Greenhouse gases and related concepts (*carbon, methane, emit*)
- Pollution and contamination (*contaminated, pollutant*)
- Climate/weather and meteorological events (*cyclone, extreme*)
- Protection and conservation (*endangered, protect*)
- General scientific domains (*biology, chemistry*) and experts (*biologist*)

Afterwards each annotator proceeded to validate the list of candidates separately. They could use different resources (terminological databases and corpora), but they could not consult each other during the validation process.

4.5.4. Annotation scheme

In order to obtain optimal results, we decided to use a simple annotation scheme where annotators classified GEL candidates in four different categories represented by a single letter. Keeping in mind that the "good" candidates are those that belong to the GEL), our scheme includes:

- B: the candidate is part of the GEL. *energy, emission, temperature, water, waste*
- M: the candidate is not part of the GEL. *cell, high, include, show, year*
- I: the candidate is part of the vocabulary of the environment; however, the annotator hesitates to classify it as topic specific or as part of the GEL. *model, range, turbine, wave, wind*
- P: the candidate is not valid. *bacterium, recharg, semi, specie, trolleybuses*

All GEL candidates proposed to the annotators had to be classified using the 4 previous codes. The P code is used to classify all forms that are mainly related to tokenizing errors and NLP errors (for example, erroneous part-of-speech tagging). Items classified using the M code could either be members of TSL, TL or general language (GL) and relevant for the current study which is solely focused on GEL.

5. Results and evaluation

5.1. Results

The extraction process on our 4.6M word specialized corpus led an impressive amount of GEL candidates. Table 3 gives an overview of the results broken down by part-of-speech.

Part of speech	Number of GEL candidates
Nouns	11,725
Adjectives	4,817
Verbs	1,722
Total	18,265

Table 3: Number of GEL candidates by part-of-speech.

5.2. Inter-annotator agreement results

The inter-annotator agreement was evaluated using a free online tool (Geertzen, 2012), which provides both the Fleiss kappa (Fleiss and others, 1971) and the Krippendorff's alpha (Krippendorff, 2004) scores (See Table 4). Detailing these measures is beyond the scope of this paper, but both measures consider pairwise agreement of the annotators.

Fleiss	Krippendorff
A_obs = 0.797	D_obs = 0.203
A_exp = 0.471	D_exp = 0.529
Kappa = 0.616	Alpha = 0.616

Table 4: Inter-annotator agreement.

Although both scores indicate that our annotators are not in total agreement, they lead us to believe that the agreement level is nevertheless fairly high.

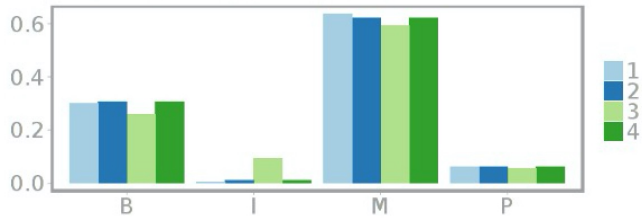

Figure 3: Inter-rater agreement evaluation.

Figure 3 clearly shows that agreement is higher for items that are not part of the GEL (M in Figure 3). We can also note that one of the annotators (the more experienced one) had more problems classifying some candidates than others ("I" in Figure 3). This is an interesting fact and it leads us to believe that more experienced annotators might be more cautious in their classification process.

In order to assess the suitability of the indices to identify the lexical items that interested us, we measured the accuracy of each index for a group of specificity and idf scores. Precision is usually defined as the fraction of relevant instances among the retrieved instances. In other others words, in our case, it corresponds to the number of GEL entries in each group compared to the total of entries in each group.

Figure 4 indicates that the specificity scores are useful to identify terminologically interesting lexical items. However, for our current goal, which is to identify GEL entries, the usefulness of this measure is mitigated by the fact that valid candidates are scattered throughout the score range. This is in line with our hypotheses that specificity scoring

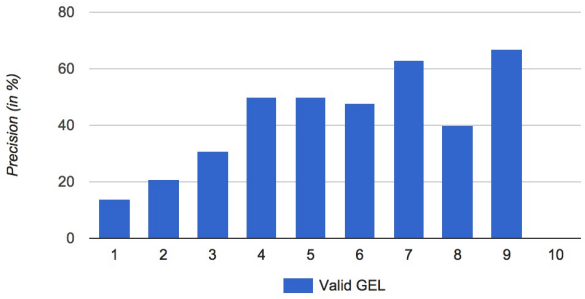

Figure 4: Precision for each group of specificity scores.

cannot, by itself, allow to identify precisely GEL entries from a list of candidates.

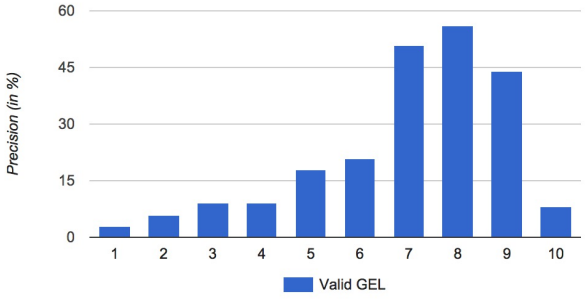

Figure 5: Precision for each group of idf scores.

In order to complete the information provided by the specificity scores, we resorted to using idf. As Figure 5 shows, higher distribution (higher values in our figure correspond to lower original idf scores) is obviously linked to the identification of valid GEL entries in a list of candidates. This observation is also in line with our initial hypotheses. The heatmap in Figure 6 combines both scores in the 10x10 matrix used for the sampling and evaluation. Some of the cells of the matrix contained no candidate and are thus empty (light green). All non-empty cells contain a precision score and are color-coded: red cells have a precision of 0 while green cells have various levels of precision with higher precision levels being darker.

As one can see in Figure 6, most of our candidates are distributed in cells 1-7 for the specificity score and 2-9 for the idf score. Our results show that our valid GEL items are mainly located in the range of specificity 4-9 and the idf 7-10. The relation between higher specificity scores and idf scores can be clearly seen as higher idf scores[7] allow to complete the information provided by specificity.

Figure 7 contains the details of the precision measures for Figure 6. Each cell where data was retrieved shows a ratio of the number of valid GEL items over the number of items in the same group. As can be observed, higher specificity leads to a lower number of candidates while the same observation cannot be made about idf. Restricting our results to high specificity (4-9) and high idf (7-10) values would mean discarding quite a few valid GEL items (88 total). On

[7] We need to remind our readers that our idf scores are reversed from the original idf measure. See section 3.3.2

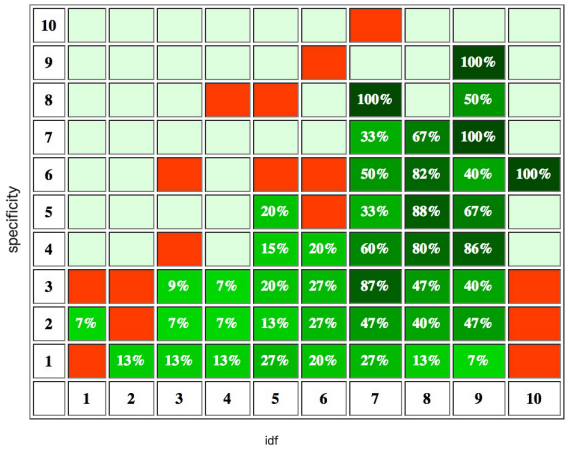

Figure 6: Specificity - idf heatmap - precision.

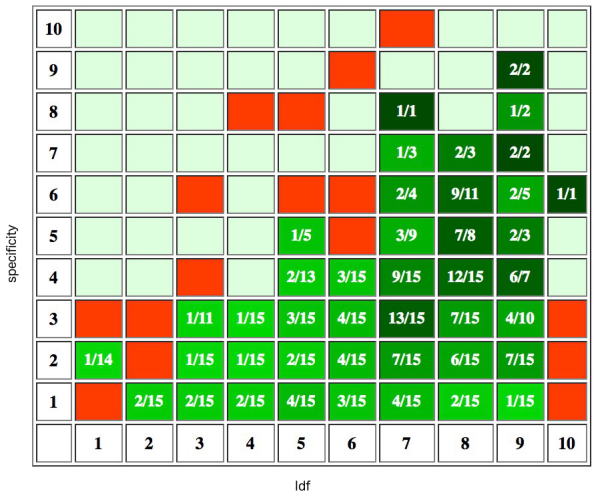

Figure 7: Specificity - idf heatmap - ratio.

the other hand, this would mean that we can obtain a precision of 68% for the same area of the matrix above, which is an interesting performance. Our specificity measure seems to consider far too few GEL items as being specific to our environment corpus.

6. Future Work

Although we limited our investigation to single-word lexical items for the current project, it could be easily applied to multi-word lexical items. In fact, this is not in itself a limitation our approach as much as a methodological decision on our part based on the terminological work being done in our research group. One avenue that could be explored is to measure the impact or the benefit of taking into consideration multi-word lexical items on the validation process.

As could be seen from the inter-annotator evaluation, the annotators seem to strongly agree on what is and what is not a valid GEL item. This was a surprising result given the difficulty of the task and the overlap between lexical that is often assumed by researchers. We would like to investigate what led to that strong agreement in order to see if an algorithm could somehow capture this knowledge. If so, it could be built into further experiments so as to increase precision and complement the method reported in this paper. Idf scores allow us to capture the behaviour of the GEL items adequately while the specificity scores do not seem to be a good indicator as valid forms are scattered throughout the specificity groups. Using a different measure to model the concept of specificity might lead to better results.

Our validation process was carried out using a sample of 15 GEL candidates taken from each cell of our 10x10 matrix. Using a larger number of candidates from each cell might allow us to observe more accurate precision levels. The method was tested on corpora linked to the domain of the environment, a domain that is quite unique since it encompasses a wide variety of topics. An interesting extension would be to test our method with corpora from other domains and see if we can obtain similar results. We will also devise a methodology for implementing this method (in this form or in a modified version) in the compilation process of terminological resources.

7. Conclusion

In this paper we proposed a method to automatically distinguish terminologically relevant lexica in the subject area of the environment. More specifically, we devised a technique to identify the general environmental lexicon (GEL) and distinguish it from other lexical layers that co-exist in specialized corpora. Our basic hypotheses were that lexical items from the GEL were both very specific to our environmental corpus and distributed evenly throughout the same corpus. In order to verify these hypotheses, we used a term extractor relying on the specificity score proposed by Lafon (1980) (criterion 1), and a reversed standard idf measure to quantify the distribution of GEL candidates (criterion 2). Our results validated our hypotheses to a large extent and that candidates with both a higher specificity level and a higher distribution tend to be lexical items of the GEL.

8. Acknowledgements

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada.

We also would like to thank the annotators who contributed to the validation of the GEL candidates.

9. Bibliographical References

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2):213–238.

Drouin, P. and Doll, F. (2008). Quantifying termhood through corpus comparison. In *Terminology and Knowledge Engineering (TKE-2008)*, pages 191–206, Copenhague, Danemark, Août. Copenhagen Business School, Copenhagen, Copenhagen Business School, Copenhagen.

Drouin, P. and Langlais, P. (2006). Évaluation du potentiel terminologique de candidats termes. In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles. (JADT 2006)*, pages 389–400, Besançon, France.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Drouin, P. (2006). Termhood experiments: quantifying the relevance of candidate terms. *Modern Approaches to Terminological Theories and Applications*, 36:375–391.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2):45–64.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Geertzen, J. (2012). Inter-rater agreement with multiple raters and variables. `https://nlp-ml.io/jg/software/ira/`. September 28, 2017.

Hatier, S. (2016). *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*. Ph.D. thesis. Thèse de doctorat dirigée par Tutin, Agnès Sciences du langage Spécialité Informatique et sciences du langage Grenoble Alpes 2016.

Indurkhya, N. and Damerau, F. (2010). *Handbook of Natural Language Processing, Second Edition*. Chapman & Hall/CRC machine learning & pattern recognition series. CRC Press.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage.

Lemay, C., L'Homme, M.-C., and Drouin, P. (2005). Two methods for extracting specific single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics*, 10(2):227–255.

Paquot, M. (2014). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. Corpus and Discourse. Bloomsbury Publishing.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Tutin, A. (2008). Sémantique lexicale et corpus : l'étude du lexique transdisciplinaire des écrits scientifiques. *Lublin Studies in Modern Languages and Literature*, (32):242–260.

10. Language Resource References

BNC Consortium. (2007). *British National Corpus, version 3 BNC XML edition*. British National Corpus Consortium, ISLRN 143-765-223-127-3.

Drouin, Patrick. (2018). *TermoStat 3.0*. http://termostat.ling.umontreal.ca.

L'Homme, Marie-Claude. (2018). *DiCoEnviro : Le dictionnaire fondamental de l'environnement*. http://olst.ling.umontreal.ca/dicoenviro.

Reppen, Randi and Ide, Nancy and Suderman, Keith. (2005). *American National Corpus (ANC) Second Release*. Linguistic Data Consortium, ISLRN 797-978-576-065-6.