A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents

Ayla Rigouts Terryn, Véronique Hoste, Els Lefever
LT3 Language and Translation Technology Team,
Department of Translation, Interpreting and Communication, Ghent University
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
{ayla.rigoutsterryn, veronique.hoste, els.lefever}@ugent.be

Abstract

Terms are notoriously difficult to identify, both automatically and manually. This complicates the evaluation of the already challenging task of automatic term extraction. With the advent of multilingual automatic term extraction from comparable corpora, accurate evaluation becomes increasingly difficult, since term linking must be evaluated as well as term extraction. A gold standard with manual annotations for a complete comparable corpus has been developed, based on a novel methodology created to accommodate for the intrinsic difficulties of this task. In this contribution, we show how the effort involved in the development of this gold standard resulted, not only in a tool for evaluation, but also in a rich source of information about terms. A detailed analysis of term characteristics illustrates how such knowledge about terms may inspire improvements for automatic term extraction.

Keywords: terminology extraction, gold standard, comparable corpora

1. Introduction

Automatic term extraction (ATE) has been a productive and successful field of research within natural language processing, yet the evaluation of ATE remains particularly difficult. The main difficulty of the task lies in the ambiguous nature of terms; there are no objective rules to distinguish terms from non-terms. While there are many definitions or terms found in the literature, such as "words that are assigned to concepts used in the special languages that occur in subject-fields or domain-related texts" (Wright, 1997, p. 13), these definitions, however accurate, aren't always helpful when deciding whether or not a lexical unit can be considered a term. Not only does this make ATE a challenge, it also poses a problem for the evaluation. The typical way to evaluate such a task is to compare the automatic output against a manually constructed gold standard (GS), in other words, to compare automatic versus human performance. In this way, precision (how many of the automatically extracted terms are correct) and recall (how many of the terms in the text are automatically extracted) can be calculated. To construct this GS, text must be manually annotated. The corpus must be large and domain-specific enough to be used as input for ATE and it must be annotated entirely to calculate recall. Owing to the ambiguous nature of terms and the necessary volume of text, this is an arduous, time consuming task. Moreover, it results in low inter-annotator agreement scores (Rigouts Terryn et al., Submitted), which are supposed to be an indication of the objectivity and quality of the annotations.

Recently, research on ATE has shifted from monolingual ATE, to bilingual ATE, first from parallel corpora and currently also from comparable corpora. ATE from comparable corpora (ATECC) attempts not only to recognise terms in a text, but also to find equivalent terms in the different languages of a comparable corpus. Comparable corpora are collections of texts in different languages, on the same subject (and preferably in the same style), but the texts are not each other's translations. Using comparable corpora is much more difficult than using parallel corpora, since it is impossible to know beforehand where to look for term translation equivalents or even whether appropriate equivalents are available in the corpus. However, comparable corpora have the great advantage of availability: it is much easier and less costly to collect comparable corpora (manually or automatically) than parallel corpora, which require aligned human translations. ATECC can therefore be used for languages with fewer resources or rare and specialised domains for which data is too scarce to compile a parallel corpus.

The result of ATECC is usually an ordered list of potentially equivalent candidate terms in the target language, for each candidate term in the source language. Consequently, the evaluation needs to include an evaluation of both term extraction, and term linking. Current research in ATECC is mostly evaluated by using reference translations from a source other than the input corpus or by using only a limited set of manually evaluated term equivalents from the input corpus. However, to accurately evaluate the entire output and be able to trace mistakes back to their source, a new type of GS is needed. To the best of our knowledge, we are the first to undertake the challenge of constructing a completely manually annotated GS for ATECC, thus requiring the development of a novel methodology to annotate and structure the data. To address the problem of low inter-annotator agreement due to the subjective nature of terms, a new term annotation scheme was developed and tested in combination with detailed annotation guidelines. The investment of time and effort is not to be underestimated, but the result is both an informative instrument for evaluation and an invaluable source of information about the nature of terms.

The remainder of this contribution is dedicated, first, to a summary of the state-of-the-art, subsequently, to a description of the GS and, next, to a discussion of what can be learnt about terms and term equivalents from this GS. The results will be recapitulated in the conclusions.

2. State of the Art

Researchers have been creative in finding ways to evaluate ATE, especially since the traditional way, i.e. calculating precision and recall, requires a fully annotated corpus. For instance, in the EVALDA-CESART project (Mustafa El Hadi et al., 2004), existing reference word lists were completed by domain specialists to calculate precision, but

recall remained a problem, since it requires a complete GS. Another problem emerged because of the need for term annotation instructions, resulting in a plethora of diverging annotation schemes and guidelines, rendering re-use and comparison problematic. For instance, Bernier-Colborne (2012) developed detailed, but very domain-specific annotation guidelines. Another example is the limitation to certain parts-of-speech (POS) patterns: sometimes, only nouns and noun phrases are annotated (Bernier-Colborne & Drouin, 2014); others allow more POS patterns (Schumann & Fischer, 2016).

Similar problems exist for the evaluation of ATECC. Constructing a complete record of all translation equivalents for all terms in a comparable corpus is a daunting task, so alternative solutions have been invented. For instance, Laroche and Langlais (2010) use translations from an existing thesaurus; Kontonatsios (2015) also uses a limited set of reference translations and sets the maximum system performance to an estimate of the percentage of translations that are present in the corpus. In the TTC project (Loginova et al., 2012), a GS was created based on the input corpus, but it was limited to ±100 term pairs.

As for the investigation of the structure of terms, surprisingly little empirical research can be found on term length or structure based on manual annotations. Justeson and Kats (1995) started from dictionaries of technical terminology and selected 200 technical terms from four different dictionaries. Only 35 of the resulting 800 terms weren't noun phrases, which led them to focus on noun phrases alone. Out of 800 noun phrase terms, 30% were single-word terms (SWTs), 55% were 2WTs, 12% 3WTs and the remaining terms were multi-word terms (MWTs) of four or more words. Only in the medical domain did they find more SWTs than 2WTs, which they attribute to the presence of more Latin or Greek single-word compounds. Around the same time, Nkwenti-Azeh (1994) had reached similar results.

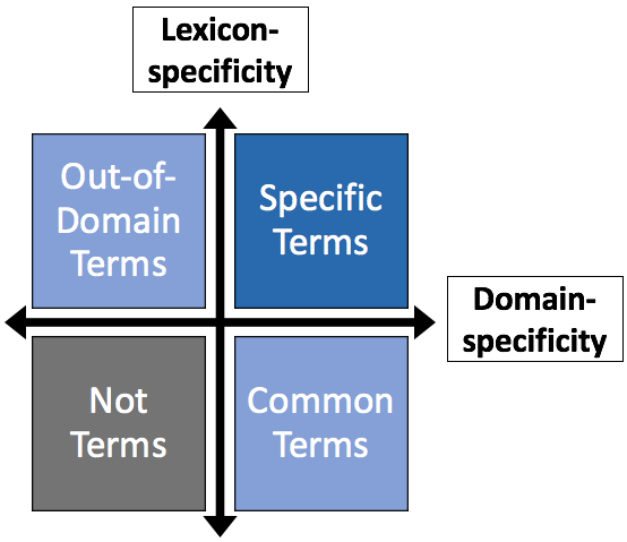

Figure 1: Annotation scheme

3. Gold Standard

3.1 Monolingual Gold Standard for ATE

Three languages were included in this project: English (EN), French (FR) and Dutch (NL). They provide a good contrast between well- and less-resourced languages (EN/FR vs. NL), Romance and Germanic languages (FR vs. EN/NL) and, particularly, languages with very different compounding strategies. Compound terms in English are often concatenations of nouns, separated by a whitespace, whereas, in French, the different parts of the compound are typically connected by prepositions and Dutch compound terms are characteristically one, long compound word. Besides language, the structure of terms may also be influenced by domain, so three different domains were selected: medical (heart failure), technical (wind energy) and juridical (corruption). Each corpus was manually checked and enhanced, but they were all based on pre-existing resources. The medical corpus of medical abstracts and short papers was based on previous research about terminology (Hoste et al., Accepted), as was the technical corpus (Daille, 2012). The juridical corpus was assembled based on a collection of titles provided by the DGT of the European Commission. The corpora are the same size for each language. Per language, the medical corpus has ± 46k words, the technical one ± 310k and the juridical corpus contains ± 670k words per language.

The medical corpus has been completely annotated and large parts of the other corpora have been annotated as well (see Table 1). The annotation of the juridical and technical corpora is an ongoing work, but a sufficient portion has already been annotated to provide a useful resource for the evaluation of ATE. There were two main concerns for the development of the annotation scheme. First, it should be intuitive and uncomplicated for the annotators and improve inter-annotator agreement. Second, the need for a "highly parametrizable" (Vivaldi & Rodríguez, 2007) GS had been expressed before, to encourage a more detailed analysis of terms and ATE. These considerations led to the development of three term labels: *Specific Terms*, *Out-of-Domain Terms* and *Common Terms*. These are defined by splitting termhood into two parameters: lexicon-specificity and domain-specificity. By representing these on two sliding scales, it results in the matrix shown in Figure 1.

	EN	FR	NL
Heart Failure (HF)	45.788	46.751	47.888
Corruption (CR)	50.322	49.180	50.676
Wind Energy (WE)	76.488	83.259	84.207

Table 1: Number of annotated tokens per corpus

Lexicon-specificity is defined as the degree to which a term belongs to either the general language or to the lexicon of specialists. Domain-specificity shows how relevant the term is to the subject. According to the strictest definitions of terms, they should score high on both scales. Terms in this category were labelled *Specific Terms*. In the domain of heart failure, "ejection fraction" would be an example of a *Specific Term*. However, no matter how well constructed the corpus, there may also be terms which are lexicon-specific, but not domain-specific. These are called *Out-of-Domain Terms*, or, abbreviated, *OOD Terms*. For instance, the medical corpus contains some terms about statistics, such as "p value", which is, in this case, an *OOD Term*. Finally, the opposite may be true as well: *Common Terms* are relevant to the domain, but are also part of the general vocabulary. In the heart failure corpus, a good example would be "heart", which is clearly domain-specific, but, since non-specialists are familiar with the term as well, not

lexicon-specific. While this doesn't eliminate the factor of subjectivity, previous experiments have shown that the labels help the annotators and increase inter-annotator agreement (Rigouts Terryn et al., Submitted). This is especially true for *Specific Terms*, which are often the most relevant terms to ATE users.

The annotation scheme is helpful for deciding whether a linguistic unit is a term. However, there is a second difficulty in term annotation, namely deciding the term boundaries. Elaborate annotation guidelines were constructed to address such added annotation difficulties (https://biblio.ugent.be/download/8503113/8517085.pdf). These guidelines also provide more information on how to determine the degree of lexicon- or domain-specificity as objectively as possible.

ID	**112**
Annotation	beta-blockers
Label	Specific Term
Frequency	4
Texts	144; 096
EN	EN
FR	2801; 3664; 4738; 5268; […]
NL	7558; 5774; 6015; 5998; […]
Lemma	Beta-blocker
Synonym	1971; 1450
Abbreviation	2567
Alt. Spelling	2099; 1509; 2243
Hypernym	87; 393; 1430; 1893; 1303; 111 […]
Hyponym	235; 1577; 2441; 2324; 2669; 222
Other	1027; 1035; 2462; 1563; 776; […]

Table 2: Example term record in the GS for ATECC

3.2 Multilingual Gold Standard for ATECC

For the evaluation of ATECC, the GS should provide more information than only termhood. Ideally, it would also include a record of all possible translation equivalents in the corpus and even additional semantic relations, to encourage a more nuanced evaluation. With such information, it would be possible to tell whether the suggested target language candidate term is a correct translation equivalent and, if not, if it is still in some way related to a correct equivalent. Most importantly, a wrong term suggestion could be traced back to its origins: either the system simply was not able to find the correct equivalent in the target language corpus, or the correct translation was not present in the corpus. It is important to remember that, since comparable corpora aren't in any way aligned, there is no guarantee that the translation equivalents for all terms are present in the corpus. Being able to trace the origins of mistakes in the term linking module of ATECC can be a useful tool to identify areas of improvement. To accommodate all this information in a single document, each unique annotation got an ID number, which could be used as a reference. The annotation, its label, frequency and the texts in which it was found were automatically extracted from the monolingual gold standards. Three fields for each language were added to indicate the source language and refer to the IDs of any equivalents in the other languages. To identify term variants, lemma, synonyms, abbreviations and alternative spellings were added manually. Other semantic links could be indicated as hypernyms, hyponyms or 'other'. Table 2 is an example of the term record for "beta-blockers". All numbers (except frequency and texts) refer to the IDs of other term records. In total, 6818 unique terms were thus created, with an additional 567 records for named entities.

4. Term Analysis

Besides being a useful tool for the evaluation of ATECC, this GS contains a wealth of information about comparable corpora, term frequency, term variation, differences per domain and language etc. In this contribution, we will focus on what can be learnt about the structure of terms, more specifically: term length and term POS patterns. In the following analyses, only term annotations are considered (no named entities) and the numbers we report are calculated on unique terms (one count per term record) instead of absolute frequencies (one count per term occurrence). Nevertheless, both calculations were made, to rule out any discrepancies. Barring some minor variations, they both lead to the same conclusions. It is also important to note, that, according to the annotation guidelines, all content words can be terms and no minimum or maximum term length was stipulated.

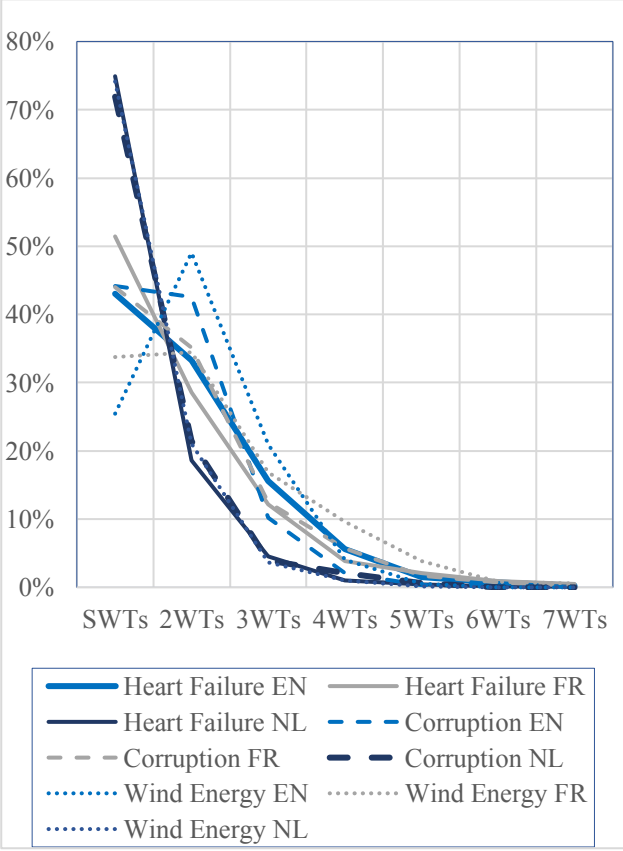

Figure 2: Percentage of terms per term length, comparing languages and domains

4.1 Word Length

When distinguishing solely between SWTs and MWTs, it became quickly apparent that both language and domain have an impact on term length. In this analysis, complex single-word compound terms were processed as SWTs. The percentage of SWTs ranged from 25% (English corpus on wind energy), to 75% (Dutch corpus on heart failure).

Figure 2 shows a chart of the term length for each language and domain. A first observation is that very few terms contain more than five words. In Dutch, over 90% of terms are no longer than 2 words. In English and French, there are more three-word terms, but no more than 10% of the terms are longer than that. Dutch term length (dark blue) appears very consistent, with hardly any variation per domain. French term length (grey) has more variation, especially for wind energy, where there is an equal percentage of SWTs and two-word terms (2WTs). In English, however, the variation per domain is more apparent. For the English corpus on wind energy, there are even more 2WTs (49%) than SWTs (25%). However, apart from the variations for SWTs and 2WTs, the numbers are rather consistent for all languages and domains. As the term length reaches 7 words, the number of terms of that length becomes negligible.

Another interesting observation regarding term length is the difference between the different term categories. There were too few *OOD Terms* to be relevant, but there was a notable difference in term length between *Specific* and *Common Terms*. As shown in Figure 3, *Common Terms* are more often SWTs, and *Specific Terms* are more likely to be longer. Only 1% of *Common Terms* are longer than 3 words, compared to almost 10% of *Specific Terms*. A potential explanation is, that *Specific Terms* are made up of several *Common Terms*, which, when combined, become more lexicon-specific.

Figure 3: Percentage of terms per term length, comparing term categories (averages over all languages and domains)

4.2 POS patterns - monolingual

For a more in-depth analysis of term structure, the LeTs preprocess toolkit (van de Kauter et al., 2013) was used for the automatic linguistic processing of all corpora. This included tokenisation and POS-tagging (i.e. automatically assigning a POS to each token). Only the results for term POS-patterns with large frequencies will be discussed here to avoid overgeneralisation. This is necessary, since automatic POS-tagging isn't flawless and some (parts of) terms were not assigned a POS, since they weren't tokenised separately. For instance, some terms were annotated that were connected to other words by "-", e.g. within the term "angiotensin-converting enzyme", "angiotensin" was annotated, as well as the full term. Since LeTs doesn't tokenise words connected by a hyphen separately, they weren't assigned any POS. Whenever the same term had received different POS-tags in the processed texts, the most frequent tag was used. When there was any ambiguity about the tag, the decision was made manually. For instance, the tagger had difficulties distinguishing between the nouns and named entities, especially for abbreviations, which sometimes lead to different tags for the same abbreviation in different sentences (e.g. "cTnT" occurred in the corpus six times, was tagged as a noun four times and as a named entity twice).

Figure 4 shows how, on average, over 80% of all terms (in all languages and domains) are one of eight POS patterns: single nouns (N), a noun and an adjective (N+A), a single adjective (A), a named entity (NE), two nouns (N+N), two nouns separated by a preposition (N+P+N), two adjectives and a noun (N+A+A) or a single verb (V). The order of nouns and adjectives varies depending on the language. Verbs are not often extracted by ATE, since the frequency of terminological verbs is considered so low, that attempting to extract them introduces more noise in the output than improved recall. Justeson and Katz (1995), for instance, found that only 3 out of 800 technical terms chosen from a dictionary were verbs. Therefore, it was surprising to find single verbs as a rather common POS pattern in our corpus. However, verbs rarely appear within MWTs. This may be explained by the fact that verbs aren't often combined with other words in the exact same way and that the lack of these set combinations leads annotators to only annotate the verb separately, but not as part of a larger MWT.

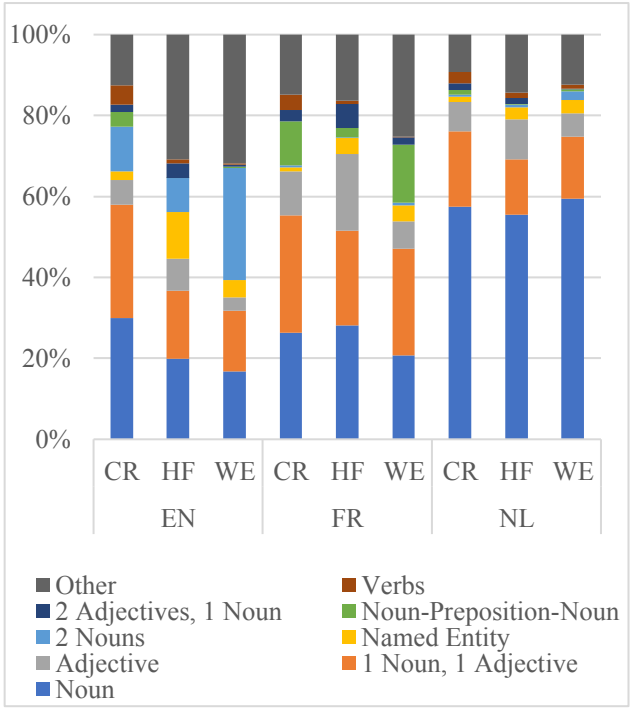

Figure 4: POS patterns

While there are some shared characteristics across all corpora (there are always many N and N+A patterns), substantial differences can be observed between different domains and, especially, between different languages. Some of these differences are easily explained. For instance, the different compounding strategies of the three languages are clearly visible. In English and French, the N+N pattern, which is almost non-existent in the other

languages, is often used for complex terms. In French, a similar phenomenon can be seen with the N+P+N pattern, which is rare in English and Dutch. In Dutch, complex terms are mostly formed by one, long, single-word compound, so there are much more N terms. Another feature may be explained by these compounding rules, namely, that Dutch seems to be less creative with different term POS patterns. A higher percentage of all Dutch terms falls into the above stated eight categories and, when comparing the number of unique POS patterns to the number of unique terms, Dutch has, on average, less different POS patterns (5%) than English and French (both 7%). The most likely explanation is that a single-word Dutch compound can be a combination of several nouns and adjectives, even though the assigned POS tag is N.

During the analysis of term length in the previous section, it was found that *Specific Terms* are less likely to be SWTs than *Common Terms*. The POS pattern analysis confirms this and provides a more detailed picture of the differences. The most common POS pattern for *Common Terms* is invariably N, whereas *Specific Terms* are more likely to be N+A (except in Dutch). Some other observations are more difficult to explain, such as the popularity of N+N terms in the English corpus on wind energy, or the fact that there are less A terms in English than in the other two languages. While the corpora are small enough that such variations may be due to chance, such peculiarities are worth keeping in mind when defining the parameters for ATE(CC).

4.3 POS patterns – translation equivalents

Analysing the term POS patterns per language already inspired some hypotheses about potential translation patterns (e.g. N+N in English ~ N+P+N in French ~ N in Dutch) and, thanks to the multilingual GS on heart failure, it is possible to substantiate these. Again, to avoid overgeneralisation, we only look at the most frequent patterns. For every POS pattern, the POS patterns of all available equivalents in the other languages were analysed. This revealed, for instance, that, out of the 365 Dutch translations for English N terms, 303 of them were also N terms. So, to find Dutch term equivalents for English N terms, the search should be focussed on Dutch N terms. However, in the opposite direction (English translations for Dutch N terms), the search should probably be widened to include other POS patterns, since only 260 out of 564 of the English term equivalents for Dutch N terms are also N terms. Other common POS patterns for the English equivalents are A+N (75), N+N (59) and NE (51). A similar pattern can be discerned between Dutch and French and for longer POS patterns, where English and French terms can often be mapped to shorter Dutch terms. In combination with the frequency of the N pattern, as discussed in the previous section, it is safe to say that single noun compound terms are a common occurrence in Dutch. A potential conclusion for the improvement of Dutch ATECC could be to incorporate automatic decompounding.

While this isn't very surprising, it is a good example of how the GS can help to define the parameters for ATECC. Another, more striking example, is the N+A pattern in English and French. Since the pattern is very common in both languages and single nouns are very rarely compounds in either language, it could be expected that the term equivalents of these patterns would correspond nicely. However, this is only true for 60% of the French equivalents found for English N+A terms and for only 43% of the English equivalents for French N+A terms. Some peculiarities may be due to differences in the automatic POS tagging for the different languages. For instance, there is no special tag for abbreviations in English, but there is one for Dutch and French. Equivalents for English terms with the NE tag do not often have the same tag in French (18%) or Dutch (12%) and there are more of NEs in English as well. A detailed look at the terms tagged as NEs in English revealed that at least part of this incongruity is due to the lack of a special tag for abbreviations in English, since it appears that these abbreviations are often tagged as NEs and, as was already discovered in a previous analysis, there are much more abbreviations in the English texts.

This is only a selection of some of the observations resulting from the GS, since the conclusions differ for each pattern and each language. However, the examples presented in this contribution do illustrate the usefulness of all the annotation work performed for the GSs and the potential of a bottom-up approach for ATE(CC).

5. Conclusion

Terms are a very ambiguous concept, which makes the development of algorithms for ATE a challenge, but also provides difficulties for the evaluation of the task. ATECC, which involves not only a term recognition module, but also a multilingual term linking module is even more difficult to evaluate. Therefore, several corpora in three different languages and domains were manually annotated. Based on the monolingual annotations about heart failure, a multilingual GS was created for ATECC. These resources will be made publically available in due course, just like the annotation guidelines, which already are. As illustrated in this contribution, such a GS can be used for more than evaluation purposes alone, as it is also a rich source of information about terms, which may inspire ideas for the improvement of ATECC.

We showed that, in general, terms are mostly SWTs or 2WTs and that very few terms are longer than 5 words. However, *Specific Terms* tend to be longer and less likely SWTs than *Common Terms*. Next, the POS patterns of the annotated terms were analysed. Apart from the popularity of N and N+A terms, there are notable differences across the three languages and even some differences per domain. Most of the differences per language were related to the language structure, such as compounding rules, but other differences may be the result of differences in the automatic POS tagging. Finally, it was shown how the data also provide information about the POS patterns of term equivalents in the different languages, which could be useful for the term linking module of ATECC.

Some of the ideas for the improvement of ATE(CC) based on these data are a maximum term length to improve ATE precision, Dutch decompounding for term linking and the inclusion of terms that aren't noun phrases, such as adjectives and even verbs and adverbs.

6. Bibliographical References

Bernier-Colborne, G. (2012). Defining a Gold Standard for the Evaluation of Term Extractors. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey: ELRA.

Bernier-Colborne, G., & Drouin, P. (2014). Creating a test corpus for term extractors through term annotation. *Terminology*, *20*(1), 50–73.

Daille, B. (2012). Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012*. Istambul, Turkey.

Hoste, V., Vanopstal, K., Rigouts Terryn, A., & Lefever, E. (Accepted). The trade-off between quantity and quality. Comparing a large web corpus and a small focused corpus for medical terminology extraction. *Across Languages and Cultures*.

Justeson, J., & Katz, S. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, *1*(1), 9–27.

Kontonatsios, G. (2015). *Automatic Compilation of Bilingual Terminologies from Comparable Corpora* (Doctor of Philosophy). University of Manchester, Manchester.

Laroche, A., & Langlais, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 617–625). Beijing, China.

Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., & Heid, U. (2012). Reference Lists for the Evaluation of Term Extraction Tools. In *Proceedings of the 10th International Congress on Terminology and Knowledge Engineering*. Madrid, Spain: ACL.

Mustafa El Hadi, W., Timimi, I., & Dabbadie, M. (2004). EVALDA-CESART Project: Terminological Resources Acquisition Tools Evaluation Campaign. In *proceedings of LREC 2004* (pp. 515–518). Lisbon, Portugal.

Nkwenti-Azeh, B. (1994). Positional and Combinational Characteristics of Terms: Consequences for Corpus-based Terminography. *Terminology*, *1*(1), 61–95.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (Submitted). In No Uncertain Terms: A Gold Standard for Automatic Terminology Extraction from Comparable Corpora. *Terminology*.

Schumann, A.-K., & Fischer, S. (2016). Compasses, Magnets, Water Microscopes. In *Proceedings of LREC 2016* (pp. 3578–3584). Portorož, Slovenia: ELRA.

van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, *3*, 103–120.

Vivaldi, J., & Rodríguez, H. (2007). Evaluation of terms and term extraction systems: A practical approach. *Terminology*, *13*(2), 225–248.

Wright, S. E. (1997). Term Selection: The Initial Phase of Terminology management. In *Handbook of Terminology management* (Vol. 1, pp. 13–23). Amsterdam: John Benjamins.