In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora

Ayla Rigouts Terryn, Véronique Hoste & Els Lefever

Language Resources and Evaluation

ISSN 1574-020X
Volume 54
Number 2

Lang Resources & Evaluation (2020)
54:385-418
DOI 10.1007/s10579-019-09453-9

Springer

Lang Resources & Evaluation (2020) 54:385–418
https://doi.org/10.1007/s10579-019-09453-9

ORIGINAL PAPER

In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora

Ayla Rigouts Terryn[1] · **Véronique Hoste**[1] · **Els Lefever**[1]

Published online: 26 March 2019

Abstract Automatic term extraction is a productive field of research within natural language processing, but it still faces significant obstacles regarding datasets and evaluation, which require manual term annotation. This is an arduous task, made even more difficult by the lack of a clear distinction between terms and general language, which results in low inter-annotator agreement. There is a large need for well-documented, manually validated datasets, especially in the rising field of multilingual term extraction from comparable corpora, which presents a unique new set of challenges. In this paper, a new approach is presented for both monolingual and multilingual term annotation in comparable corpora. The detailed guidelines with different term labels, the domain- and language-independent methodology and the large volumes annotated in three different languages and four different domains make this a rich resource. The resulting datasets are not just suited for evaluation purposes but can also serve as a general source of information about terms and even as training data for supervised methods. Moreover, the gold standard for multilingual term extraction from comparable corpora contains information about term variants and translation equivalents, which allows an in-depth, nuanced evaluation.

Keywords Automatic term extraction · Terminology · ATR · Comparable corpora · Term annotation

✉ Ayla Rigouts Terryn
ayla.rigoutsterryn@ugent.be

Véronique Hoste
veronique.hoste@ugent.be

Els Lefever
els.lefever@ugent.be

[1] LT3 Language and Translation Technology Team, Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

1 Introduction

Automatic term extraction (ATE), also often referred to as automatic term recognition (ATR), is the automated process of identifying terms in specialised texts, where terms can be described as the linguistic representations of domain-specific concepts. ATE is meant to alleviate the time- and effort-consuming task of manual terminology management by providing a ranked list of candidate terms identified in a given domain-specific corpus. Moreover, it has become an important pre-processing step in many natural language processing (NLP) tasks (Zhang et al. 2018), such as automatic indexing (Jacquemin and Bourigault 2003), automatic text summarisation (Zhang et al. 2004) and machine translation (Wolf et al. 2011).

Despite abundant interest from the research community, there is still plenty of room for improvement (Astrakhantsev 2017). Two often-cited obstacles are the difficulty to obtain datasets (Astrakhantsev et al. 2015) and the lack of a clear definition of what terms actually are (Pazienza et al. 2005). To evaluate ATE against human performance, a manually annotated gold standard is needed, which requires a lot of time and effort to create and often has low inter-annotator agreement due to the lack of a clear boundary between terminology and general language. Nevertheless, such datasets remain invaluable for accurate evaluation and are also needed as training data with the current evolution towards supervised learning and deep learning methodologies (Drouin et al. 2018a, b).

For multilingual ATE, the addition of the controversial concept of translation equivalence (Panou 2013) presents an added difficulty for evaluation and annotation (Le Serrec 2012). Since parallel corpora can be difficult to obtain, especially for the specialised domains that are interesting for ATE, research into multilingual ATE has recently shifted towards comparable corpora instead (Daille 2012; Delpech et al. 2012; Hazem and Morin 2016b; Kontonatsios 2015), i.e. texts in different languages, which are not translations but contain much of the same vocabulary because of a comparable topic (and style). ATE from comparable corpora (ATECC) attempts to identify the terminology in a comparable corpus and to suggest translation equivalents for these term candidates. In contrast to ATE from parallel corpora, there is no way to know where these equivalent terms might be found in the corpus, or even if they exist in the corpus at all. This complicates not just the task itself, but the evaluation as well.

The original aim of this research was to construct a gold standard for ATECC, based on manual annotations in an entire specialised corpus. However, since the first part of ATECC is monolingual ATE, the monolingual annotations can be used as a gold standard for ATE as well. Moreover, the inclusion of three different languages and four domains provides many opportunities for comparative evaluations and makes this a rich resource to discover term characteristics and term behaviour. Finally, the amount of manually annotated terms qualifies this dataset as a potential training corpus for supervised machine learning approaches.

The remainder of this paper is divided into five sections. First, the state-of-the-art will be presented, first for the evaluation of monolingual ATE, then considering term annotation and finally for the evaluation of multilingual ATECC. Section 3

provides a detailed description of all corpora and a summary of the annotators' profiles. Section 4 is dedicated to the monolingual annotation, containing a description of the annotation scheme and guidelines and the inter-annotator agreement experiments. This section concludes with a use case with the ATE system TExSIS (Macken et al. 2013) to illustrate the dataset's usefulness as a gold standard. Next, the multilingual gold standard for ATECC is presented, with an explanation of how it was constructed and a discussion of the results. The final section is devoted to the conclusion and ideas for future research.

2 State-of-the-art

2.1 Evaluation of monolingual ATE

Monolingual ATE provides the user with a list of term candidates identified in a given domain-specific corpus. Often, the first step towards identifying the terms is a linguistic preprocessing, during which words and sequences of words are matched to pre-defined part-of-speech patterns (e.g. Macken et al. 2013). The following step is to calculate (statistical) features which represent termhood and unithood, as defined by Kageura and Umino (1996). Finally, the resulting list of candidate terms is usually ranked and a definitive selection is made by determining cut-off thresholds for certain features, by selecting the top-n best candidates or, more recently, by applying machine learning to find the optimal combination of features (Fedorenko et al. 2013).

The traditional method of evaluation for ATE is to compare against human performance and calculate *precision* (how many of the candidate terms are actually terms), *recall* (how many of the terms in the text were correctly extracted) and *F-score* (the harmonic mean between precision and recall). These are closely related to the concepts of *noise* (incorrectly extracted terms) and *silence* (terms that should have been extracted but were not). Other evaluation methods, such as receiver operating characteristic (ROC) curves may be used as well, but they are less common in this domain (Azé et al. 2005). Since these metrics only measure performance, some researchers argue that a more holistic evaluation protocol is necessary. One of the very first evaluation protocols for ATE (L'Homme et al. 1996) broadly defines five pre-evaluation criteria to complement the performance metrics, ranging from an evaluation of the basic design to the way the results are presented in lists. In other early research, performance is measured by precision and recall, but with the disclaimer that "low scores do not mean inferiority" (Kageura et al. 1999), since the essential question of what terms are is still unsolved. Instead, they state that consistency of the results is as important as performance. In other work, Sauron (2002) proposes a quality model that measures not just precision or recall, but also suitability, reliability, usability, efficiency, maintainability and portability. Within the framework of the CESART project (Mustafa El Hadi et al. 2004; Mustafa El Hadi et al. 2006), a user- and application-oriented evaluation protocol is presented, but, as stated by Nazarenko and Zargayouna (2009), application-oriented evaluations are much more difficult to set up and it is difficult

to weigh the impact of the ATE on other tasks, such as indexing or thesaurus building. While a complete evaluation of any system should indeed be more elaborate, the remainder of this paper is dedicated to the performance aspect of ATE evaluation.

The two most popular ways to evaluate ATE performance are *reference term lists* and *manual validation* (Pazienza et al. 2005; Astrakhantsev et al. 2015). The former consists of using a list of terminology in the relevant domain to compare to the output, in order to calculate precision and recall. This reference term list (or gold standard) may be an adaptation of a pre-existing list (Enguehard 2003; Dobrov and Loukachevitch 2011; Wermter and Hahn 2005), a small sample of some of the terms in the corpus (Baroni and Bernardini 2004; Loginova et al. 2012), or a complete list of all terms in the corpus, identified through manual term annotation (Kim et al. 2003). Unless the entire corpus was manually annotated, only approximations of precision and recall can be calculated this way. The other strategy, manual validation, means manually validating the top-n candidate terms. This is usually done by either a domain-expert or a terminologist (Chen and Yan 2017; Gurrutxaga et al. 2013; Drouin 2003; Haque et al. 2018; Frantzi and Ananiadou 1999). The obvious drawback of this method is that it only evaluates precision, not recall.

Several strategies have been developed to counter the drawbacks of these evaluation protocols, while still avoiding the manual annotation of an entire corpus. The most obvious strategy is to combine the two strategies and use a pre-existing reference term list to calculate (approximate) recall, combined with manual validation for precision. Term Evaluator (Inkpen et al. 2016) uses a different strategy; this tool was specifically designed to facilitate the comparative evaluation of ATE systems. They validate candidate term lists and improve the consistency by, e.g. checking against previous annotations. Moreover, they calculate *relative recall* by comparing against the union of all correctly extracted terms by all systems. While this is a very practical tool with many useful features, relative recall is flawed in the sense that different ATE systems are likely to make some of the same mistakes. Thus, when different systems have trouble identifying the same terms, these will not be included in the calculation of relative recall. The following section discusses one of the biggest obstacles for ATE evaluation: manual term annotation.

2.2 Term annotation

The first reason for opting not to work with a fully annotated corpus is probably how time- and effort-consuming term annotation is. The volumes of the corpora currently used for ATE range from some 10 k tokens (Patry and Langlais 2005; Vivaldi and Rodríguez 2007), to several hundreds of thousands of tokens (Ghazzawi et al. 2018; Kim et al. 2003; Pazienza et al. 2005) to a million or more tokens (Inkpen et al. 2016; Zhang et al. 2008; Loginova et al. 2012). Nevertheless, with the rise of machine learning strategies for ATE (Conrado et al. 2013), the need for annotated corpora is becoming even more pressing, not only for evaluation, but also since "one of the major problems in applying machine learning to ATE is the availability of reliable training data" (Zhang et al. 2018).

There are a few annotated corpora available. One of the most popular resources is the GENIA corpus (Kim et al. 2003), which has been used in multiple ATE evaluations (Zhang et al. 2018; Zhang et al. 2008; Nenadić and Ananiadou 2006; Nenadic et al. 2004; Bordea et al. 2013; Fedorenko et al. 2013). GENIA is a collection of 2000 abstracts from the MEDLINE database in the domain of biomedicine, specifically about "transcription factors in human blood cells" (Kim et al. 2003, 180). All biologically relevant terms have been manually annotated by two domain experts, with additional linguistic annotations and labels for the GENIA ontology. In total, 93,293 terms were annotated in over 400 k tokens. Another biomedical dataset is the CRAFT corpus (Bada et al. 2010, 2012), for which the subject is very broadly defined as "biomedical journal articles". In this corpus, all terms referring to concepts that were represented in certain ontologies were strictly annotated. While this leads to a very consistent annotation, it also means that any terms not represented in any of the ontologies were not annotated. The ACL RD-TEC (Qasemizadeh and Handschuh 2014; Qasemizadeh and Schumann 2016) was designed specifically for ATE evaluation in the domain of NLP, based on the assumption that it would be a great advantage to have a dataset for which researchers in NLP could be domain experts themselves. The second version (ACL RD-TEC 2.0) contains 300 abstracts from the ACL Anthology Reference Corpus (Bird et al. 2008) with a total of 6818 term annotations. It has been used, among others, for a supervised learning approach to ATE (Hätty et al. 2017a).

There are also some smaller and/or lesser known resources, such as an automotive corpus of 224 k tokens in which all terms and term variants are annotated (Bernier-Colborne and Drouin 2014; Bernier-Colborne 2012). In an attempt to analyse the evolution of terms through time, Schumann and Fischer (2016) annotated a corpus of texts from different time periods, starting from 1665. The corpus was based on the Philosophical Transactions and Proceedings of the Royal Society of London and, using topic modelling, texts from the domain of mechanical engineering were selected for annotation. Over 10 k term occurrences were annotated in five corpora of 20–32 k tokens, spanning five time periods. A very different gold standard was created based on German online forum data about do-it-yourself (DIY) projects (Hätty et al. 2017b), which is promoted as a broad-topic corpus that contains many registers. At the time of the 2017 paper, they were aiming at a 80 k token corpus, fully annotated by 3 annotators. Hätty and Schulte im Walde (2018) included this DIY corpus and 3 others in later experiments to test inert-annotator agreement in term annotation by lay people. In the context of the TTC project (Daille 2012; Gornostay et al. 2012; Loginova et al. 2012), short reference term lists (between 107 and 159 terms per corpus) have been collected for specialized corpora in two domains (wind energy and mobile technology) and seven languages (Chinese, English, French, German, Latvian, Russian and Spanish). However, these reference term lists were created by annotating the output of ATE tools, not the source texts. Within the framework of the TermITH project (Billami et al. 2014; Projet TermITH 2014), a French corpus of scientific texts, specifically in the field of language sciences, is preprocessed using the TTC-Termsuite (Daille 2012) and the automatically generated candidate term list is manually validated, but the candidate terms are presented with their context of the original text. Enguehard

(2003) also presents a corpus annotated with the help of ATE tools: a 104 k token corpus in the domain of metallurgy. This corpus was already accompanied by a list of 6582 terms and the list was further enriched with the manually validated results of two ATE tools. Nazar (2016) had student linguists and domain experts annotate around 200 terms in an English corpus on psychiatry. Another noteworthy example is the research by Judea et al. (2014), who addressed the lack of resources by developing an unsupervised method of labelling training data, based on existing term identification and ranking methods, and high accuracy for automatic identification of terms in figure references in English patents. Additionally, more ad-hoc term annotation has been performed with the specific aim of evaluating an ATE tool, e.g. annotation of English mathematics terms (Amjadian et al. 2016) or Spanish medical terms (Vivaldi and Rodríguez 2007).

The annotation process for all these datasets varies considerably, so it is difficult to compare inter-annotator agreement scores. Moreover, not all researchers are able to calculate inter-annotator agreement scores, due to the expense of having multiple annotators go over the same text (Bernier-Colborne and Drouin 2014). To complicate matters further, the type of agreement score that is reported varies as well. However, the majority (especially the ones that do not start from pre-generated candidate terms) do report the difficulty and unavoidable subjectivity of the task, which often results in low inter-annotator agreement. No inter-annotator agreement scores could be found for the GENIA corpus. The inter-annotator agreement scores for the CRAFT corpus (Bada et al. 2010, 2012) are exceptionally high for most of the ontologies, especially after an initial period of adaptation. They express the scores in percentages and, for some of the corpora, can maintain an agreement of over 90%. This can probably be attributed to elaborate guidelines and, most of all, the strict link to the existing ontologies. For the ACL RD-TEC 2.0 (Qasemizadeh and Schumann 2016), F-score was calculated over four iterations, with discussions and elaborations of the guidelines between iterations. Average F-score started at 0.49 after the first iteration, climbing up to 0.74 after the fourth iteration. They also calculated self-agreement (same annotator, same text, two days in a row) and found that even self-agreement was no higher than 0.88, illustrating the difficulty of the task. Nazar (2016) also found "more than expected disagreement", reporting a Fleiss' Kappa index op 0.319 for the psychiatrist annotators (domain-experts) and 0.454 for the student annotators (linguists). For the DIY corpus (Hätty et al. 2017b) Fleiss' Kappa was reported as well, but with a very different approach, using IOB and additional labels, resulting in 9 labels per annotation. They reached substantial agreement with a Fleiss' kappa of 0.81. Schumann and Fischer (2016) report inter-annotator agreement with occasional discussion between annotators at an average F-score of 0.655, but with considerable variations per corpus (between 0.376 and 0.933). Agreement scores are generally high when a list of term candidates is validated, rather than terms in running text, e.g. kappa agreement between 0.53 and 0.84 (Amjadian et al. 2016). One more aspect that differs across all these evaluations, is how strict the *match* between two annotations or between an annotation and an extracted term candidate should be: some work solely with full span matches (Qasemizadeh and Schumann 2016), others with partial matches (Sauron 2002) or even more elaborate systems (Hätty et al. 2017b).

Clearly, there is a lot of variation in both the term annotation tasks themselves and how they are evaluated. One suggestion is that "a more fine-grained distinction between different types of terms […] might be helpful, at least partly, in alleviating the difficulty of the annotation task" (Schumann and Fischer 2016, 3582) and the reported agreement scores do appear to show that elaborate guidelines lead to higher agreement. However, there is a lot of disagreement about what those guidelines should be. A first point of disagreement is whether to extract only nouns and noun phrases (e.g. Bernier-Colborne and Drouin 2014), or also adjectives (Projet TermITH 2014), adverbs (Bada et al. 2012) and verbs (Schumann and Fischer 2016). Second, should there be a minimum and/or maximum term length? Some researchers do not limit term length at all (e.g. Bernier-Colborne and Drouin 2014), while others focus only on unigram (single-word) terms (SWTs) (Conrado et al. 2013; Estopà et al. 2000), multiword terms (L'Homme et al. 1996), or only on very specific combinations of part-of-speech patterns and term lengths (Vivaldi and Rodríguez 2007; Haque et al. 2018; Pazienza et al. 2005; Wong 2009). Another point of disagreement concerns the labels. In most research, only a binary evaluation (term vs. not term) is used. Bernier-Colborne and Drouin (2014) add additional information to the term annotation concerning term structure and variation (e.g. acronym, simple or complex term, etc.). Schumann and Fischer (2016) add confidence scores to the annotations, depending on the agreement between the annotators. The original ACL RD-TEC (Qasemizadeh and Handschuh 2014) distinguishes between technology and non-technology terms and the second version (Qasemizadeh and Schumann 2016) groups terms into 7 semantic categories. Term annotation in the DIY corpus (Hätty et al. 2017b) consists of three labels: *domain, domain-zusatz* or *ad-hoc*. The CRAFT (Bada et al. 2012) and GENIA (Kim et al. 2003) corpora both have very elaborate annotation schemes based on diverse ontologies.

While there are surely other term annotation projects, the sample discussed here already shows some interesting trends. First, it shows how diverse the methodologies for term annotation can be in every aspect: type of annotation (candidate term list or source text), annotation scheme (binary or multi-label) and annotation guidelines (e.g. term length and part-of-speech patterns). Even the evaluation of these resources (inter-annotator agreement) is very diverse. These differences make any comparison between corpora extremely difficult, especially when there is a lack of meta-information about the corpora. Second, it becomes clear that the resources are mainly in English and monolingual (with few exceptions). Third, only the TTC project includes multiple languages and domains, but their term lists are limited and not based on manual annotation of the corpus. Finally, high inter-annotator agreement scores are often attributed to detailed and elaborate guidelines. Our dataset was constructed with these observations in mind and takes into account remarks made during previous annotation projects, such as:

> In terms of annotation guidelines, a more fine-grained distinction between different types of terms (e.g. topic keywords, scientific standard vocabulary, foreign language words, unknown or "strange" words,…) might be helpful, at

> least partly, in alleviating the difficulty of the annotation task. (Schumann and Fischer 2016, 3582)
>
> Due to the great variability of TE scenario and the low agreement between terminologists and domain experts on what term candidates should be treated as terms, such gold standard should be highly parameterizable and should integrate (partial) evaluation pieces (and evaluators) (Vivaldi and Rodríguez 2007).

2.3 Annotation and evaluation of multilingual ATECC

ATECC generally consists of two steps. First, terminology is identified monolingually in the separate parts of a comparable corpus. Then, one of the monolingual lists of candidate terms is interpreted as the source language and, for each candidate term in the source language, suggestions will be made for potentially equivalent terms in the list of candidate terms in the target language. It is beyond the scope of this paper to discuss the various methodologies, but the evaluation challenges will be discussed in more detail. Since the first step of ATECC is monolingual ATE, it faces the same challenges as mentioned above. Additionally, an evaluation of the suggested translation equivalents is needed. This presents its own set of problems. First, since comparable corpora are not aligned, equivalents may be found anywhere in the corpus or not at all. Second, evaluating translation equivalence is no easy task, since *equivalence* is still a very controversial subject in translation studies (Panou 2013; Le Serrec 2012): how semantically related does a term have to be, to be considered a good equivalent? Finally, since this is still a relatively new area of research, there are very few available datasets and there is no consensus yet about the best way to evaluate ATECC.

The most common strategy for the evaluation of ATECC is to use reference term lists based on existing resources. For instance, Laroche and Langlais (2010) use 5000 English-French pairs of nominal terms from the Medical Subject Heading (MeSH) thesaurus. Similarly, Morin and Hazem (2014) selected single word terms that appeared more than 4 times from the UMLS meta-thesaurus and constructed English/French reference term lists (169 pairs for breast cancer and 244 pairs for diabetes corpus). A consideration when using reference term lists with tools that include a dictionary-lookup methodology, is that there should be no overlap. In such cases, reference terms might be selected that do not occur in the dictionary and equivalents for these terms could be searched in different resources (Saralegi et al. 2008). The EU term thesaurus EUROVOC has been used as well to train and test a classifier for bilingual terminology by Aker et al. (2013). They further calculated precision by manually evaluating 600 English-German candidate term pairs, using one of four categories: *equivalence* (for exact translations), *inclusion* (when the correct translation is part of the suggested translation), *overlap* (when an equivalent of at least one word of the source term can be found in the target term) or *unrelated* (when none of the above apply). This categorisation illustrates the need for a fine-grained evaluation. When using existing resources, similar problems apply as with monolingual ATE: without additional human control, systems may be unfairly

evaluated for correct translations which are not in the reference list or for wrong suggestions when the correct translations were not available in the corpus. These shortcomings are handled in creative ways, such as by Kontonatsios (2015), who, first, limited his set of reference translations to those found in his source language corpus and, subsequently, used a reference dictionary to estimate the percentage of terms in the source language corpus for which translations were available in the target language corpus. This percentage was used as the upper-bound for translation accuracy. Another methodology was used within the framework of the TTC project, where a GS was constructed based on the input corpus (Loginova et al. 2012). Using automatically extracted monolingual lists of candidate terms, SWTs and MWTs were selected as a starting point for the bilingual reference list. After validation, the monolingual reference term lists were used to create bilingual reference term lists of ca. 100 term pairs. Only terms which appeared both in the source and target corpus were included and the minimum term frequency was 10 for SWTs and 5 for MWTs. This GS is freely available and has been used in other research as well (Hazem and Morin 2016b). As with monolingual ATE, there are some researchers who stress the importance of an application-oriented evaluation, such as Delpech (2011) who wanted to test the use of ATECC for translation. In her evaluation protocol, translators were asked to make translations containing terminology with and without using ATECC output as a resource, after which other translators judged the quality of the translations. The judges evaluated each potentially problematic term translation as correct, acceptable or wrong. This *acceptable* judgement left room for interpretation when translations were not 100% correct, but could be acceptable in certain contexts or were very closely related to the correct translation. However, she stated that "some hitches in our procedure prevent us from clearly demonstrating the added-value of terminologies acquired from comparable corpora". In conclusion, while researchers have been resourceful in inventing evaluation protocols for ATECC, so far, to the best of our knowledge, no completely manually annotated and evaluated gold standard has been developed for ATECC and no methodology to do so has been suggested.

3 Corpora and annotators

3.1 Corpora

Based on our observations from the state-of-the-art and our ultimate goal of creating re-usable (evaluation) datasets for both monolingual ATE and multilingual ATECC, several corpora were carefully constructed. The first requirement was that multilingual comparable corpora were needed for ATECC. Additionally, a parallel corpus was constructed as well, with the same subject as a different comparable corpus, so that the performance of ATECC might be compared to that of ATE from parallel corpora. Since ATE systems are generally language-dependent and term characteristics and ATE performance may differ between languages, we included three languages: English and French as large, well-resourced languages with many opportunities for comparison and Dutch as a less-resourced language. Having one

Table 1 Overview of corpora with token count

	English	French	Dutch
Corruption (comparable)	489,191	475,244	470,242
Corruption (parallel)	176,314	196,328	184,541
Dressage	102,654	109,572	103,851
Heart failure	45,788	46,751	47,888
Wind energy	314,618	314,681	308,744

Romance and two Germanic languages also allows a comparison between languages of different families and, especially important for ATE(CC), between languages with very different compounding strategies. In English, compound terms are typically separated by a whitespace (e.g. "evaluation criteria). French versions of these terms often include a preposition (e.g. "critères d'évaluation"), whereas in Dutch, complex terms are concatenated into a single, long compound (e.g. "evaluatiecriteria"). Like language, domain also influences term characteristics, so four very different domains were included: corruption, dressage, heart failure and wind energy. Another potentially influential factor for ATE performance is the size of the corpora, so the corpora are of different sizes across the domains, yet similar sizes per language to improve comparability. An overview is presented in Table 1.

The corpora about corruption were chosen to represent the juridical domain. They are based on a collection of titles provided by the Dutch terminology department of the European Commission and the texts were manually collected. They contain mostly legal documents and texts about corruption policies, but also relevant newspaper and Wikipedia articles. A large portion of the texts are from the EU, the United Nations or Transparency International (a global organisation against corruption). Because of the availability of texts about corruption by the EU and the United Nations, it was possible to construct a parallel, trilingual, sentence-aligned corpus, in addition to the comparable corpus. The contents in the comparable and parallel parts of the corpora are very similar, but there is no overlap: no texts are used for both corpora.

The corpus about dressage was not based on any previous data and was included, first, to have a corpus for which the main annotator was a domain-expert (see also Sect. 3.2). Another motivation is a that such a completely different corpus related to sports and hobbies might offer interesting new insights, as demonstrated by Condamines (2017), who analysed fishing terminology. The corpus was constructed completely manually and contains mostly text from online magazines and blogs about horseback riding. Only texts that were written in standard and correct language were included. It is very focussed, since it contains only texts about one branch of the horseback riding sport: dressage.

The corpus about heart failure was based on previous research about the influence of corpus quality and size on ATE (Hoste et al. 2019). Based on the titles crawled in that previous research, abstracts about heart failure were manually collected. Since the previous research did not include French, similar abstracts were manually added for French. Due to the limited available number of medical abstracts about heart

failure, a few short papers were included as well, but the majority of the corpus exists of published medical abstracts that have a strong link to heart failure.

Finally, the corpus about wind energy was also based on previous research to improve the options for comparison with other state-of-the-art research. The English and French parts of the corpus were freely available on the TTC project (Loginova et al. 2012) website.[1] A comparable Dutch corpus was manually added, based on the descriptions and the content of the English and French parts. The texts in this corpus range from technical descriptions of wind turbines, to academic papers about the engineering behind turbines and reports about the impact of wind turbines on the environment.

The English and French parts of the TTC corpus on wind energy were left unchanged. All other corpora were semi-automatically cleaned (e.g. removal of content tables, bibliographies and footnote numbers) and since corpus collection was done manually, the corpora should contain very little out-of-domain data.

3.2 Annotators

This research was performed in the context of a small project and, considering the amount of data, there were insufficient resources to hire domain-experts to annotate the corpora. To increase consistency across corpora, the same annotator(s) had to be able to annotate all corpora. Since it would be near impossible to find an annotator who is a domain-expert in all four domains and proficient in all three languages, it was decided that the annotators only needed to be language experts. While parts of the corpora have been annotated by several annotators to calculate inter-annotator agreement (see Sect. 4.2) and while language-students assisted with the annotation, most of the annotation work was performed by a single annotator. Moreover, all annotations made by other annotators were reviewed by this main annotator, who was fluent in all three languages and an experienced terminologist. Since consistency could not be guaranteed by having multiple annotators go over all texts and only keeping the terms on which multiple annotators agreed, the next best (possible) option seemed to work with one main annotator: the annotations will unavoidably still be subjective, but at least they will be as consistent as possible.

While many researchers have claimed that it is necessary to have domain-experts validate the terminology, we argue that a thorough knowledge of the language and experience with terminology might be equally, if not more important. Of course, annotators will spend a lot of time researching some of the terminology (maybe even needing to consult with domain experts on occasion), but having some distance from the topic may allow annotators to recognise non-general vocabulary more easily. The main annotator experienced this while annotating the corpus on dressage, on which she was a domain-expert herself. The annotation process went faster because she had no trouble understanding the terminology, but it was often more difficult to distinguish between general vocabulary and terms, simply because these terms had become part of her personal general vocabulary. For now, these observations are only based on impressions, but it would be interesting to research

[1] http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html.

further. For instance, Nazar (2016) found higher inter-annotator agreement between student annotators, than between domain-experts (psychiatrists). In other research, inter-annotator agreement between laypeople who annotated corpora in several domains showed that "laypeople generally share a common understanding of termhood and term association with domains" (Hätty and Schulte im Walde 2018, 325). A final argument in favour of our methodology is that, after annotating several tens of thousands of words in domain-specific corpora in three languages (as was the case in this project), any layperson would gain a minimal understanding of the terminology in that domain. This means that the annotation was not always a linear process: the annotator regularly went back to change previous annotations, always attempting to find the most logical and consistent way of labelling.

4 Monolingual annotation

4.1 Annotation scheme

To make the dataset fit for different applications of ATE, yet also domain- and language-independent, an annotation scheme was developed with three term labels, based on two parameters. In the pragmatic school of terminology, "two broad classes of distinctions are made, the first using the criterion of known/unknown and the second distinguishing between subject-specific and non-subject-specific terms" (Pearson 1998, 1:21). However, Pearson rejects both classes as too vague and does not define different term categories, believing that users will be more interested in identifying terms than distinguishing between different types of terms. Nevertheless, such domain- and language-independent distinctions between terms may prove helpful for more application-oriented evaluations of terms, since it has long been argued that different users require different terms. For instance, Estopà (2001) had four different groups of professionals annotate terms in a medical corpus: doctors, archivists, translators and terminologists. She found great differences between their annotations, e.g. terminologists annotated most terms and translators annotated much fewer terms (only the ones that did not belong to their general vocabulary and might present translation difficulties). Warburton (2013) similarly remarks that translators are not interested in any terms that belong to the general lexicon. In this sense, having different term labels based on the two classes mentioned by Pearson (1998) might improve customisation options for different applications. The first parameter will be called *domain-specificity and* represents the degree to which a term is related to the researched domain. This has been mentioned before, e.g. "All specialized languages show a gradient of domain-specificity" (Loginova et al. 2012) and the TermITH project guidelines (Projet TermITH 2014) specify that terms from the transdisciplinary domain and from a different domain should be rejected. The second parameter will be called *lexicon-specificity*, i.e. the degree to which terms are either part of the specialised lexicon used only by domain experts or part of general language. Drouin (2003) mentioned this term in earlier work and, in a more recent paper (Drouin et al. 2018b), a scale is presented of four degrees of lexicon-specificity: from topic-specific, to subject-specific (in their case: environmental), to

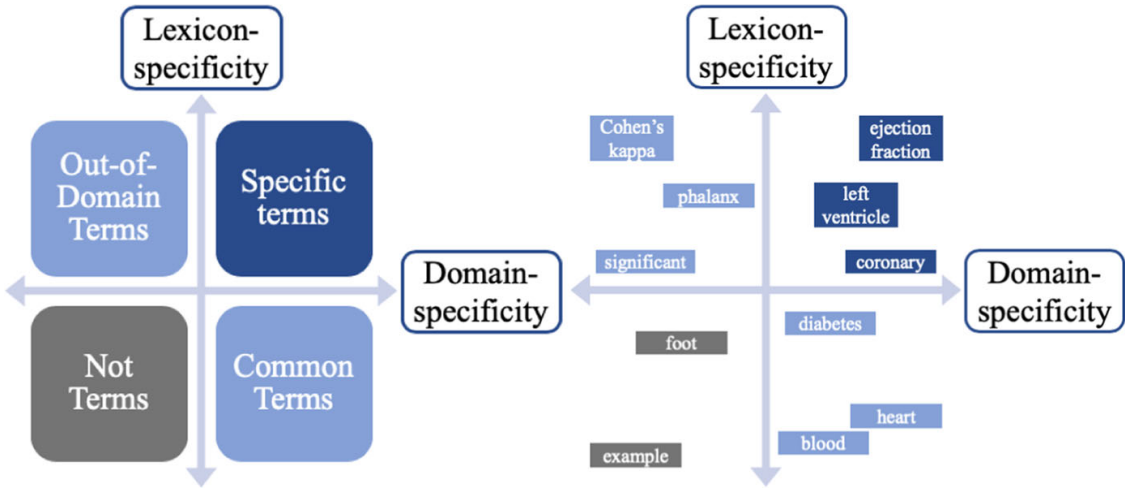

Fig. 1 Annotation scheme with parameters and term labels (left), with examples in the domain of heart failure (right)

transdisciplinary, to general lexicon. By combining lexicon- and domain-specificity in a matrix, as shown in Fig. 1, three term categories can be defined.

The three categories of terms are labelled: *Specific Terms, Out-of-Domain (OOD) Terms* and *Common Terms*. Examples in the domain of heart failure are shown in Fig. 1. Specific Terms are both lexicon- and domain-specific and are terms according to the strictest definitions of the concept. An example in the domain of heart failure would be *ejection fraction*, which is not part of general language and laypeople probably do not know its meaning. At the same time, it is strongly related to the heart failure, having something to do with the volume of blood pumped with each heartbeat. OOD Terms are lexicon-specific, but not domain-specific. For instance, in the corpus about heart failure, some of the medical abstracts contained terminology related to statistics, such as *p value*, which is not part of the general lexicon, but it is not very specific to the domain of heart failure either. This category contains, among others, what Hoffmann (1985) called "allgemeinwissenschaftlicher Wortschatz", which can be translated as "non subject-specific terms" (Pearson 1998). The final label, Common Terms, is meant for the opposite case, when terms are strongly related to the domain, but are not very lexicon-specific, such as *heart* in the domain of heart failure. This may be related to what Hazem and Morin mean when they describe "technical terms that have a common meaning in the general domain" (Hazem and Morin 2016a, 3406). While we do not deny that domain experts will have a much more intricate idea of the concepts behind Common Terms like *heart* and *blood*, generally, all laypeople do have at least a basic idea of the concept and know the words. These categories could be used to customise the data to the application, so that, for instance, translators could ignore any terms that are not lexicon-specific, since they would likely be part of the translator's known vocabulary.

An additional label was included for *Named Entities* (NEs), since they can be very closely related to terms, as shown by the fact that they are often mentioned in term annotation guidelines with specific instructions (e.g. Projet TermITH 2014;

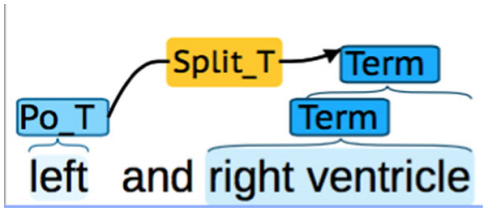

Fig. 2 Example of a split term annotation in BRAT

Schumann and Fischer 2016). Another problem that has been mentioned in many related research (Hätty et al. 2017b; Bernier-Colborne and Drouin 2014; Kim et al. 2003) are the *Split Terms*, i.e. terms that are somehow interrupted by other words or characters. Two common causes are abbreviations (e.g. *left ventricular (LV) hypertrophy*, where there are two split terms: *left ventricular hypertrophy* and *LV hypertrophy*), and coordinating conjunctions (e.g. *left and right ventricle*, where both *left ventricle* and *right ventricle* are terms). This was solved by creating *Part-of* term labels, which could be connected to each other, as shown in Fig. 2. All annotations were made in the BRAT online annotation tool (Stenetorp et al. 2011).

The annotation scheme is accompanied by elaborate guidelines which were constructed during the annotation process and after discussions between annotators. They contain instructions on as many recurring problems as possible. Like the annotation scheme itself, the guidelines are language- and domain-independent, though examples are cited from the corpora discussed here. Since the complete guidelines are freely available online,[2] only a sample of some of the most important instructions will be discussed here. The annotation scheme only provides a basis for whether or not to annotate a term and with which label, so many of the instructions in the guidelines concern term boundaries, viz. which span should be annotated. The first important rule is that each occurrence of all terms must be annotated, even if it is embedded within a longer term. This can be seen in Fig. 2, where both the multiword term or complex term *right ventricle* and the simple, single-word term *ventricle* are annotated. Moreover, there is no minimum or maximum term length and all content words may be annotated: nouns and noun phrases, but also adjectives, adverbs and verbs. Another notable issue concerns the distinction between different labels, since "decisions on the 'generalness' of a term candidate are somewhat subjective" (Warburton 2013, 99). An example regarding the difference between Common Terms and Specific Terms is to check whether the term is used in publications which are addressed to a large, non-domain-expert audience, such as tabloids. If the term is used, without any further explanations, in such a source, it is safe to assume it is part of the general lexicon and therefore more likely a Common Term than a Specific Term. An example here could be the term *heart failure*. There is no doubt about this term being domain-specific enough, since it was literally the subject of the corpus. However, the lexicon-specificity is more difficult. Is *heart failure* part of general vocabulary or not? Intuitively, one could assume that many people have at least heard of the term before and have some basic understanding of what it means, but is that only because the term is so descriptive? To decide, we looked at occurrences of the term in a Google News search. Since it

[2] http://hdl.handle.net/1854/LU-8503113.

appeared regularly and without further explanations in newspapers and magazines which aim at a very large, general audience, it was decided that *heart failure* would get the label Common Term. While this method is certainly not perfect, it provides a somewhat objective strategy in case of doubt. More examples and strategies can be found in the guidelines online.

A final consideration was that annotators were instructed to annotate the terms as they appeared in the text., irrespective of whether the terms were accepted in the field or if they were spelled according to the latest conventions. As long as they were used as terms in that text, they should be annotated as such. Since the annotators were no domain-experts, this was the most manageable approach. It is also the most logical one if the purpose is to compare human performance against ATE performance, since the annotators were only identifying the terms in the data that was there, without reference to some external ontology to which an ATE system might not have access. Indeed, one of the primary uses of ATE is identifying terms that are not in any databases yet, so it is important to identify terms as they appear in the texts.

4.2 Inter-annotator agreement

4.2.1 Pilot study

In a preliminary pilot study, inter-annotator agreement was calculated between three annotators who each annotated around 3 k tokens per language in the corpora about corruption, heart failure and wind energy (total ± 40 k tokens). All possible aids could be used, especially since the annotators were no domain-experts. They were, however, all fluent in the three languages. Similar to the procedure followed during the annotation of the ACL RD-TEC 2.0 (Qasemizadeh and Schumann 2016), there were two annotation rounds, with discussions of the results between each round. First, F-score was calculated to test agreement on term span annotations, without taking into account the given label, where:

$$\text{Precision of Annotator A versus B} = \frac{\text{Annotator A} \cap \text{Annotator B}}{\text{Annotator A}}$$

$$\text{Recall of Annotator A versus B} = \frac{\text{Annotator A} \cap \text{Annotator B}}{\text{Annotator B}}$$

$$\text{F-score of Annotator A versus B} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{recall}}$$

Agreement was calculated on type, not token. Consequently, when an annotator gave a label to a certain term, but forgot to accord the same label for a later occurrence of the same term, agreement did not decrease. Average F-score after the first iteration was 0.641, which was already good considering the task, but not great. 4207 unique annotations were found in this first round and only 33% was annotated by all annotators, 26% was annotated by two and 41% was annotated by a single annotator. These results are similar to those reported by Vivaldi and Rodríguez (2007). Discussing annotations in detail, improving the guidelines and then

returning (separately) for the second iteration resulted in a drastic improvement to an average F-score of 0.895. To determine agreement on the labels, Cohen's Kappa was calculated on all shared term annotations. These results were already very promising after the first iteration, with an agreement of 0.749 and improved after the second iteration to 0.927. While this was a good indication of the validity of the procedure and a great way to optimise the guidelines, the methodology was imperfect since specific cases were discussed in detail between rounds and the same dataset was re-used. Consequently, more rigorous experiments were organised.

4.2.2 Inter-annotator agreement evaluation

The purpose of this experiment was to see if the proposed annotation scheme and guidelines improved inter-annotator agreement. For this purpose, annotators were asked to annotate terms in a part of the heart failure corpus in three tasks with different instructions:

Test group:

- Task 1: single label (*Term*) annotation with only term definitions from literature (e.g. Cabré 1999; Faber and Rodríguez 2012) as guidelines.
- Task 2: term annotation with the four labels as specified above and with an explanation of the annotation scheme, but no further guidelines.
- Task 3: term annotation with the four labels like in task 2, but with the full guidelines.

Control group:

- For all texts: annotate all terms (without any additional information about terms).

Two different abstracts were chosen for each task, all with a similar word count (so six different texts in total). Texts were chosen without any Split Terms, to avoid the added difficulty. Moreover, readability statistics (De Clercq and Hoste 2016) were calculated to ensure that the texts were all of a comparable difficulty. The annotators all came from different backgrounds and the only requirement was that they knew English well enough to read a scientific text. While we expect to obtain much lower agreement scores than would be desirable due to the diverse annotator profiles, the main goal in this experiment was to compare agreement with and without our annotation scheme and guidelines. Therefore, in a control group, annotators were asked to annotate the same texts, but all with the same instructions.

There were 8 annotators in the test group and 6 annotators in the control group. All annotators were between 20 and 30 years of age and knew sufficient English to understand the texts. Other than that, there were few similarities between annotators. Seven of them were students with a language-related degree, but the others all came from very different backgrounds, including a medical student, a music teacher and an engineering student. While there are many other possible

Table 2 Average inter-annotator agreement scores per group and per task

	Task 1	Task 2	Task 3
Test group			
Average F-scores	0.48	0.56	0.59
Average Cohen's Kappa	− 0.36	0.06	0.11
Control group			
Average F-scores	0.44	0.46	0.40
Average Cohen's Kappa	− 0.29	− 0.26	− 0.26

patterns in these data, the analysis in this contribution will focus only on the validation of the annotation scheme and guidelines.

Agreement was calculated between all annotator-pairs as described in Sect. 4.2.1: first, F-score was calculated, then Cohen's Kappa. Since chance-corrected agreement scores, like kappa, can only be calculated when the total number of annotations is known (which is impossible for term annotation in the full text), this is usually only calculated on the intersection of annotations made by both annotators (Qasemizadeh and Schumann 2016). However, this would mean having to exclude the first task from the comparison, since only one label was used in this task. Similar to the methodology proposed by Vivaldi and Rodríguez (2007), we instead take the union of all terms annotated by both annotators as an approximation. Still, comparisons between the first task and the other two will have to be carefully interpreted, since kappa score was calculated on a different number of categories (two categories in task 1: term or not-term; vs. five categories in task 2 and 3: Specific Term, Common Term, OOD Term, Named-Entity or not-term).

In Table 2, it can be observed that agreement scores, especially kappa scores, are low, as expected. However, a first indication in favour of the annotation scheme and guidelines is that agreement increases per task in the test group. While the difference is small, the results are further validated by the fact that agreement in the control group stays roughly the same for all tasks and even decreases. The difference in agreement between the second and third task is very small, which may be due to the fact that the guidelines are too elaborate to be helpful for inexperienced annotators for such a small annotation task. It can even be seen as a sign in favour of the annotation scheme, i.e. that it works well on its own, even without elaborate guidelines. Since the improvement in agreement can be seen for both F-scores and kappa-scores, we carefully conclude that (1) the annotations scheme improves consistent term annotation when compared to annotation based on no more than term definitions, (2) the guidelines may be a further help to annotators, and (3) including multiple labels does not decrease agreement. Finally, while agreement is expectedly low among annotators with such diverse profiles, we are optimistic that experienced annotators/terminologists can be more consistent, as indicated by the pilot study.

A final remark concerning inter-annotator agreement is that, as mentioned before, the final annotations were all made or at least checked by one experienced annotator and terminologist, to improve consistency. Additionally, other semi-automatic

Table 3 Number of tokens annotated per domain and language

	English	French	Dutch
Corruption (parallel)	45,234	50,429	47,305
Dressage	51,470	53,316	50,021
Heart failure	45,788	46,751	47,888
Wind energy	51,911	56,363	49,582

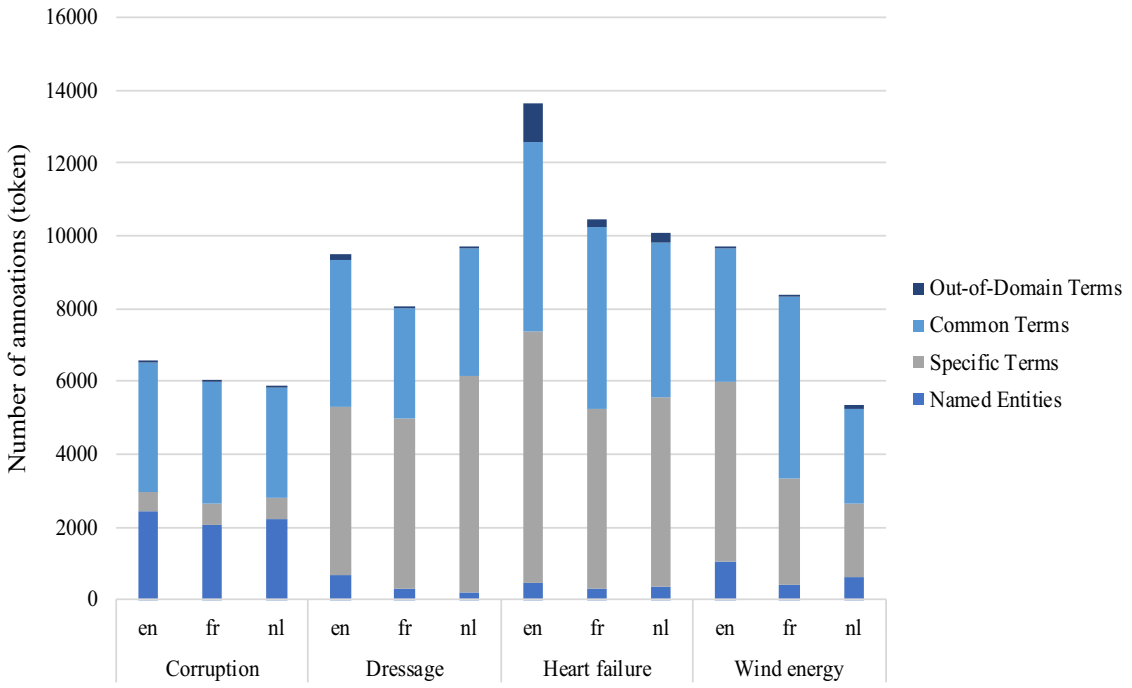

Fig. 3 Number of annotations (token) per language, domain and category

checks were performed to ensure the annotations would be as consistent as possible. For instance, when the same word(s) received a different label at different instances, the annotator double-checked whether it was an inconsistent annotation, or a polysemous term.

4.3 Results and analysis

Around 50 k tokens have been manually annotated per domain and language, leading to a total of 596,058 annotated tokens across three languages and four domains, as represented in Table 3. Only the parallel corpus on corruption was annotated; not the comparable part. This resulted in 103,140 annotations in total and 17,758 unique annotations (= 17.2 unique annotations per 100 tokens). For comparison: in the ACL RD-TEC 2.0 (Qasemizadeh and Schumann 2016) 33,216 tokens were annotated, resulting in 4849 unique terms (= 14.6 unique annotations per 100 tokens). Since only nominal terms were annotated for the ACL RD-TEC 2.0, this difference was to be expected.

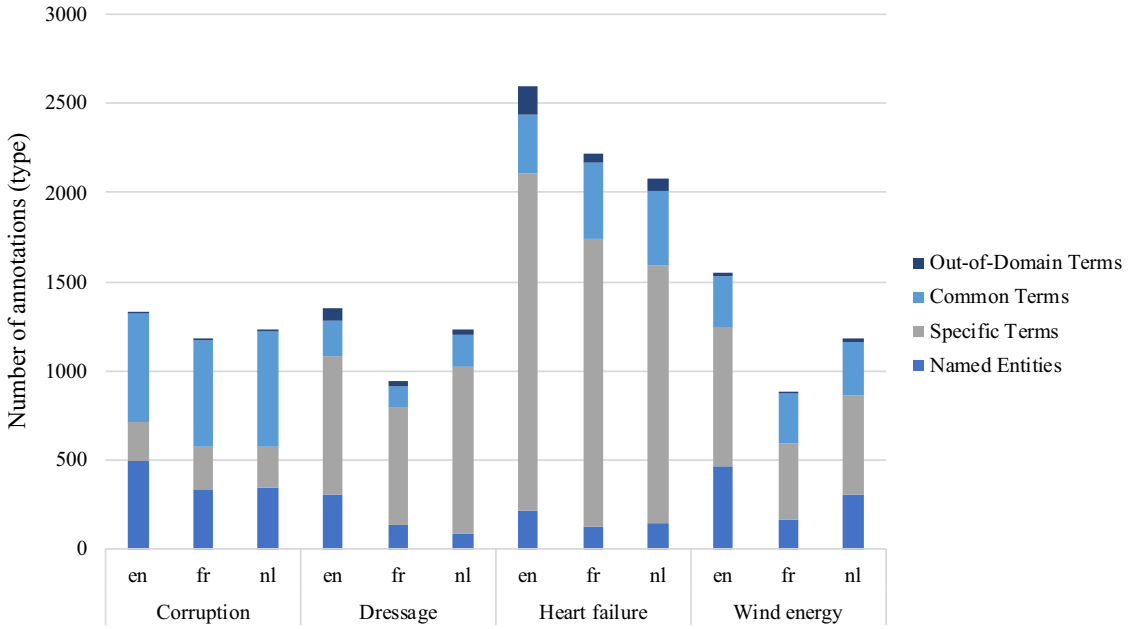

Fig. 4 Number of annotations (type) per language, domain and category

The first observation concerns the number of annotations per language, domain and category. This is shown in Fig. 3 for tokens and Fig. 4 for types. As can be seen in these graphs, the largest differences are between corpora in different domains. Within each domain, the total number of annotations in all languages is reasonably similar, as is the distribution over the different term categories. This is encouraging, since the corpora should be as comparable as possible. Of course, since the corpus on corruption is a parallel corpus, the differences there are smallest. The fact that more tokens were annotated in the English corpus on heart failure than in the other languages, despite it being the smallest in number of tokens, may be related to the fact that English is so predominant in this type of literature. Maybe terms are coined more easily in English or maybe it is related to the fact that, due to less available data in French and Dutch, more abstracts in the alpha sciences, e.g. regarding patient care and quality-of-life were included. Such abstracts may, in this context, contain less terminology than the medical abstracts in the beta sciences, though this is no more than a hypothesis. The corpus on dressage is one of the most focussed corpora in terms of subject, which would explain why, while there are quite a lot of annotations when looking at tokens, there are fewer when types are considered: there are a lot of terms in the corpus and many recurring terms. The same seems to be true for the French corpus on wind energy. Otherwise, both views lead to roughly the same conclusions.

Concerning the distributions over the different term categories, there is one corpus that stands out, namely the one about corruption. In this corpus, there are many NEs and very few Specific Terms when compared to the other corpora. This can be logically explained by the fact that (1) the legal texts often contain many person and place names, in addition to titles of laws etc., and (2) juridical terms are more likely to find their way to general language. Juridical proceedings are often reported in the news and many people get confronted with some legal jargon when,

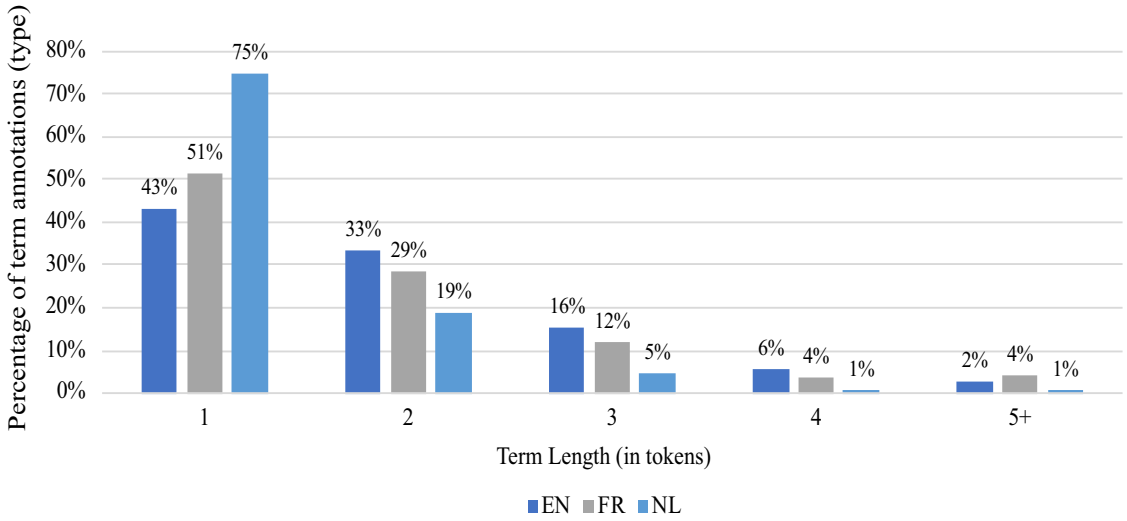

Fig. 5 Term length (in number of tokens) of term annotations (excluding NEs), counted per type

e.g. buying or renting a house, paying taxes, signing any type of contract, etc. The percentage of OOD Terms in the medical corpora can be explained by the prevalence of statistical terms. Since statistics are often required in scientific research, such terms may appear in the abstracts, even though they are not directly related to heart failure. There are relatively more Common Terms when looking at tokens than at types, since there are often few general language terms that are related enough to the domain to be included, but these do occur quite often, e.g. *heart* and *blood* in the domain of heart failure. The opposite is true for NEs, which do not occur very often and are not repeated often, so type counts are relatively higher.

The next analysis concerns term length, as shown in the graph in Fig. 5. While the differences are slightly less extreme when counting per token, the general conclusions remain the same. While there are some differences between the different domains as well, most differences are between languages. A first conclusion is that terms are generally quite short, with few exceeding a length of five tokens. The longest term was ten tokens long in the French corpus on heart failure: *inhibiteurs de l'enzyme de conversion de l'angiotensine II*. There are more single-word terms than two-word terms in all languages, even though the difference is very small for English. There are exceptions, e.g. the English corpus on wind energy has more two-word terms than single-word terms. Still, these findings are surprising when compared to some other research. In the ACL RD-TEC (Qasemizadeh and Handschuh 2014), there are many more two- and even three-word terms than single-word terms. In earlier work (Justeson and Katz 1995), two-word terms are also found to be much more common than single-word terms, except in the medical domain. However, there are also some findings that are more similar to ours. Estopà's (2001) finds that 42.91% of the terms in her medical corpus are simple noun terms. A German annotation experiment (Hätty and Schulte im Walde 2018) arrived at similar conclusions, with 46.7% single-word terms. One potential explanation for the differences is the inclusion of terms other than nouns and noun

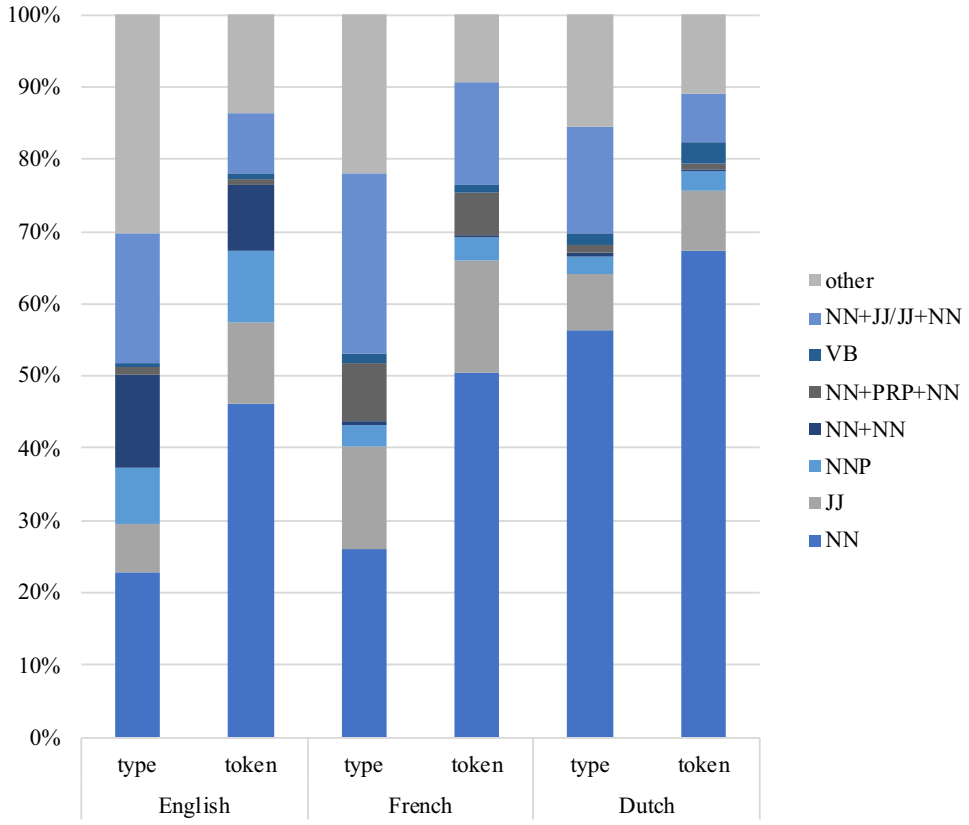

Fig. 6 Part-of-speech patterns per language, excluding NEs

phrases. The corpus itself may also have a considerable influence. A final observation is that there are many more single-word terms in Dutch. This is more easily explained by the pervasiveness of single-word compounds in Dutch.

The final part of this discussion will focus on the part-of-speech patterns that were found in the corpora, as shown in Fig. 6. All corpora were automatically tagged using LeTs Preprocess (van de Kauter et al. 2013). These results have been discussed in more detail in (Rigouts Terryn et al. 2018). The main conclusions were, that (1) nouns and noun phrases are important, but adjectives and even verbs are not uncommon; (2) there are a few common patterns, as can be seen in the graph, but 10–30% of the annotations have other, often quite complicated part-of-speech patterns; (3) the patterns vary considerably per language and domain.

4.4 Use case with TExSIS

In the previous section, a sample was presented of the type of information that could be gained from the dataset. In this chapter, the practical use of the dataset as a gold standard will be illustrated by means of a use case with the monolingual pipeline of the hybrid ATE system TExSIS (Macken et al. 2013). For this experiment, the threshold cut-off values of TExSIS were set very low, so there was a clear focus on recall over precision. Moreover, TExSIS currently only extracts nouns and noun phrases, which will impact recall, since the gold standard does include other part-of-speech patterns. NEs were included in all analyses, since TExSIS includes a named

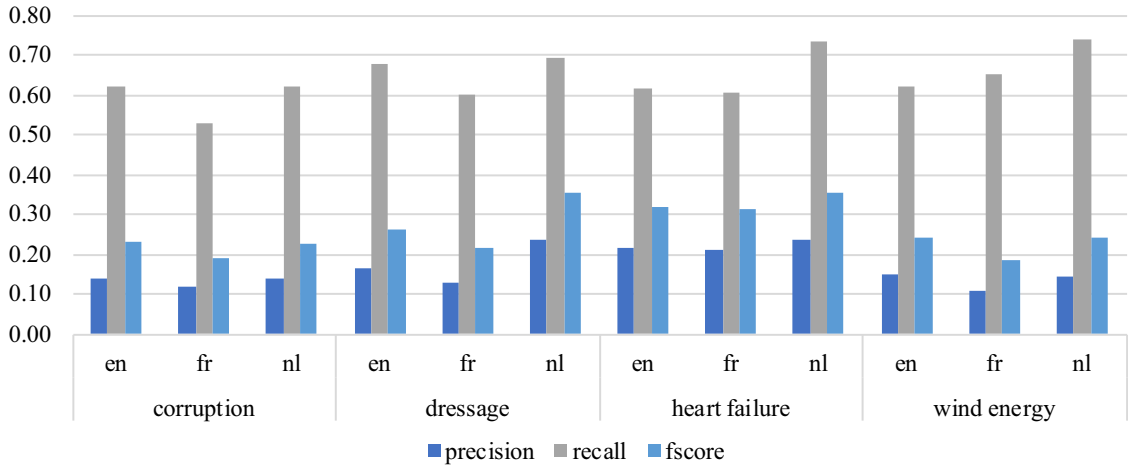

Fig. 7 Precision, recall and F-score per corpus

entity recognition (NER) module. Split Terms, however, were excluded, since TExSIS cannot handle interrupted occurrences of terms. It should also be noted that is not the aim of the use case to provide an elaborate evaluation of TExSIS, but rather to illustrate the usefulness of the dataset for evaluation purposes.

First, precision, recall and F-scores were calculated for all corpora, for which the results are presented in Fig. 7. As expected, recall is much higher than precision. There are also considerable differences between the different domains and languages. For instance, the three corpora with the worst F-scores are all French, though the French corpus on heart failure scores fourth best. This may be because the system seemed to work best on this domain: the three corpora on heart failure have the best (Dutch), third best (English) and fourth best (French) F-scores. In all domains, the French corpora score worse than their counterparts in English and Dutch.

Next, the impact of the different termhood measures was analysed. This included Vintar's termhood measure (2010) and log-likelihood ratio (llr). For the correctly extracted terms, the overall average termhood score was 12.13 and the average llr was 56.21. For the incorrectly extracted terms (noise), these averages were only 4.23 and 6.84 respectively, which means a difference of 7.89 points for the termhood measure and 49.37 points for llr. These findings confirm that both measures are informative of termhood. However, since TExSIS sorts based on Vintar's termhood measure, rather than on llr, it was surprising to find that llr seemed so much more informative in this comparison. To examine this in more detail, the evolution of precision, recall and F-score was plotted for the best ranked terms, first when sorted by Vintar's termhood measure, then when sorted by llr. Since the minimum number of extracted terms was 3884, we looked at the best ranked 3884 terms for each corpus. The results are presented in Figs. 8 and 9. Precision, recall and F-score were averaged over all corpora. A logical pattern would be for precision to start very high and recall very low, changing to cross each other at some point. While this is true for recall, precision does not start high in either case and varies only slightly throughout; it even starts to increase slightly towards the end. Precision for the termhood measure starts with at least a very small

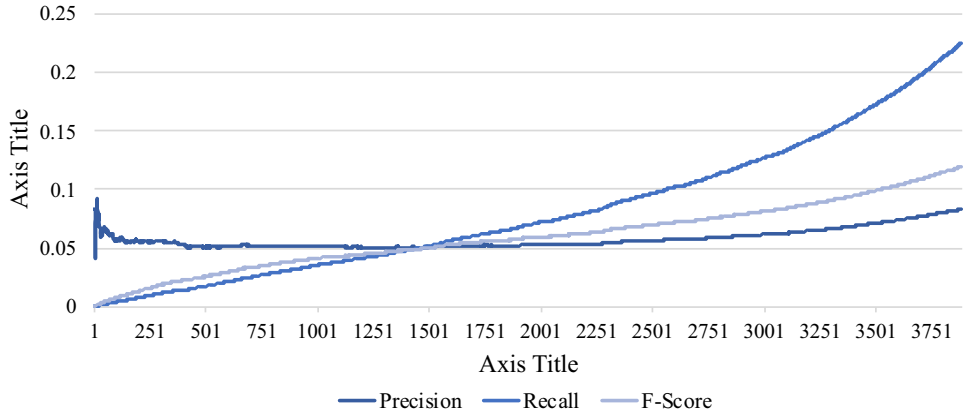

Fig. 8 Evolution of precision, recall and F-score when candidates are sorted by Vintar's termhood measure

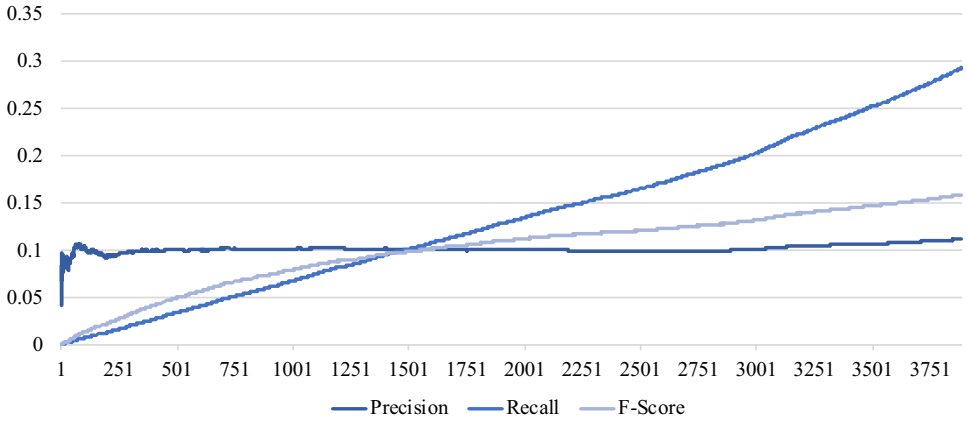

Fig. 9 Evolution of precision, recall and F-score when candidates are sorted log-likelihood ratio

peak, but overall, performance is slightly better when term candidates are sorted based on llr. The fact that even the highest ranked term candidates do not reach a higher precision is an indication that these statistical measures fail to capture important term characteristics.

Next, the distribution of the different term labels is compared in the gold standard, versus the correctly extracted terms and the terms that should have been extracted, but were not (silence). The greatest difference was found for the NEs. On average (across all languages and domains), 21% of all unique annotations in the gold standard were NEs. Of the correctly extracted terms, this was only 17%, versus 28% on average for the silence, indicating that TExSIS is worse at identifying NEs than other terms. This is hardly surprising, since TExSIS was mainly designed for ATE and the NER module was not the focus of the tool. Conversely, TExSIS does seem to perform well for the Specific Terms and Common Terms, with larger proportions of each among the correctly extracted terms than among the silence (average difference of 3% for each).

Many other automatic evaluations could be performed by comparing the ATE output to the dataset, such as evaluations of the number of terms extracted, the term

length, the part-of-speech patterns or variations between the different domains and languages. However, the analyses presented suffice to show the practical usefulness of the datasets as gold standards for the evaluation of ATE.

5 Multilingual gold standard for ATECC

5.1 Gold standard construction

Some of the main questions a good gold standard for ATECC should ideally be able to answer are:

1. Are the suggested source language and target language term candidates both actual terms?
2. Is the suggested translation equivalent correct?
   If not:

   (a) How wrong is the suggested equivalent? Is it at least semantically related to the source term?
   (b) Was a correct equivalent even available in the corpus?

   If so:

   (c) Are there other translation equivalents in the corpus for this source term? Maybe a lower-ranked translation suggestion is also correct?

The first question can already be answered by using the datasets presented in the previous section. The second question is more difficult, especially when considering the three additional, related questions. Knowing the answer to question 2a would be useful for a more fine-grained analysis. Suggesting *ventricular* instead of *ventricle* as an English equivalent for the Dutch *ventrikel* does not seem as wrong as suggesting *heart failure* would be. Additionally, judging translation equivalence is not always a simple binary decision (Le Serrec 2012), so including strongly semantically related terms may provide a way to capture other acceptable translations. To evaluate this would require knowledge of all related terms in a corpus. The importance of question 2b has been discussed before: knowing whether a correct equivalent was available in the corpus at all would help to identify which needs to be improved: the input corpus or the system. This question can only be answered when all potential equivalents have been identified in an entire corpus. This same information could be used to answer the final question, 2c, which is especially useful considering that most ATECC systems provide a ranked list of translation suggestions for each source term.

With these considerations in mind, a methodology was developed for the annotation of a gold standard for ATECC. The corpus on heart failure (including all three languages) was chosen for this purpose since its moderate size made the task manageable, but the density of terms still makes the corpus relevant for ATECC.

Moreover, terminology research is often performed in the medical domain, and annotators all felt the medical corpus was least difficult to annotate for the monolingual term annotation, which likely benefited the quality of the annotations. This corpus also has the highest number and proportion of Specific Terms, which are likely the most interesting for most applications. The magnitude of this task did not allow us to use multiple annotators or calculate inter-annotator agreement. As before, the entire task was performed by a single annotator and semi-automatic checks were performed to make the annotations as consistent as possible. Finally, a domain-specialist was consulted for the most difficult cases to compensate for the annotator's lack of domain-knowledge.

Since term translations in the three languages rarely consist of one option for each language, aligning the translations in three columns was no option. Consequently, IDs (unique identification numbers) were assigned to each term, so translations could be indicated by referring to the ID(s) of the respective term(s). Annotations that were the same except for capitalisation were combined when they had also received the same term label. Term variation was another factor to consider. Many researchers have somehow integrated term variation into their annotations (Bernier-Colborne 2012; Loginova et al. 2012; Schumann and Fischer 2016) and, combining insights from previous research, three categories of term variants were defined: *synonyms*, *abbreviations* and *alternative spellings*. Additionally, the *lemma* was manually added, so variations of the terms with the same lemma could easily be connected as well. To enable an even more fine-grained evaluation of the translations, three additional categories were included: *hypernyms*, *hyponyms* and *other*. The latter was included because many terms are in some way related, but difficult to capture in specific categories. This category is only vaguely defined and leaves room for interpretation, but was necessary to facilitate annotation and avoid the unnecessary accumulation of categories. It includes related terms sharing a similar root but a different POS (e.g. "ventricle" and "ventricular"), adjectives and noun phrases including those adjectives (e.g. "left ventricular ejection fraction" and "ventricular") and other, undefined connections. Room was also left for *notes*, in case the annotator felt the annotations needed further explanation. Finally, the information which could be extracted automatically from the monolingual annotations was included as well (*label*, *frequency* and *texts* in which the annotation occurs), to make the gold standard as comprehensive as possible.

Table 4 is an example of one of the resulting annotations. The annotation *beta-blockers* received the ID *112*. It was labelled as a Specific Term and appeared four times in this form in two different texts. It was an English term and many different equivalents were found in French and in Dutch, such as annotations 2801 and 5998, which refer to *β-bloquant* and *bètablokkeerder* respectively. The lemma was manually added, so the link to the singular form *beta-blocker* could be made automatically. Examples of synonyms, abbreviations, alternative spellings, hypernyms, hyponyms and others are, in that order, *1971 = beta receptor blockers*, *2567 = BB*, *2099 = beta blockers*, *87 = medication*, *235 = nonselective beta-blockers* and *1027 = beta blockade*. The annotator left no notes for this example.

Table 4 Example annotation in gold standard for ATECC

ID	112
Annotation	Beta-blockers
Label	Specific term
Frequency	4
Texts	144; 096
English	EN
French	2801; 3664; 4738; 5268; 4851; 4867; 2953; 4076; 3870; 4769
Dutch	5998; 7558; 5774; 6015; 6329; 6559; 7040; 6183; 7129; 5354; 6327; 6391; 5299; 6052; 7158
Lemma	Beta-blocker
Synonyms	1971; 1450
Abbreviations	2567
Alternative spellings	2099; 1509; 2243
Hypernyms	87; 393; 1430; 1893; 1303; 111; 1926; 1752
Hyponyms	235; 1577; 2441; 2324; 2669; 222
Other	1027; 1035; 2462; 1563; 776; 724; 882; 1632; 1789; 633; 1916; 789; 2214; 2531; 2430; 748
Notes	None

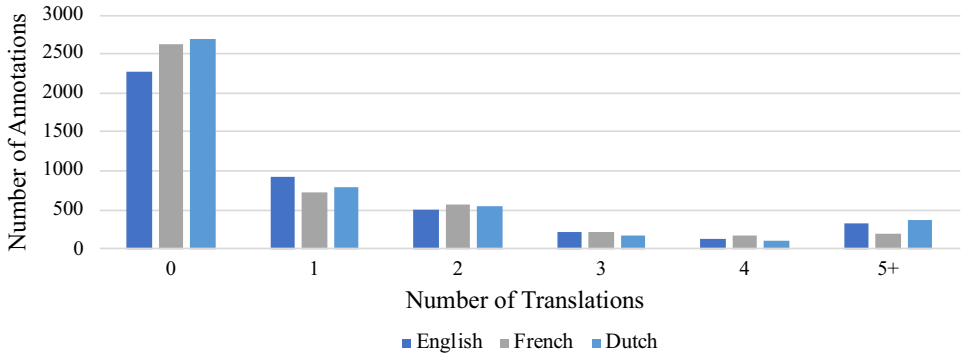

Fig. 10 Number of annotations per number of translations in both target languages, per source language and excluding NEs

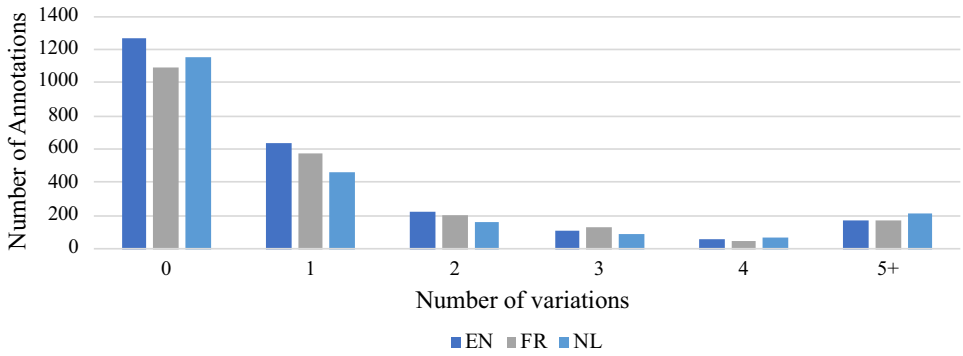

Fig. 11 Number of annotations per number of variations per source language, excluding NEs

To summarise, the time and effort required for this task should not be underestimated. The time spent on the annotation is extremely variable, depending on many factors, such as the annotator (experience with annotation, the subject, the language and concentration level), the text (subject, language, difficulty, concentration of terms) and other factors like how often the annotator had to return to previous annotations and the speed of the annotation software etc. As an example: simply annotating an average text of about 800 tokens in ideal circumstances (very experienced annotator, domain-specialist, native language), without going back for additional corrections or checks, took approximately 10 min for 137 annotations among those 800 tokens. That could really be considered the upper bound, since conditions are rarely that good and it doesn't take into account setting up the annotation, the learning process, reviews and edits, potential technical problems or any other distractions. The total process was the work of many months. Despite this investment of time and our best efforts to create the optimal setup, human error and a certain degree of subjectivity cannot be completely eliminated. However, this was kept to a minimum by working with a single, experienced annotator and a format that allowed several automatic checks to detect errors and inconsistencies, e.g. checking if references in the category *French* correctly refer to French annotations, removing doubles after manual checks, etc.

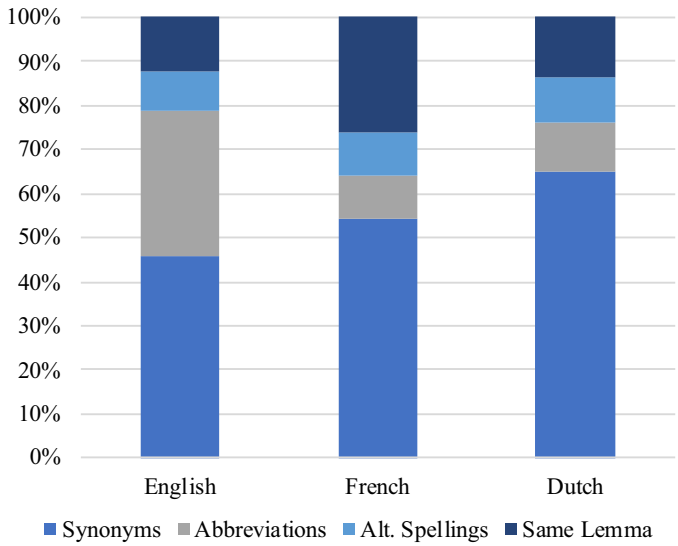

Fig. 12 Types of variations per language

5.2 Analysis

7385 unique annotations were extracted from the monolingual GSs. The first and most striking observation is the high term variation. All term annotations (excluding NEs) have, on average, 1.29 term variations (synonyms, abbreviations, alternative spellings and annotations with the same lemma). This has a direct impact on the number of translations per annotation (term variants normally have the same translations), resulting in an average number of 2.40 translations per annotation (± 1.2 per language). A more detailed look at the numbers revealed a Zipf-like distribution, as illustrated in Figs. 10 and 11.

As can be seen, many annotations have no variations or translations. However, some have more than ten, e.g. the English term "beta-blockers" (see also Table 4). The French translation for this annotation has ten variations: *β-bloquant, ß-bloquants, ß-bloquants, bétabloquants, bêtabloquant, bêtabloquants, bêta bloquants, bêta-bloquant, bêta-bloquants* and *bêtabloqueur*. Even removing those with the same lemma, there are still seven variants. These variations, while closely related, are written differently and an ATE system would therefore not automatically know they are related. Since term frequency is such an important factor, a system that can connect variants would have a great advantage. The joint frequency of all French variants is 29, but separately, only one variant has a frequency of more than five. Variation is common in all languages, though slightly more in Dutch (1.47 variations on average, versus 1.14 in English and 1.30 in French). The translations were studied in more detail in (Rigouts Terryn et al. 2018), which includes an investigation of how different part-of-speech patterns are translated.

A more in-depth analysis of the types of variations was performed, revealing clear differences between the three languages. As shown in Fig. 12, the four types of variations (synonyms, abbreviations, alternative spellings and terms with the same lemma) occur in different proportions in the three languages. Only the proportion of alternative spellings was somewhat consistent; abbreviations were most popular in English, terms with the same lemma occurred most in the French corpora and, in

Dutch, synonyms were used more than in the other two languages (though they were the most common type of variation in all languages). The popularity of variation in the form of terms with the same lemma in French can reasonably be explained by the presence of more morphological variation (e.g. male and female forms), but the other two are more difficult to explain. Maybe the popularity of synonyms in Dutch can be attributed to the fact that it is a smaller language, where there is less standardised terminology, leading to the use of more synonymous terms. The same reasoning could apply to the abundance of abbreviations in English: it is the biggest language in scientific communication, so there is more standardised terminology, which is still recognisable when abbreviated.

While these observations are limited to a modest corpus of specific texts and cannot be generalised, they may provide inspiration on handling variation for ATECC. If there is this much variation in a clean, focussed corpus of published medical abstracts and short papers, it may be even more prevalent in corpora including texts that were not written by professionals or subjected to an editing procedure before publication. One possible conclusion is that improving lemmatisation and normalisation could greatly benefit ATECC. Moreover, since even a clean and focussed comparable corpus such as the one used here only contained translations for a limited portion of all terms, then expecting perfect translation coverage from comparable corpora is not realistic. Therefore, the possibility to detect the source of mistakes is a valuable feature of the gold standard.

Besides term variation, the gold standard also provides information about hypernyms, hyponyms and "other" connections. On average, annotations have eleven of these connections. Almost all terms have at least one connection; only 0.06% of all annotations are completely isolated. However, as with variations and translations, most annotations have only a couple of connections and the average is higher because of a few terms with a lot of connections (e.g. *maladie* (English: *disease*)) has 435 connections, because all diseases mentioned in the corpus are hyponyms). Not only does this information allow a more fine-grained evaluation of ATECC output (e.g. if a hyponym of the correct translation is suggested, a system could be penalised less than when a completely unrelated translation is suggested), it is also a useful resource for related tasks, such as hypernym detection (Rigouts Terryn et al. 2016). In conclusion, the gold standard enables a fine-grained evaluation of ATECC, both for individual systems and for benchmarking in comparison with other systems, and it is a valuable source of information about the nature of terms.

6 Conclusions and future research

Automatic term extraction is a productive field of research and a preprocessing step for many other NLP tasks. However, there are two major obstacles related to data, namely the shortage of well-documented, domain- and language-independent gold standards and the lack of good training datasets for machine learning approaches. For multilingual automatic term extraction from comparable corpora, there is an even greater shortage of gold standards and constructing a gold standard for this task

presents an even greater challenge. The aim of the research presented in this paper was to construct detailed, manually annotated, high-quality gold standard datasets for both tasks, which were specifically designed to be easily re-usable.

Corpora were collected and described in three languages (English, French and Dutch) and four domains (corruption, dressage, heart failure and wind energy). An annotation scheme was developed and tested with three term labels (Specific Terms, Common Terms and OOD Terms) based on two parameters (Domain-specificity and Lexicon-specificity) and with an additional label for Named Entities. Around 50 k tokens were annotated per corpus according to this scheme and with elaborate guidelines, resulting in over 100 k annotations. These datasets can serve as rich sources of information about terminology, as training data for machine learning approaches or as gold standards, which was demonstrated by a use case with the TExSIS system.

An entirely new methodology was developed for the construction of a gold standard for ATECC, which was designed to allow a fine-grained and detailed evaluation. The annotation was performed on the trilingual comparable corpus about heart failure. The gold standard contains information about all terms and NEs in the corpus, all possible translation equivalents among the annotations, variations found of each annotation in the corpus and strongly related terms, such as hypernyms and hyponyms. Similar to the monolingually annotated datasets, this gold standard cannot only be used for evaluation purposes, but also as a rich source of information about terminology. At the end of this research project, all datasets will be made publicly available.

The next step will be to test supervised machine learning approaches on the datasets and to explore customisation options for different languages, domains and applications. Which features can be used to extract terms in which settings? Can term features be combined to automatically distinguish between different types of terms, based on the proposed annotation scheme? Furthermore, a use case should further validate the multilingual gold standard for ATECC.

Acknowledgements This research has been carried out, first, in the framework of the SCATE project, funded by the Flemish agency for Innovation and Technology (IWT) under Project Number 130041 and, second, as part of a Ph.D. fellowship on the EXTRACT project, funded by the Research Foundation—Flanders.

References

Aker, A., Paramita, M., & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 402–411). Sofia: ACL.

Amjadian, E., Inkpen, D., Sima Paribakht, T., & Faez, F. (2016). Local-global vectors to improve unigram terminology extraction. In *Proceedings of the 5th international workshop on computational terminology* (pp. 2–11). Osaka, Japan.

Astrakhantsev, N. (2017). ATR4S: Toolkit with state-of-the-art automatic terms recognition methods in Scala. *Language Resources and Evaluation, 52,* 853–872.

Astrakhantsev, N., Fedorenko, D., & Turdakov, D Yu. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software, 41*(6), 336–349.

Azé, J., Roche, M., Kodratoff, Y., & Sebag, M. (2005). Preference learning in terminology extraction: A ROC-based approach. *Proceeedings of Applied Stochastic Models and Data Analysis* (pp. 209–2019). France: Brest.

Bada, Michael, Eckert, Miriam, Evans, Donald, Garcia, Kristin, Shipley, Krista, Sitnikov, Dmitry, et al. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics, 13,* 161–180.

Bada, M., Eckert, M., Hunter, L. E., & Palmer, M. (2010). An overview of the CRAFT concept annotation guidelines. In *Proceedings of the fourth linguistic annotation workshop, ACL 2010* (pp. 207–211). Uppsala: ACL.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*. Lisbon, Portugal.

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: ELRA.

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

Billami, M., Camacho-Collados, J., Jacquey, E., & Kister, L. (2014). Annotation Sémantique et Validation Terminologique En Texte Intégral En SHS. In *Proceedings of TALN 2014* (pp. 363–376). Marseille, France.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., et al. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the international conference on language resources and evaluation, LREC 2008*. Marrakech, Morocco: ELRA.

Bordea, G., Buitelaar, P., & Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *Proceedings of the 10th international conference for terminology and artificial intelligence (TIA)*. Paris, France.

Cabré, M. T. (1999). Terminology. Theory, methods and applications. In J. C. Sager (Ed.), *Terminology and lexicography research and practice*. Amsterdam: John Benjamins.

Chen, Z., & Yan, E. (2017). Domain-independent term extraction & term network for scientific publications. In *IConference 2017 Proceedings* (pp. 171–189).

Condamines, A. (2017). The emotional dimension in terminological variation. The example of transitivization of the locative complement in fishing. In P. Drouin, A. Francœur, J. Humbley, & A. Picton (Eds.), *Multiple perspectives on terminological variation* (pp. 11–30)., Terminology and Lexicography Research and Practice 18 Amsterdam: John Benjamins.

Conrado, M. S., Salgueiro Pardo, T. A., & Rezende, S. O. (2013). A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the NAACL HLT 2013 student research workshop* (pp. 16–23). Atlanta, GA: ACL.

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.

De Clercq, O., & Hoste, V. (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics, 42*(3), 457–490.

Delpech, E. (2011). Evaluation of terminologies acquired from comparable corpora: An application perspective. In *NODALIDA 2011* (pp. 66–73). Riga, Latvia.

Delpech, E., Daille, B., Morin, E., & Lemaire, C. (2012). Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *Proceedings of the 24th international conference on computational linguistics (COLING 2012)* (pp. 745–761). Mumbai: ACL.

Dobrov, B., & Loukachevitch, N. (2011). Multiple evidence for term extraction in broad domains. In *Proceedings of the international conference recent advances in natural language processing 2011* (pp. 710–715). Hissar, Bulgaria.

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

Drouin, P., Grabar, N., Hamon, T., Kageura, K., & Takeuchi, K. (2018a). Computational terminology and filtering of terminological information: Introduction to the special issue. *Terminology, 24*(1), 1–6.

Drouin, P., L'Homme, M.-C., & Robichaud, B. (2018). Lexical profiling of environmental corpora. In *Proceedings of LREC 2018* (pp. 3419–3425). Miyazaki: ELRA.

Enguehard, C. (2003). CoRRecT: Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes. In *Actes de la 10eme conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)* (pp. 339–345). Nancy, France.

Estopà, R. (2001). Les unités de signification spécialisées élargissant l'objet du travail en terminologie. *Terminology, 7*(2), 217–237.

Estopà, R., Vivaldi, J., & Cabré, M. T. (2000). Extraction of monolexical terminological units: Requirement analysis. In *Workshop proceedings second international conference on language resources and evaluation. Terminology resources and computation* (pp. 51–56). Athens: ACT.

Faber, P., & Rodríguez, C. I. L. (2012). Terminology and specialized language. In P. Faber (Ed.), *A cognitive linguistics view of terminology and specialized language* (pp. 9–32). Berlin: Walter de Gruyter GmbH & Co.

Fedorenko, D., Astrakhantsev, N., & Turdakov, D. (2013). Automatic recognition of domain-specific terms: An experimental evaluation. In *Proceedings of the ninth spring researcher's colloquium on database and information systems* (Vol. 26, pp. 15–23). Kazan, Russia.

Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing, 6*(3), 145–179.

Ghazzawi, N., Robichaud, B., Drouin, P., & Sadat, F. (2018). Automatic extraction of specialized verbal units: A comparative study on Arabic, English and French. *Terminology, 23*(2), 207–237.

Gornostay, T., Gojun, A., Weller, M., Heid, U., Morin, E., Daille, B., et al. (2012). Terminology extraction, translation tools and comparable corpora: TTC concept, midterm progress and achieved results. In *LREC 2012 workshop on creating cross-language resources for disconnected languages and styles (CREDISLAS)* (p. 4).

Gurrutxaga, A., Leturia, I., Saralegi, X., & Vicente, I. S. (2013). Automatic comparable web corpora collection and bilingual terminology extraction for specialized dictionary making. In S. Sharoff, R. Rapp, P. Zweigenbaum, & P. Fung (Eds.), *Building and using comparable corpora* (pp. 51–75). Berlin: Springer.

Haque, R., Penkale, S., & Way, A. (2018). TermFinder: Log-likelihood comparison and phrase- based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation, 52,* 365–400.

Hätty, A., Dorna, M., & Schulte im Walde, S. (2017). Evaluating the reliability and interaction of recursively used feature classes for terminology extraction. In *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics* (pp. 113–121).

Hätty, A., & Schulte im Walde, S. (2018). A laypeople study on terminology identification across domains and task definitions. In *Proceedings of NAACL-HLT 2018* (pp. 321–326). New Orleans: ACL.

Hätty, A., Tannert, S., & Heid, U. (2017). Creating a gold standard corpus for terminological annotation from online forum data. In *Proceedings of language, ontology, terminology and knowledge structures workshop (LOTKS 2017).*

Hazem, A., & Morin, E. (2016a). Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3401–3411). Osaka, Japan.

Hazem, A., & Morin, E. (2016b). Improving bilingual terminology extraction from comparable corpora via multiple word-space models. In *Proceedings of LREC 2016* (pp. 4184–4187). Portorož: ELRA.

Hoffmann, L. (1985). *Kommunikationsmittel Fachsprache*. Tübingen: Gunter Narr Verlag.

Hoste, V., Vanopstal, K., Terryn, A. R., & Lefever, E. (2019). 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, 20*(2).

Inkpen, D., Sima Paribakht, T., Faez, F., & Amjadian, E. (2016). Term evaluator: A tool for terminology annotation and evaluation. *International Journal of Computational Linguistics and Applications, 7*(2), 145–165.

Jacquemin, C., & Bourigault, D. (2003). Term extraction and automatic indexing. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 599–615). Oxford: Oxford University Press.

Judea, A., Schütze, H., & Brügmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 290–300). Dublin, Ireland.

Justeson, J., & Katz, S. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1*(1), 9–27.

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

Kageura, K., Yoshioka, M., Tsuji, K., Yoshikane, F., Takeuchi, K., & Koyama, T. (1999). Evaluation of the term recognition task. In *Proceedings of the first NTCIR workshop on research in japanese text retrieval and term recognition* (pp. 417–434). Tokyo, Japan.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics, 19*(1), 180–182.

Kontonatsios, G. (2015). *Automatic compilation of bilingual terminologies from comparable corpora*. Doctor of Philosophy, University of Manchester, Manchester.

L'Homme, M.-C., Benali, L., Bertrand, C., & Lauduique, P. (1996). Definition of an evaluation grid for term-extraction software. *Terminology, 3*(2), 291–312.

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.

Le Serrec, A. (2012). *Analyse Comparative de l'équivalence Terminologique En Corpus Parallèle et En Corpus Comparable: Application Au Domaine Du Changement Climatique*. Doctor of Philosophy, Université de Montréal.

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: ACL.

Macken, L., Lefever, E., & Hoste, V. (2013). TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology, 19*(1), 1–30.

Morin, E., & Hazem, A. (2014). Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 1284–1293). Baltimore, MA: Association for Computational Linguistics.

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.

Mustafa El Hadi, W., Timimi, I., Dabbadie, M., Choukri, K., Hamon, O., & Chiao, Y.-C. (2006). Terminological resources acquisition tools: toward a user-oriented evaluation model. In *Proceedings of LREC 2006* (pp. 945–948). Genoa: ELRA.

Nazar, R. (2016). Distributional analysis applied to terminology extraction. *Terminology, 22*(2), 141–170.

Nazarenko, A., & Zargayouna, H. (2009). Evaluating term extraction. In *Proceedings of the international conference RANLP-2009* (pp. 299–304). Borovets: ACL.

Nenadić, G., & Ananiadou, S. (2006). Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing, 5*(1), 22–43.

Nenadić, G., Ananiadou, S., & McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. In *Proceedings of COLING 2004* (pp. 604–610). ACL.

Panou, D. (2013). Equivalence in translation theories: A Critical evaluation. *Theory and Practice in Language Studies, 3*(1), 1–6.

Patry, A., & Langlais, P. (2005). Corpus-based terminology extraction. In *Terminology and content development—Proceedings of the 7th international conference on terminology and knowledge engineering* (pp. 313–321). Copenhagen, Denmark.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis (Ed.), *Knowledge mining* (Vol. 185, pp. 255–279). Berlin: Springer.

Pearson, J. (1998). Terms in context. In E. Tognini-Bonelli (Ed.), *Studies in corpus linguistics* (Vol. 1). Amsterdam: John Benjamins.

Projet TermITH. (2014). *Annotation Sémantique et Terminologique Avec La Plateforme SMARTIES*.

Qasemizadeh, B., & Handschuh, S. (2014). The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of COLING 2014: 4th international workshop on computational terminology* (pp. 52–63). Dublin, Ireland.

Qasemizadeh, B., & Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of LREC 2016* (pp. 1862–1868). Portorož: ELRA.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2018). A gold standard for multilingual automatic term extraction from comparable corpora: Term structure and translation equivalents. In *Proceedings of LREC 2018*. Miyazaki: ELRA.

Rigouts Terryn, A., Macken, L., & Lefever, E. (2016). Dutch hypernym detection: does decompounding help? In *Proceedings of the second joint workshop on language and ontology & terminology and knowledge structures (LangOnto2 + TermiKs)* (pp. 74–78). Portorož: ELRA.

Saralegi, X., San Vicente, I., & Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of building and using comparable corpora workshop* (pp. 27–32). Marrakech, Morocco.

Sauron, V. (2002). Tearing out the terms: Evaluating term extractors. In *Proceedings of the twenty-fourth international conference on translating and the computer*. London: ASLIB.

Schumann, A.-K., & Fischer, S. (2016). Compasses, magnets, water microscopes. In *Proceedings of LREC 2016* (pp. 3578–3784). Portorož: ELRA.

Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.-D., & Tsujii, J. (2011). *BioNLP Shared task 2011: Supporting resources*. In *Proceedings of BioNLP shared task 2011 workshop*.

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.

Vintar, S. (2010). Bilingual term recognition revisited. *Terminology, 16*(2), 141–158.

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

Warburton, K. (2013). Processing terminology for the translation pipeline. *Terminology, 19*(1), 93–111.

Wermter, J., & Hahn, U. (2005). Finding new terminology in very large corpora. In *Proceedings of the 3rd international conference on knowledge capture (K-CAP 2005)* (pp. 137–144). Alberta: ACM Press.

Wolf, P., Bernardini, U., Federmann, C., & Sabine, H. (2011). From statistical term extraction to hybrid machine translation. In M. L. Forcada, H. Depraetere, & V. Vandeghinste (Eds.), *Proceedings of the 15th conference of the European association for machine translation* (pp. 225–232). Leuven, Belgium.

Wong, W. (2009). Determination of unithood and termhood for term recognition. In *Handbook of research on text and web mining technologies* (pp. 500–529). IGI Global.

Zhang, Z., Gao, J., & Ciravegna, F. (2018). SemReRank—Improving automatic term extraction by incorporating semantic relatedness with personalised PageRank. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/3201408.

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *Proceedings of the international conference on language resources and evaluation, LREC 2008* (pp 2108–2113). Marrakech, Morocco.

Zhang, Y., Milios, E., Zincir-Heywood, N. (2004). A comparison of keyword- and keyterm-based methods for automatic web site summarization. In *Technical report WS-04-01, papers from the AAAI'04 workshop on adaptive text extraction and mining* (pp. 15–20). San José, CA: ACL.