Reference Lists for the Evaluation of Term Extraction Tools

Elizaveta Loginova Clouet, Anita Gojun, Helena Blancafort, Marie Guegan, Tatiana Gornostay, Ulrich Heid

Reference Lists for the Evaluation of Term Extraction Tools

Elizaveta Loginova[1], Anita Gojun[2], Helena Blancafort[3], Marie Guégan[3], Tatiana Gornostay[4], and Ulrich Heid[2]

[1] UN Lina, University of Nantes
elizaveta.loginova@univ-nantes.fr
[2] IMS, University of Stuttgart
{gojunaa,heid}@ims.uni-stuttgart.de
[3] Syllabs, France
{blancafort,guegan}@syllabs.com
[4] Tilde, Latvia
tatiana.gornostay@tilde.lv

Abstract. In this paper, we discuss practical and methodological issues of the creation of reference term lists (RTLs) for the evaluation of monolingual and bilingual term candidate extraction from comparable corpora in the domains of wind energy and mobile technology. These reference term lists are intended to serve as a "gold standard" for the qualitative and quantitative evaluation of automatic term extraction tools. We present the preliminary results of the evaluation of the monolingual term extraction. Using the manually collected RTLs, we evaluated monolingual term candidate lists which are automatically extracted from the Spanish texts in the domain of wind energy.

Keywords: terminology extraction, multilingual context, comparable corpora, reference lists, evaluation

1 Introduction

1.1 Context

In the FP7 EU project *Terminology Extraction, Translation Tools and Comparable Corpora* (TTC), tools for the automatic extraction of bilingual terminology from the domain-specific corpora are being developed. Since domain-specific parallel corpora are scarce, our tools aim at extracting bilingual terminology from comparable corpora. These are easier to find automatically, e.g., on the Web, also for under-resourced languages. In TTC, we deal with 7 languages from different language families: Germanic: German (DE) and English (EN); Romance: Spanish (ES) and French (FR); Baltic: Latvian (LV); Slavonic: Russian (RU); Sino-Tibetan: Chinese (ZH).

In order to handle all mentioned languages with tools that follow one and the same architecture, we avoid using deep linguistic knowledge within our term extraction processing chain.

In the following, we give a brief overview of the processing steps:

1. **Text crawling**
   The domain-specific texts are collected with a thematic web crawler (also called focused or topic crawler [3]) *Babouk* [10] which has been developed in the TTC project. The crawler takes a list of domain-specific words, called seed terms, as input and outputs the texts found on the Web which deal with the the domain of interest. Seed terms are usually terms representative of the domain for which we want to retrieve the web documents. During the first iteration of the crawling process, the given seeds are expanded to a large terminology using the BootCaT procedure [2]. The output of Babouk are text documents provided in text format with utf8 encoding. Babouk[5] also provides meta data about the crawled texts following the Dublin Core Metadata Initiative[6] , such as information about the source of the text, the publisher, the publishing date, etc. The texts are derived from HTML sites, as well as from PDF and Word files. Since in TTC, we mainly deal with the domain of wind energy and mobile technology, the corpora were compiled for these two domains.
2. **Text pre-processing**
   The domain-specific texts are tokenized, annotated with part-of-speech (POS) tags and with lemmas. For all languages except of ES, LV and ZH, we use the *TreeTagger* [21] for tagging and lemmatization. For ES and ZH we use a proprietary tagger developed by one of the industrial partners. For Latvian, we use the web service which provides the procedures for tagging and lemmatization of Latvian texts based on the proprietary POS tagger for Latvian developed by Tilde [18].
3. **Monolingual terminology extraction**
   In this step, term candidates are extracted from the pre-processed domain-specific texts. Term candidates may be single-word terms (SWTs), as well as multi-word terms (MWTs). The extraction relies on POS patterns which describe nouns (e.g. *"energy"*) or nominal phrases, such as adjective + noun (e.g. *"renewable energy"*), noun + noun (e.g. *"wind energy"*), etc.
   The extraction patterns were collected manually within the project for all seven languages and encoded in the extraction tools developed within TTC. In addition to the POS-pattern based extraction, we developed a knowledge-poor tool which learns POS and POS sequences, and then automatically annotates the noun phrases in a given text. We use [4]'s tool for the POS induction step instead of a supervised PoS tagger and CRF++ for the noun phrase training and tests. The contribution of part-of-speech induction to shallow parsing is reported by [11].
   The identified term candidates are subsequently filtered in order to separate domain-specific terms from general ones. As a filtering measure, we use *weirdness ratio*, the domain specificity value defined by [1].

[5] `http://greenhouse.syllabs.com/ttc/`
[6] `http://dublincore.org/`

4. **Variant recognition**
   Using a set of manually collected language specific equivalence patterns, such as, for example, (EN) N1 N2 ↔ N2 of N1 (*"energy production"* ↔ *"production of energy"*), the term candidate lists are processed further in order to find term variants (cf. e.g. [6] and [23]). The output of the monolingual term extraction and variant recognition step are lists of term candidates with information about their frequencies, domain specificity, POS and variants. The output may be provided in a tab-separated (TSV) format, as well as in the TBX[7] format.
5. **Bilingual term alignment**
   In this step, the equivalent term candidates in two different languages are identified. For the alignment of SWTs, we use a standard context-based approach [19], as well as an approch for aligning words with neoclassical stems [12]. For MWTs, we use compositional alignment as described in [16]). The output of the term alignment step are lists with source language term candidates and their target language equivalents (cf. table 1).

Table 1. Example of the alignment output. The French term *"énergie renouvelable"* is aligned with three English terms which are sorted by their *alignment scores* which indicates how reliable the alignments are.

énergie renouvelable	renewable energy	[1.0]
	sustainable energy	[1.0]
	renewable power	[1.0]

The result of the bilingual term extraction process[8] described in the preceding paragraphs may be fed into computer-assisted translation (CAT) tools, as well as into machine translation (MT) systems. First experiments regarding the integration of bilingual terminology lists into a standard statistical MT system already showed improvements in the quality of the generated translations.

1.2 Data Used in the Evaluation Experiments

To evaluate extraction tools, we use comparable corpora from the domains of wind energy and mobile technology crawled with Babouk. The same corpora were used to create the RTLs. For some languages, such as LV and DE, the crawled data was not sufficient or not specific enough, therefore we enriched the collection with manually compiled corpora. Size of our reference corpora are between 4,263,336 and 220,823 words, depending on the language and domain (cf. tables 2 and 3).

[7] `http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf`

[8] The TTC tool *TermSuite* implements the entire extraction pipeline for all TTC languages. The tool can be downloaded from: `http://code.google.com/p/ttc-project/`.

Table 2. Wind energy corpora

	Chinese	English	French	German	Latvian	Russian	Spanish
nb of tokens	4,263,336	750,855	710,702	1,700,000	220,823	2 328,609	1,297,338

Table 3. Mobile technology corpora

	Chinese	English	French	German	Latvian	Russian	Spanish
nb of tokens	2,435,232	308,263	302,634	474,316	306,878	372,459	473,273

For the evaluation of both monolingual and bilingual term extraction, we use monolingual and bilingual RTLs, respectively. The RTLs are created manually using the reference corpora from the two mentioned domains. They serve as a "gold standard" of what we consider to be relevant terms of the corpora. The overall approach is thus similar to that of all other natural language processing evaluation tasks: a manually constructed "gold standard" is derived from a set of texts that are then also processed by the tools under evaluation. In the special case of terminology, a number of problems are encountered, some technical, others more theoretical (most prominently the notion of "termhood"). Against the experience from TTC, we will address these issues for both monolingual and bilingual term extraction evaluation. In total, we collected 14 monolingual RTLs and 24 bilingual RTLs which are publicly available.[9]

This paper is structured as follows: in the section 2, we describe the method used to manually collect RTLs and in the section 3, we discuss problems encountered. In section 4, we present the evaluation results of the monolingual term extraction obtained using the manually collected RTLs.[10] We draw a few methodological conclusions in section 5.

2 Creating Monolingual and Bilingual RTLs

2.1 Requirements for Monolingual RTLs

RTLs are intended to evaluate term extraction on the basis of the corpora mentioned above (related to the domains of wind energy and mobile technology). A basic requirement in order to make a quantitative evaluation possible is to ensure that all RTL terms do appear in the reference corpora.

Then, RTLs should reflect the properties of the target technical terminologies, as well as the capabilities of the tools. This includes the treatment of both of both single-word terms (SWTs) and multi-word terms (MWTs), base terms and variants, as well as their linguistic properties.

Variants are either synonymous to base terms or semantically related to them, e.g. through coordination, adjectival modification, etc. Our typology (based on

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

[10] The evaluation of the term alignment procedures is still ongoing and thus not a part of this paper.

[6]) includes graphical variants (EN *offshore* vs. *"off-shore"*), morphological variants (e.g. DE *"solare Energie"*↔*"Solarenergie"* (*"solar energy"*), EN *"synchronous"*↔*"asynchronous"*) and syntactic variants (*"wind energy"* vs. *"wind and solar energy"*). We are aware that other variant types exist in specialized texts (anaphorical, transposition, etc.), but currently they can not be detected by the tools.

We also included paradigmatic variants of MWTs formed by substitution of one of the elements by a synonym: *"wind park"* - *"wind farm"* into RTLs since a term in the source language can be translated by more than one term in the target language. If a tool does not translate a multi-word term by the base term listed in our RTL, but by its paradigmatic variant, it seems correct to count it as a good translation (to illustrate, *"wind farm"* can be translated into Russian as *"ветровой парк"*, i.e. *"wind park"*, because the form *"ветровая ферма"* - *"wind farm"* - is rare in our corpora).

RTLs have to contain various corpus-observable linguistic features of each term in order to allow for the automatic comparison of the tool output and the manual work.

We included the term frequencies, POS-patterns (e.g. *Noun + Adjective*), morphological annotation, e.g. for gender and number, (for most languages we use the Multext tagset [11] annotation), all inflected forms of the term found in the corpus, frequent collocations, and for compound terms, we mark the nature of their components: native vs. neoclassical. A sample bilingual entry is attached in the appendix (table 6).

Initially, for the sake of homogeneity of our RTLs, we determined a fixed distribution of different term types : 20% of SWTs, 20% of single-word compounds and 60% of MWTs. However, during our work, we realised that this requirement is difficult to respect due to the differences between the languages under study. For example, in German, the proportion of compounds is much more important than in English or French since compounding is a very productive word formation process in German. Very often, the equivalence of an English or French MWT is a German compound, a single graphical unit (EN *"rotor blade"* - DE *"Rotorblatt"*, EN *"surface of a rotor"* - DE *"Rotorfläche"*). Thus, we had to reject this initial distribution and take into account the reality of each language. Also, the term distribution over the POS-patterns depends on the language, the domain and likely the text type, so we can not a priori fix proportions.

We had to introduce, though, a constraint of minimum term frequency to be sure that our tools can at all find all the RTL terms in the corpora. We fixed a threshold of 10 for SWTs and of 5 for MWTs.

2.2 Practical Aspects of the Construction of Monolingual RTLs

A simple approximation of the "termhood" of a lexical item is its distribution in the domain-specific texts, comparing its frequency in a specialized corpus with the frequencies in a general corpus (cf. [1]). We use general language

[11] `http://sites.univ-provence.fr/\simveronis/donnees/index.html`

corpora of 10 to 15 Million words depending on the language, mainly consisting of newspaper articles, combined with known reference corpora such as Europarl [14], Wortschatz[12] and the Spanish corpus Ancora[13].

To choose SWTs and MWTs for our RTLs, we used the output of the term extraction tools developed within TTC that sort the terms by frequency of occurrence and relative frequency. Sometimes to take a final decision we examine the context of the candidate terms in the corpora or on the Internet. We examine the collocational usage of SWT candidates which are much more frequent in the specialized corpus than in a general corpus, but not specific enough for being added individually in the SWT list. For example, the English adjectives *"vertical"* and *"horizontal"* are very frequent in the wind energy corpus, but they are not to be considered as domain-specific terms. However, their co-occurrences *"vertical axis"* and *"horizontal axis"* are valid multi-word terms characterizing a wind turbine.

For Latvian, the process of the compilation of reference term lists was fourfold: (1) initially a linguist extracted term candidates manually, (2) then a terminologist validated the list, (3) then the list was checked against another list of automatically extracted term candidates to ensure the frequency of the manually extracted term candidates in the corpus, (4) and finally, a domain specialist was consulted on the termhood and/or unithood of term candidates.

In addition, we apply some of [15]'s linguistic criteria:

- The derivation products of a term are also frequent in the specialized corpus. For instance: FR *"rotor"* – *"rotorique"* (*"(of a) rotor"*) or *"pompe"* – *"pompage"* (*"pump(ing)"*); EN *"wind"* – *"upwind"*, *"downwind"*, *"windmill"*;
- The candidate is considered as a term if it has a paradigmatic relation with the terms already admitted. Such relationships can be found in terminological definitions: the definition of FR *"hélice"* (*"propeller"*) in the *Grand Dictionnaire Terminologique*[14] *("partie du rotor de l'éolienne constituée de l'ensemble des pales et du moyeu")* refers to the terms *"rotor"* (*"rotor"*), *"éolienne"* (*"wind turbine"*), *"pale"* (*"blade"*), *"moyeu"* (*"hub"*); thus it can be admitted as a term.
- Termhood often leads to specialized collocations which are not found in general language: FR *"arbre"* is ambiguous: its general meaning is *"tree"* and its specialized meaning is *"shaft"*, and it appears in this specialized meaning in combinations such as *"arbre lent"* (*"low speed shaft"*) which are atypical for its general meaning *"tree"* (**"low speed tree"*).

We also noticed that some words of foreign origin (mostly English words) can be very frequent in the specialized corpora. For example, the English words *wind, energy, power, speed, system* appear at the top of the French frequency list. The reasons are the following: (i) some recent terms do not have a conventional translation yet, (ii) the English equivalents of important terms used in

[12] `http://wortschatz.uni-leipzig.de/wortschatz`
[13] `http://clic.ub.edu/corpus/en`
[14] `http://www.granddictionnaire.com`

French texts are usually given, and (iii) the texts can include a part in a foreign language (the abstracts of scientific articles, theses, etc). The translations of such frequently used foreign words are potentially good term candidates.

Finally, we checked whether the chosen term candidates are listed in one or more of the large terminology banks or specialized dictionaries (e.g. TERMIUM[15], Grand Dictionnaire Terminologique, IATE[16], EuroTermBank[17].

2.3 Creating Bilingual RTLs

Bilingual RTLs contain equivalents of the monolingual terms identified previously; we created bilingual RTLs for 12 language pairs involving the seven project languages. Given the nature of the monolingual RTLs, the objective of the creation of bilingual RTLs is twofold: (i) to harmonize the monolingual lists and (ii) to provide a correct alignment of term candidates.

The bilingual RTLs include only the terms which appear both in source and target language corpus. Concerning the additional information, the bilingual RTLs contain the same data as the monolingual RTLs, i.e. term variants, POS, corpus frequencies, etc. (see appendix).

In terms of workflow, the monolingual RTLs are created independently for each language by language experts. The harmonization step leads to a reassessment of the contents of these monolingual lists, from a bilingual viewpoint, and items from each input list may be removed during the process. Similarly, target language (TL) equivalents of relevant source language (SL) terms may need to be manually added if they are not in the TL reference list, but found in the TL corpus.

We started from monolingual RTLs with ca. 130 terms per language, with the aim to produce bilingual RTLs with ca. 100 term pairs. However, even a margin of 30% is not enough to have a resulting hundred terms appearing in both monolingual RTLs. The reasons are the following: (i) in the domain of terminology and translation, a term in a source language does not always have an equivalent in the target language, (ii) the frequency of occurrence of the term in the source and target language may vary a lot: a frequent term in one language may only occur once in another language, and (iii) our RTLs are not of a big size (only 100 terms), thus, some terms may appear in a SL reference list, but not in the chosen TL reference list, even if their equivalents are present and frequent enough in the TL corpus.

To resolve this problem, we had to complete the bilingual RTLs with terms appearing in both corpora, but not in both monolingual RTLs. So we had to look for translations of some terms belonging to the source RTL in the corpus of TL.

---

[15] `http://termiumplus.gc.ca`

[16] `iate.europa.eu/`

[17] `http://www.eurotermbank.eu`, etc.

3 Problems in the Creation of RTLs and Some Practical Issues

During the compilation of RTLs we faced a certain number of difficulties that are not new for terminological practice, but nevertheless always problematic. In this section, we describe our difficulties and share some working conventions that we adopted in order to avoid or reduce them.

A major issue, as in all terminology work, is the difficulty to apply the notion of termhood [13]: it might be difficult to decide whether a term corresponds to a specific domain or not, because the notion of termhood depends on several criteria, such as the domain-specific expertise of the linguist or terminologist, the application foreseen, the point of view, etc. [5]. Under *termhood*, we understand "the degree to which a stable lexical unit is related to some domain-specific concepts" [13]. However, "there exists a lack of formal or precise rules which would help us to decide between a term and a non-term. Domain experts (who are not linguists or terminologists) do not always agree on termhood" [8].

All specialized languages show a gradient of domain-specificity; most domains are interdisciplinary, with terms from other domains interfering (e.g. EN *"solar energy"* or French *"géothermie"* - *"geothermic"* - are found in most texts of our corpora on wind energy). So we decided to authorise the inclusion of some very important terms from adjacent domains.

Then, our corpora are not very big and thus, the frequency figures are low which makes the calculation of the domain specificity much harder. In some cases, the terms relevant for the domain have a low frequency in some of our corpora, and we can not add them in RTLs. We keep a constraint of a minimum frequency to assure getting enough contexts for automatic term extraction.

The term status of certain lexical objects is controversial. For example, initially we did not want to count abbreviations as independent term entries. But for the domain of mobile technology, it seems very difficult not to include them, because some abbreviations are very important for the domain and nowadays almost never used in a full form (e.g. *"IP"*, *"GSM"*, *"WLAN"*, etc.).

Unithood [13] and determining whether a multi-word unit is a term or not is another issue we faced with. The boundaries between MWTs and collocations of terms are vague (e.g. is *"energy production"* a term or a collocation?). Even human term identification is not homogenous, for example, in the Latvian language a collocation "vēja enerģijas ražošana" ("wind energy production") is considered to be a term by an experienced terminologist.

To make a decision, we apply some criteria: (i) the extensions of type "Noun + Noun + Noun" of a MWT with a structure "Noun + Noun" are usually collocations (*"wind energy"* → *"wind energy production"*, *"wind energy market"*, *"wind energy sector"*); (ii) the term is a frequent element of other collocations in the corpus (or in the web): *"coal production"*, *"energy production"*, etc.; (iii) we also check the translation of the word in another language: wind energy production" → "Windenergieproduktion" (DE), "producción de energía eólica" (ES), "production d'énergie éolienne" (FR), "vēja enerģijas ražošana" (LV). If in all languages the term is translated while keeping the same elements, (*"wind*

energy" and *"production"* separately), so the whole unit is a collocation, and its elements are terms.

It is not possible to determine an a priori distribution of specific properties for all the languages. We have already mentioned the example of the distribution of compounds. Another example comes from Chinese: in this language the linguistic features included in the RTL with regard to the origin of compound elements (native vs. neoclassical element) do not apply, as neoclassical elements do not exist. The types of encountered variants vary according to the language as well. Thus, the monolingual lists for different languages are not parallel but rather comparable.

Concerning the choice of the main term entry, we select the lemma form, which corresponds to the output of the tools. So, all components of MWTs are lemmatized and appear in the singular form, all adjectives are in masculine (FR *énergie.Fem.Sg, éolien.Masc.Sg = EN "wind energy"*), and not in the canonical form (*énergie.Fem.Sg éolienne.Fem.Sg*). This was a controversial issue as the lemma form, especially for MWTs, does not always correspond to the canonical form which is present in lexicons or traditional terminologies. However, this choice is quite common in the domain of computational terminology, because the lemma is needed for the automatic evaluation of tools for automatic terminology extraction. The canonical form appears in our RTLs in the column "inflected forms" or "most frequent form used".

However, the final choice of the relevant terms for a reference list is still a decision based on domain and specialized language expertise.

4 Using RTLs in Term Extraction Evaluation

In this section, we report on preliminary results of the evaluation of the monolingual candidate terms lists (CTLs) produced by our term extraction tools. We illustrate the challenges of evaluating large terminologies extracted automatically from large data by using a small reference term list.

Experiments were performed on the Spanish data in the domain of wind energy. To generate the candidate term lists (CTLs), we used the domain-specific corpus crawled with Babouk (see section 2.2) containing 1,297,338 tokens. To compute domain-specificity of a term based on the quotient of frequency [1], we applied a general language corpus (newspaper data) with 10,959,833 tokens. The extracted term candidates were subsequently sorted by their domain-specificity (cf. section 2.2).

The reference term list contained 121 reference terms (RT), as well as their lemmatized variants. In total, this leads to 160 reference terms and variants, as a reference term can include variants that are also correct, which means that the evaluation script takes variants into account. Concerning the CTL, our tool output a list of 68,156 terms. The CTL contains all lemmatized single word and multi-word term candidates found in the corpus, including terms with a frequency of 1 (hapax legomena) that represent 28% of the terms occurring in the domain-specific corpus. This CTL could be reduced to 10,845 term candidates,

both MWT and SWT, if we required the minimum frequency of the extracted candidates to be 5. Assumed that the minimum frequency is 10, we get a CTL containing 5,509 term candidates.

The evaluation of large CTLs against monolingual RTLs raised a number of practical choices which should be made beforehand. In this paper we evaluate precision and recall, that are the most frequent used measures to evaluate terminological output [22], and have been borrowed from Information Retrieval [20]. Precision measures the degree of correctness of the term candidates that are suggested as terms while recall measures the degree of comprehensiveness of the list of term candidates. Both measures are useful to assess the output. Recall is the hardest figure to calculate, since it implies reading the whole corpus and selecting manually all the relevant terms.

When the CTL is built using a large corpus, it may contain thousands of term candidates which are then compared with a short list of 160 reference terms. We thus focus on recall, as precision gets rather low in this case. However, as precision is an important value for users, we also measured the precision based on the RTL as well as on a second manual evaluation of the top 500 candidates.

4.1 Evaluation Results

The figure 1 illustrates the results on recall and precision. As expected, recall increases with a higher number of candidates while precision decreases.

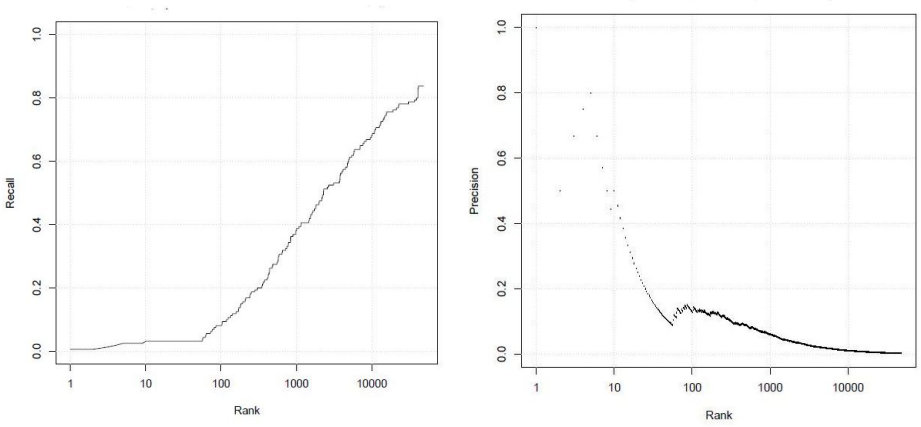

Fig. 1. Recall and Precision on matched output terms

Recall Overall, 90.91% of the reference terms were extracted as candidate terms, as well as 89.38% of the reference terms and their variants. This means that 11 reference terms and 11 reference terms and their 6 variants were missed by our tools. These terms are missed for several reasons:

1. Pos-tagging errors: unknown word to the pos-tagger and not guessed correctly, this is the case for "hélice" (EN propeller) and "buje" (EN hub), that were not given the right pos-tagged (noun).
2. Lemmatization of MWTs: one of the components is not correctly lemmatized by the tool. This is the case for "anemométricos" (EN anemometric) in "dato anemométrico". The used form in the corpus is always the plural "datos anemométricos" that is lemmatized to "dato anemométricos".
3. Mixed entities: our program does not handle mixed entities which are terms including a named entity as "distribución de Weibull" (EN Weibull distribution) and "límite de Betz"(EN Betz' law) and "turbina Daerrius" (EN Daerrius turbine). Mixed entities are not identified by the terminology extractor in Spanish to reduce the noise and avoid a longer list of term candidates.
   We are aware that this case is an exception to the rule stated in section 2.1., where we require the RTLs to reflect the extraction capabilities of the tools under test.
4. Term variants: in this case, the corpus does not contain the reference term, but a variant that is not correct. As corpora are crawled from the web, they are not error-free. For example, the term candidate "rosa de viento" occurs 40 times in the corpus. However, this form is not correct; the correct form is "rosa de los vientos" (EN wind rose), but unfortunately does not occur in the corpus. Here, the linguist did not include in the RTL the form present in the corpus, but the correct form instead.
   Obviously, a more detailed evaluation that would be fully in line with our basic requirement from section 2.1, would have (i) the variant as a part of the RTL and (ii) wanted to count extraction results separately for base terms and variants.

A detailed overview of the growth of recall is illustrated in figure 1, which shows the coverage of the CTL depending on the number of candidate terms. To obtain coverage of 70% of the RTL, 10,000 candidate terms are needed.

Precision We evaluated the precision of the reference terms and their variants according to the number of top candidates, that varies from top 5 to top 500. In the table 4, we can see that precision decreases with a longer list of top candidates and that only 8 RTs appear in the top 100 list. These results were expected, because as we said before, we are comparing a long list of CTs to a short list of RTs. Therefore, we run a second evaluation that consisted in evaluating manually the first 500 CTs and deciding whether CT not included in the RTL could be considered as a term of the domain or not. Here, the precision increased, and we obtained a rate of 100% for the top 5 and top 10 candidates, and 44% for the top 100 list. This means that the output of the tool is quite useful to compile large terminologies with a large number of SWT. From the first top 20 candidates, 17 are SWT. Table 5 shows the first top 20 candidates output by the tools and judged as domain-specific by a linguist (native speaker).

Table 4. Precision evaluation of reference terms and their variants depending on the number of top-candidates

	Evaluation based on the RTL		Evaluation based on human judgement	
Number of top candidates	Precision (%)	Number of reference terms	Precision (%)	Number of correct candidate terms
5	40	2	100	5
10	30	3	100	10
20	30	6	80	16
50	16	8	62	31
100	9	9	44	44
200	4.5	9	32	65
300	3.7	11	30	91
400	4.3	17	26	104
500	3.8	19	23	117

Table 5. Human evaluation of top 20 candidate terms

Rank	**Candidate term (CT)**	**Domain-specific**
1	Energía	Yes
2	**Eólico**	**RTL**
3	Viento	Yes
4	Potencia	Yes
5	**Aerogenerador**	**RTL**
6	Velocidad	Yes
7	Sistema	Yes
8	Eléctrico	Yes
9	Parque	Yes
10	Turbina	Yes
11	Generador	Yes
12	**Parque eólico**	**RTL**
13	Solar	No
14	Tensión	Yes
15	Agua	No
16	**Energía eólico**	**RTL**
17	Red	No
18	**Energía renovable**	**RTL**
19	Instalación	Yes
20	Figura	No

4.2 Discussion

Impact of Lemmatization and Pos-tagging Errors. Coming back to the comparison of a large CTL to a short RTL, the evaluation is based on a lemma comparison, which means that we consider a candidate term as correct if it matches any reference term in the RTL, both lemmatized terms and their lemmatized variants. This means that only matches between lemmas are considered as correct and that a lemma candidate matching a form of the RTL is incorrect. The main drawback of a comparison based on lemma is the strong dependence on lemmatization. The lemmatization is performed automatically and is not error-free. This explains why a CT may be a form and not a lemma and therefore is not evaluated correctly by the program. Typical errors of the linguistic pre-processing step of the corpus can be due to bad lemmatization as well as bad POS tagging and/or POS guessing. [9] report on the positive impact of lemma correction on the quality of term candidate lists. In our Spanish CTL, the term "hélice" (EN propeller) is missing, as it is missing in the lexicon used by the POS tagger and has been guessed as a verb instead of a noun. Moreover, when working with lemmas, the lemmatization procedures of the CTLs and RTLs should be the same. This caused some problems when starting the evaluation experiments, as some errors were found in the first version of the RTLs and had to be corrected, especially concerning MWTs. For the linguist it seems not natural, even if specified in the guidelines, to write MWT in the lemmatized form instead of the canonical form, e.g. we found "energía eólica" instead of "energía eólico" and the canonical form "energía generada" (noun + participle) instead of the lemma "energía generar" (noun + verb). Hence, the RTL needed some corrections and extra reviewing.

Evaluation of Termhood by Experts. It would have been interesting to have an expert of the domain for a more precise evaluation, however, as demonstrated by [7], terminologists and domain experts do not always agree on the termhood. Even among experts there might be differences in the evaluation of termhood. To illustrate this disagreement, [22] report on an experience to evaluate a list of term candidates where only 37% of full agreement was found between the 3 experts participating in the evaluation. Moreover, 26% of terms were chosen by 2 experts (out of 3) and 37% by only one expert. It is important to note that low agreement between evaluators or annotators is a common hurdle in NLP annotation or evaluation activities, especially in word disambiguation tasks and POS tagging [22].

5 Conclusions

The methodology for RTL construction is closely related with the techniques used in term extraction evaluation and with the properties of the extraction tools under study. The following methodological considerations seem to be of particular importance:

- for the evaluation of term extraction tools, reference corpora and RTLs must be very closely related, the RTLs being derived from the reference corpora;
- if RTLs contain term patterns, term variants and corpus frequency data, they can be used as a diagnostic tool, it is then possible to identify, for example, term patterns that are not yet correctly handled by the extraction tools;
- the number of terms contained in an RTL may have an impact on the numeric evaluation results; larger RTLs should provide higher numbers; we expect, however, that the proportions, e.g., between different (variants of) tools should remain constant;
- RTL construction from comparable corpora makes a harmonization step necessary when monolingual RTLs are merged into bilingual ones. This work of RTLs' harmonization makes in evidence the difficult points of terminology extraction: all specialized languages show a gradient of domain-specificity, the terminological status of some lexical units (e.g. abbreviations) is controversial, the boundaries between multi-word term and collocations of terms are not always clear, etc.
- no a priori proportions of specific properties (POS-patterns, SWT vs. MWT, etc.) in the RTLs valid for all the languages is possible;
- the notion of "termhood" is not fully operationalizable and will always introduce an element of arbitariness into the RTLs and into the evaluation based on them, even though the above mentioned procedures are aimed at keeping this element manageably low.

The monolingual RTLs and the pertaining reference corpora are publicly available [18]. The bilingual RTLs will be added as soon as our work is finished.

6 Acknowledgements

We would like to thank the computational linguists and terminologists who participated in the compilation of the RTLs, namely Ana Laguna, Tian Tian, and Somara Seng from Syllabs, Iveta Keiša and Dzintars Skarbovskis from Tilde as well as Elena Umanskaya and Yannan Guo from University of Leeds.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248005.

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

Appendix:

Bilingual term lists

Table 6. An example of a bilingual reference term pair

	Source language	**Target language**
term lemma	Windfarm	wind farm
SWT/MWT	SWT (CP)	MWT
pattern	N N	N N
morph. tag	Ncfsn	Nc-s- Nc-s-
origin	native	native
inflected forms (IF)	Windfarm, Windfarmen	wind farm, wind farms
frequency	75	1975
most frequent IF (mfIF)	Windfarm	wind farm
frequency of mfIF	51	1423
variant	Wind-Farm	-
frequency	1	-
variant type	graphical	-
synonym	yes	-
IFs	Wind-Farm	-
most frequent IF	Wind-Farm	-
frequency of mfIF	1	
collocational use	Erweiterung von Windfarmen	offshore wind farm

References

1. Ahmad, K., Davies, A., Fulford, H. and Rogers, M.: *What is a term? The semi-automatic extraction of terms from text.* Translation Studies: An Interdiscipline, John Benjamins, Amsterdam, pp. 267-278, (1994).
2. Baroni, M. and Bernardini, S.: *BootCaT: Bootstrapping corpora and terms from the web* Proceedings of LREC 2004, (2004).
3. Chakrabarti, S., Van den Berg, M. and Dom, B.: *Focused crawling: a new approach to topic-specific Web resource discovery.* Computer Networks, 31(11-16): pp.1623–1640,(1999).
4. Clark, A.: *Combining distributional and morphological information for part of speech induction.* Proceedings of the tenth Annual Meeting of the European Association for Computational Linguistics (EACL), pp. 59-66, (2003).
5. Condamines, A.: *Variations in Terminology.* Application to the Management of Risks Related to Language Use in the Workplace. Terminology vol. 16(1), (2010)
6. Daille, B.: *Variants and application-oriented terminology engineering.* Terminology, vol. 1, pp. 181-197, (2005).
7. Estopà, R.: *Extracció de la terminologia: elements per a la construcció d'un SEACUSE.* IULA, (1999)

8. Frantzi, K. T. and Ananiadou, S., and Tsujii, J. *The C-value/NC-value Method of Automatic Recognition for Multi-word Terms.* G. Goos, J. Hartmanis and J. van Leeuwen (Eds.), *Research and Advanced Technology for Digital Libraries: Proceedings of the Second European Conference, ECDL'98* (Vol. 1513, pp. 585–604). Lecture Notes in Computer Science. Berlin / Heidelberg: Springer, (1998).
9. Gojun, A., Heid, U., Weissbach, B., Loth, C. and Mingers, I.: *Adapting and evaluating a generic term extraction tool.* Proceeding of the 8th international conference on Language Resources and Evaluation (LREC), (2012)
10. de Groc, C.: *Babouk: Focused web crawling for corpus compilation and automatic terminology extraction.* Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon, France, (2011).
11. Guégan M. and de Loupy, C.: *Knowledge-poor approach to shallow parsing: Contribution of unsupervised part-of-speech induction.* Recent Advances in Natural Language Processing (RANLP), (2011)
12. Harastani, R., Daille, B. and Morin, E.: *Neoclassical Compound Alignments from Comparable Corpora.* CICLing (2), pp. 72-82, (2012)
13. Kageura, K., and Umino, B.: *Methods notion of automatic term recognition. Terminology.* International Journal of Theoretical and Applied Issues in Specialized Communication, 3, pp. 259–289, (1996).
14. Koehn, P.: *Europarl: A Parallel Corpus for Statistical Machine Translation.* MT Summit, (2005)
15. L'Homme, M.-C.: *La terminologie: principes et techniques.* Montréal, Les Presses de l'Université de Montréal, Coll. "Paramètres", (2004)
16. Morin, E. and Daille, B.: *Compositionality and lexical alignment of multi-word terms.* Language Resources and Evaluation, volume 44, pp. 79-95, (2009)
17. Nazarenko, N. and Zargayouna, H. *Evaluating term extraction* RANLP 09, Borovets Bulgaria, (2009)
18. Pinnis, M. and Goba, K. *Maximum Entropy Model for Disambiguation of Rich Morphological Tags.* Proceedings of the 2nd Workshop on Systems and Frameworks for Computational Morphology (SFCM2011), Zürich, 26 August 2011, Springer, Heidelberg, Communications in Computer and Information Science, 1, Volume 100, pp. 14-22, (2011).
19. Rapp, R.: *Automatic Identification of Word Translation from Unrelated English and German Corpora.* Proceedings of the 37th annual meeting of the association for computational linguistics (ACL '99). College Park, Maryland, USA, pp. 519–526, (1999)
20. Salton and McGill: *Introduction to Modern Information Retrieval.* Computer Science Series. McGraw Hill, (1983)
21. Schmid, H.: *Improvements In Part-of-Speech Tagging With an Application To German.* ACL SIGDAT Workshop, (1995)
22. Vivaldi, J. and Rodríguez, H. *Evaluation of terms and term extraction systems: A practical approach.* Terminology 13:2, 225248, (2007)
23. Weller, M., Blancafort, H., Gojun, A. and Heid, U.: *Terminology extraction and term variation patterns: a study of French and German data.* GSCL 2011, Hamburg, Germany,(2011).