2014

A survey of automatic term extraction for Brazilian Portuguese

Journal of the Brazilian Computer Society, Dordrecht, v.20, p.1-28, 2014
http://www.producao.usp.br/handle/BDPI/45447

Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

A survey of automatic term extraction for Brazilian Portuguese

Merley da Silva Conrado[1,2*], Ariani Di Felippo[3], Thiago Alexandre Salgueiro Pardo[2] and Solange Oliveira Rezende[1]

Abstract

Background: Term extraction is highly relevant as it is the basis for several tasks, such as the building of dictionaries, taxonomies, and ontologies, as well as the translation and organization of text data.

Methods and Results: In this paper, we present a survey of the state of the art in automatic term extraction (ATE) for the Brazilian Portuguese language. In this sense, the main contributions and projects related to such task have been classified according to the knowledge they use: statistical, linguistic, and hybrid (statistical and linguistic). We also present a study/review of the corpora used in the term extraction in Brazilian Portuguese, as well as a geographic mapping of Brazil regarding such contributions, projects, and corpora, considering their origins.

Conclusions: In spite of the importance of the ATE, there are still several gaps to be filled, for instance, the lack of consensus regarding the formal definition of meaning of 'term'. Such gaps are larger for the Brazilian Portuguese when compared to other languages, such as English, Spanish, and French. Examples of gaps for Brazilian Portuguese include the lack of a baseline ATE system, as well as the use of more sophisticated linguistic information, such as the WordNet and Wikipedia knowledge bases. Nevertheless, there is an increase in the number of contributions related to ATE and an interesting tendency to use contrasting corpora and domain stoplists, even though most contributions only use frequency, noun phrases, and morphosyntactic patterns.

Keywords: Automatic term extraction; Statistical, Linguist, Hybrid knowledge

Background

Domain knowledge of specific areas, such as computer science, medicine, and law, is expressed with lexical units of specialized meaning, which are denominated as *terms* or *terminological units*. In this regard, terms are considered lexical units used to designate concepts in a thematically restrict scenario [1].

The identification and selection of terminological units in specialized texts are fundamental tasks for the building of two different resources: (i) traditional lexicographical resources (such as glossaries and dictionaries) and (ii) computational resources that aid the automatic processing of these texts (such as taxonomies and ontologies). These terminological units also are important for building computational tasks (such as information retrieval and text classification).

From the 1990s on, with the advance of the investigations in the areas of corpus linguistic and natural language processing (NLP), systems for automatic term extraction have been developed in countries with large tradition in terminological research. These systems aim to fasten the manual processes of identification and collection of potential terms [1].

The first system of this nature was TERMINO, developed for the French language [2]. In Brazil, the beginning of investigations on automatic extraction of (candidate) terms occurred at the end of the 1990s [3].

Although ATE has been researched for more than 20 years, there is still room for improvement since term extraction is a difficult task even when it is carried out

*Correspondence: merleyc@gmail.com
[1]Laboratory of Computational Intelligence (LABIC), Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), P.O. Box 668, 13561-970 São Carlos, SP, Brazil
[2]Interinstitutional Center for Research and Development in Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), P.O. Box 668, 13561-970 São Carlos, SP, Brazil
Full list of author information is available at the end of the article

manually by a terminologist. This is due to the characteristics of the terminological units, which will be discussed in Section 'Problems of the automatic term extraction'. As a result, the units identified by terminologists in specialized discourses are denominated 'candidate terms', and they usually undergo a manual step of validation, carried out by at least one domain expert. Only after this process may the validated candidates be called 'terms'.

Estopá and Souto [4] demonstrate that even experts in a domain do not completely agree among themselves with respect to the identification and selection of terms. On average, agreement reaches about 60%, revealing that the recognition of terminological units is a subjective task. In view of the difficulty faced even by humans on the delimitation of candidate terms, the results obtained by the automatic extractors may not be totally precise, and thus, they also undergo a validation process similar to the one used in the manual method.

There are many ATE investigations available in the literature [5-17]. However, they perform ATE using different scenarios (e.g., variation of the test corpora[1] and measures and evaluation conditions), which make it difficult to choose the best ATE system. The comparison of methods is an existing gap in itself in the area.

In spite of this difficulty in finding the average precision of the automatic extractors due to the variation, for instance, of the test corpora size and other factors, it is verified in contributions such as Cabré et al. [1] and Ioannis et al. [18] that about 80% of the linguistic units extracted by the systems of ATE are effectively validated as terminological units by domain experts. For the English language, one of the most recent contributions on ATE is the work of Nazar [13], which confirms the tendency in the literature by presenting a 75% precision. Concerning ATE based on texts in Brazilian Portuguese, the best results also reach an 80% precision [9], when using the hybrid extractor EχATO*LP* [19]. The hybrid extractor YATE [15], proposed for the extraction of candidates from the Spanish corpora on the medicine domain, is one of the few in the literature that reaches a 98% precision rate. This extractor is characterized by the use of varied linguistic knowledge, such as morphological, syntactic, and semantic, together with statistical measures, considerably improving the extraction results. The acquisition of semantic knowledge, in special, comes from the query of external resources, such as specialized dictionaries (in digital format) and lexical databases in general language that store certain terminological knowledge.

ATE systems must also be validated considering the recall rates obtained by these systems, i.e., whether they are able to extract all of the (or most of the) terminological units foreseen by the domain experts. Usually, when the systems obtain high precision, the recall is low and *vice versa*. For example, a corpus of a specific domain has 50 terms and an ATE system correctly extracted two of these terms. Considering this result, the precision rate obtained by this system was 100%, but the recall rate is 4%.

In spite of the limitations, the development of automatic extractors has allowed the tasks of term identification and extraction, previously performed manually from printed texts, to be automatized, which has considerably streamlined and, in certain cases, refined the systematization of 'terminologies'.

In this paper, we present the scenery of the automatic term extraction for the Brazilian Portuguese language. With this survey, those who are interested in the systematization of terminologies with the help of automatic term extraction may find a detailed description of the main paradigms of extraction (statistical, linguistic, and hybrid), types of knowledge (statistical and linguistic), and linguistic/computational resources/tools (for instance corpora, taggers, parsers, etc.) that are necessary for the usage of different paradigms as well as mapping of the main contributions and proposals related to the Brazilian Portuguese language. We have focused on ATE for Brazilian Portuguese; however, this survey is also relevant for other languages since many of the investigations described here may be applied for other languages. Additionally, we have focused on general ATE because, this way, anyone may adapt the extraction for a specific task (e.g., building an ontology) or domain (e.g., extraction of medical terms by using prefixial morphemes, such as '*artri/o-*' in '***artri****te*' (in English[2], '*arthr(o)-*' in '**arthr**itis').

There are some investigations that perform a review about the ATE task [18,20-22]. Although, they are not directed for the ATE in Brazilian Portuguese, which is our focus. We have focused in Brazilian Portuguese because there is a gap relation to the resources used for ATE in Brazilian Portuguese and other languages, such as English, Spanish, and French. In such languages, the term extraction task uses advanced resources, such as WordNet, specific domain ontologies, thesaurus, and different parsers and disambiguation resources. In Brazilian Portuguese, we do not have all of these resources available and it negatively affects the results. Therefore, in this paper, we describe the available resources for Brazilian Portuguese, which are at the Ontology Portal (OntoLP) and at Linguateca, both detailed in subsection 'Projects related to the Brazilian Portuguese term extraction'.

There are five main issues observed in this survey of the state of the art, which are (i) the diversity in which extracted terms are applied direct and indirectly in different tasks, (ii) the overall observation on how the extracted terms relate to the knowledge used for extraction, (iii) the advances obtained in the term extraction task, (iv) the tendency in the recent contributions in considering knowledge from domain or contrasting domains, and (v)

the increase of a practical comparison of the existing contributions for the extraction focused on the Brazilian Portuguese language.

We present this paper as follows: Section 'Problems of the automatic term extraction' introduces examples of terms and the problems faced when extracting them. Section 'Approaches for the extraction of candidate terms' details the approaches used for extraction of candidate terms. Section 'Evaluation measures for term extraction' describes the different ways to evaluate extracted candidates. The corpora available in the Brazilian Portuguese Language are described in Section 'Corpora for Portuguese'. Section 'State of the art of term extraction in Brazilian Portuguese' presents the state of art of ATE in Brazilian Portuguese, which includes the research developed for the ATE, the measures, tools, and resources used in ATE, and the projects related to ATE. Section 'Discussion about the state of the art in term extraction for the Brazilian Portuguese language' presents the discussion about the state of the art in ATE of the Brazilian Portuguese Language. Finally, Section 'Conclusions' gives the final considerations.

Methods and Results

Problems of the automatic term extraction

In spite of the importance of the ATE task, there is not a consensus what is a term. Most of the terms are nominal units since they designate concepts (e.g., to denominate/give a name to some concept). Because of that, the nouns are more studied in specific domains [23]. We sustain such statement using Sager's example [24]: 'concepts represented in terminological dictionaries are predominantly expressed by the linguistic form of nouns'. Taking into account that terms are the entries of such dictionaries, we may consider that nouns are usually used as terms. Another example is the work of Batista [25], which affirms that terms from the business domain in Brazilian Portuguese are commonly nouns.

Term extraction is not a trivial task, even when they are carried out manually by a terminologist [4]. This difficulty is due to the characteristics of the terminological units. To illustrate how terms may be extracted emphasizing its difficulty, we consider some examples of terms of the distance education (DE[3]), nanoscience and nanotechnology (N&N[4]), and ceramic coating (RC[5]) domains that are showed in (a), (b), and (c), respectively.

(a) 'One of the main points that we consider highly relevant for the configuration of [[**virtual environments**] **for learning**] is the simple and easy design. (...) The [**cyberspace**] is much more than a [**means of communication**] or media. (...) creating and producing [**printed didactics material**]] for DL is a necessary alternative'.

(b) 'A [**nanometer**] is equal to a billionth part of a meter and any measure in such scale is invisible with the naked eye'.

(c) 'The [**frits**] are used nowadays as the main constituents of the [**enamel**] employed in the national fabrication of the [**ceramic coatings**].'

One of the main characteristics of the highlighted units is their specification/meaning. In the ceramic coating domain (sub-domain of materials engineering), for instance, the term '*frita*' ('frit') and '*esmalte*' ('enamel') have a very specific meaning in the technical discourse. In this domain, the term '*fritas*' means 'grounded glass obtained from the fusion of a mix of different ingredients, such as borates, potassium, soda, chalk, alumina, etc.'[6], and '*esmalte*' means 'coating with an impermeable, white, coloured, transparent, or opaque, aspect similar to glass, which is applied to a ceramic plate for decoration and/or protection'[7].

Regarding the morphological structure, it is possible to see in (a), (b), and (c) that terms may be (i) lexically simple, i.e., formed by one singular element or (ii) lexically complex, i.e., formed by more than one element, such as the examples presented in Table 1.

The lexically simple terms might present some morphological characteristics that differentiate them from the lexical units, which are used in the general language or in a specialized domain. In the term '***ciber****espaço*' ('cyberspace') (a), for instance, it is possible to identify '*cyber-*', abbreviation of '*Cibernética*' ('Cybernetics'), used in the denomination of several concepts of DE, which are related to the Internet, i.e., virtual world or space [26]. The term '*nanômetro*' ('nanometer') (b) is another paradigmatic example of a lexically simple term whose morphological structure reveals its terminological statute. Such statute is due to the presence of the prefix '*nano-*'[8], employed to indicate the scale 10^{-9} of the indicated measure (meter) [27]. In the matter of the medical terms, they are characterized by the presence of prefixial morphemes (for example: '*artri/o-*' and sufixal morphemes (for example: '*-patia*', in English 'pathy' with Greek origin (or Latin), such as in '***artri****te*' ('arthritis') and '*cardio****patia***' ('cardiopathy'), respectively.

The denomination of concepts, however, is not always done using lexical units that present some morphological mark that characterizes the domain to which it belongs. The use of words with specialized meaning that do not manifest morphological particularities is often used. In this category, it is included, for instance, the terms '*frita*' ('frit') and '*esmalte*' ('enamel') (c). Although they do not have formal elements of specificity (prefixes, suffixes, etc.), such units have, from a conceptual point of view, highly precise specialized meanings, as highlighted in the beginning of this text.

Table 1 Examples of simple and complex terms of different domains

	Brazilian Portuguese terms	English translations	Domains
Simple terms	*frita*	frit	Ceramic coating
	esmalte	enamel	Ceramic coating
	ciberespaço	cyberspace	Distance education
	interatividade	interactivity	Distance education
	nanômetro	nanometer	Nanoscience and nanotechnology
	ácido	acid	Nanoscience and nanotechnology
Complex terms	*revestimento cerâmico*	ceramic coating	Ceramic coating
	resistência mecânica	mechanical resistance	Ceramic coating
	meio de comunicação	communication means	Distance education
	ambiente virtual (de aprendizagem)	virtual environment (for learning)	Distance education
	escala nanométrica	nanometric scale	Nanoscience and nanotechnology
	potência óptica	optical power	Nanoscience and nanotechnology

In the general language, such units are usually generic and polysemic ('*esmalte*' ('enamel'), for instance, is defined as 'an opaque or semi-transparent glossy substance that is a type of glass, applied by vitrification to metallic or other hard surfaces for ornament or as a protective coating'[9] and it has three meanings, on average, in Portuguese dictionaries). On the other hand, in the specific domains, the same units present specific meaning and are usually not polysemic.

The terms that constitute complex lexical terms are more frequently used in the denomination of concepts in specialized domains. Such terminological units are formed by different formal structures, denominated morphosyntactic patterns (POS).

Concerning the structures based on terms from the Portuguese language, the most frequent structure is [noun + adjective]. This is the case of the terms '*revestimento cerâmico*' ('ceramic coating') in (c), '*ambiente virtual*' ('virtual environment') in (a) and '*material didático*' ('didactics material') in (a). Other morphosyntactic patterns that usually characterize the complex terms are (i) [noun + adjective + preposition + noun] (for example: '*ambientes virtuais de aprendizagem*' ('virtual environments for learning') in (a)), (ii) [noun + preposition + noun] (for example: '*meios de comunicação*' ('means of communication') in (a)), and (iii) [noun + adjective + adjective] (for example: '*material didático impresso*' ('printed didactics material') in (a)), etc.

Another characteristic of lexically complex terms is the expansion of the lexical character that, in fact, coincides with another denomination, corresponding to a specialization of the generic term. For example, the term '*ambiente virtual*' ('virtual environment') is cited, which applies to the computational systems that have an advanced interface to the users and which is being related to the *learning* systems, which are used to mediate the distance learning process. This expansion generated the term '*ambientes virtuais de aprendizagem*' ('virtual environments for learning'). Examples of expansion to the right of a generic term are very common in the specialized languages and denominate a new concept corresponding to a new invention or technology. In such way, the term '*material didático*' ('didactics material'), which denotes the set of objects that are indispensable for the execution of teaching activities, has been used with the expansion '*impresso*' ('printed'). So, this term would be '*material didático impresso*' ('printed didactics material'). Note that these examples follow the Portuguese pattern. In case of the English language, the expansion would be to the 'left' of a generic term (from 'didactics material' to 'printed didactics material').

Regarding simple lexical terms, the difficulties are mainly related to the identification of candidates with no morphological marks of specificity that indicate the terminological potential, i.e., candidates that are also used in the general language by a non-expert. For example, the Portuguese unit '*esmalte*' may mean, in English, 'nail polish' for a non-expert or 'enamel' for an expert of the ceramic coating domain.

In such cases, terminologists may use the frequency criterion, i.e., the fact of repeatedly finding the unit 'enamel' in texts of the ceramic coating domain might indicate that it is a candidate. This wise, the linguistic expression is finally selected by the terminologist and sent for the appreciation of an expert in the domain. However, it must be assumed that the use of frequency does not obtain a completely satisfactory result, because there are some candidates that have a high frequency but are not terms of the domain. For example, the word '*se refere*' ('refer(s)'), '*definido como*' ('defined as'), '*nós*' ('we'), and '*aquele*' ('that').

Experts usually receive candidates organized in a conceptual structure that reflects the reality of the domain in question. In that structure, candidates are allocated in their respective notional fields in a manner that the expert evaluates the relevance of a candidate/concept with respect to the relations that it establishes with the remaining candidates of the same field. For instance, in the conceptual structure of the ceramic coating field, the candidate '*esmalte*' ('enamel') composes the notional field of the inputs and its terminological relevance will be evaluated in regard to the remaining candidates of the field (or sub-field). By adopting such procedure, we say that experts adopt the semantic criterion for the validation of a candidate.

Concerning complex lexical terms, the difficulty lies on distinguishing the candidates of this type from the free phrases. While a complex lexical term is a combination of elements, constituting a lexical-semantic unit and, thus, expressing a specific concept (for instance: '*material didático impresso*' - 'printed didactics material'); a free phrase presents, in turn, a fragile stability in the lexical system (for instance: '*intercâmbio didático dirigido*' - 'directed didactic exchange').

Usually, the criteria proposed for the identification and delimitation of the complex terms are based on the degree of lexicalization that, in turn, determines the limits of the syntagmatic units. Terminologists may identify complex candidate terms based on some of their characteristics. According to Barros (2004, p. 103), these characteristics are the following:

> (a) Non-autonomy of a component in relation to the other components that compose the lexical-semantic unit with no meaning modification; for instance: '*quinta*' and '*feira*' in '*quinta-feira*' ('Thursday');
> (b) The impossibility of commutation of a component with no meaning changing; for instance: '*mesa-redonda*' (round table, which means a discussion involving several participants) / '*mesa quadrada*' ('square table');
> (c) No separability of components; for instance: '*terra fina*' ('thin soil') / '*esta terra é fina*' ('this soil is fine');
> (d) Internal structure particularity; for example: the absence of determination means the integration of the constitutive elements: '*ter medo*' ('to be afraid'), '*fazer justiça*' ('to make justice').

In addition to these characteristics, other criteria may be applied, such as the synonymical commutation. According to this criterion, the commutation possibility of '*estrada de ferro*' for '*ferrovia*' indicates that '*estrada de ferro*' is a potential terminological unit. We highlight that, according to the Oxford dictionary[9], '*estrada de ferro*' and '*ferrovia*' means only one word in English that is 'railroad'. Another important criterion for the verification of the degree of lexicalization of a phrase is the frequency of co-occurrences, i.e., the fact of always recovering the same association of words in the study domain is normally a clue of phrase lexicalization. Finally, the identified candidates based on at least one of the mentioned criteria are selected and sent to the expert together with the simple candidate terms.

Altogether, the cases in which the automatic extraction is problematic are mainly due to the computational limitation of handling terms whose criteria of identification do not reside in formal aspects of the language, but on abstract aspects, such as, the semantic and even enunciative aspects. These aspects, which do not present formal elements that identify them in regard to the general language, are also units with specialized meaning. For example, in the medical domain, the term '*intervenção*' ('intervention') does not mean 'action or result of intervening', but 'medical procedure'. However, the identification of these units as terms requires the machine to deal with abstract linguistic knowledge, and such task is quite complex from a computational point of view.

After detailing some ways for identifying terms, showing some examples of the difficulty in performing that, we describe and discuss in the next section the existing approaches for the extraction of candidate terms.

Approaches for the extraction of candidate terms

According to Cabré et al. and Pazienza et al. [1,22], the automatic term extraction is traditionally based on one of three approaches: statistical, linguistic, or hybrid. Such approaches are characterized by the primordial type of knowledge used in the respective task. Next, we present each of these approaches.

The statistical approach

The purely statistical approach uses knowledge obtained by application of statistical measures. For this purpose, the corpus undergoes a pre-processing step, which usually involves the identification of tokens[10], removal of stopwords[11], and the representation of the texts in tables. In these tables, each row represents a document (d_i) and each column represents an n-gram[12] of document (n_j), where cell $d_i n_j$ may be filled with some measure, for instance, the absolute frequency of n-gram n_j in document d_i. Such text representation is denominated bag-of-words (BOW). In this sense, the use of statistical measures by means of a BOW ignores any structural information about the sentences of the texts, such as the order in which the n-grams occur. From the values obtained by the chosen measure, the candidate terms are ranked. In this rank, it is considered that the candidates with higher ranking have higher probability of being terms of the domain [22].

The measures usually adopted for the development of automatic extractors according to the statistical

approach are independent from the language. The language independence is an advantageous characteristic from the computational point of view, as the use of measures do not require the specification (manual or automatic) of any type of knowledge (for example: morphological, syntactic, etc.) on the language of the texts under processing, which makes the automatic extraction simpler and faster. Compared to the human extraction, the independence of language does not reflect the process used by domain experts, as they use linguistic knowledge to identify terms. A type of linguistic knowledge is the morphological, used, for instance, to identify terms composed by Greek-Latin morphemes (for example: '*artr/i* → ***artrit****e*' ('arthritis') and '*osteo****artrit****e*' ('osteoarthritis')).

The main issue with ATE systems developed according to the statistical approach is the 'silence', i.e., the non-identification and extraction of real terms in a text or collection of texts. An example of this problem is when the chosen measure is the frequency of each term in the collection of texts that is the basis for the term extraction and, thus, a determined term (for example, *'polinização'* ('pollination'), from the ecology domain) is not extracted as it has low frequency.

According to Kageura and Umino [21], the goal of statistical measures is to identify two terminological properties: 'unithood' and 'termhood'. The measures that express unithood reveal the force or stability of the complex expressions (i.e., formed by two or more elements separated by blank spaces). The measures that express termhood reveal, in turn, the degree or relation between a linguistic expression and a knowledge domain. In other words, termhood expresses how much a linguistic expression (whether it is a simple one, as *'polaridade'* ('polarity'), or complex, as *'molécula orgânica'* ('organic molecule') and *'molécula de água'* ('molecule of water')) is related to a domain.

The main statistical measures are described next. For this, D corresponds to the number of documents in a corpus and t is a candidate term. We mention 'candidate term' instead of only 'term', since a candidate only may be considered a term after being validated by at least one domain expert. Such measures return a list of candidates ordered by the obtained values for each measure. Thus, it is necessary to manually choose a minimum value for the candidates to be considered as possible terms. It is important to stress that some measures quantify the relevance of the candidate terms of a specific domain based on corpora of other domains. The corpora of other domains are known as contrasting corpora.

Measures expressing unithood Association measures are used to express the unithood property, since unithood must reveal the power or stability of complex expressions [22]. The main statistical measures are formally described next.

(a) Log ilkelihood ratio (ll)
The log likelihood ratio test aims to detect whether the combinations are more than simple casual occurrences in the documents, providing, for such, a list of all the candidate combinations. According to Manning and Schütze [28], it is necessary to formulate two hypotheses, shown next, for the elaboration of this list (for the case of bigrams k_1k_2, for example). Consider h = hypothesis, P = probability, and t = candidate term, which belongs to the combination (gram = token 1 (k_1) and token 2 (k_2)).

$$\begin{aligned} h_1 &: P(k_1|k_2) = P(k_1|\neg k_2) \\ h_2 &: P(k_1|k_2) \neq P(k_1|\neg k_2) \end{aligned} \tag{1}$$

Hypothesis 1 (h_1) is the formalization of the independence, i.e., the occurrence of k_2 is independent from the occurrence of k_1. Hypothesis 2 (h_2) is the formalization of dependency. When h_2 is satisfied, it means that an interesting combination might be found.

(b) Pointwise mutual information (mi)
Pointwise mutual information [29] measures the quantity of information that a variable contains about another one. The formal definition of mi is

$$\text{mi}_{(k_i,k_j)} = \log_2 \frac{P(k_i, k_j)}{P(k_i) \times P(k_j)} \tag{2}$$

where k_i and k_j are tokens that compose a candidate term from a corpus with W words, $P(k_i)$ and $P(k_j)$ are the probabilities of k_i and k_j, respectively, and correspond to the frequencies of these tokens in the same corpus, while $P(k_i, k_j)$ is the probability that tokens k_i and k_j occur altogether.

(c) Dice's coefficient (dice)
Dice's coefficient [30] presents a similar interpretation to the mi. As Teline [31] explains, the difference between these measures is that, contrary to the mi, Dice's coefficient does not depend on the size of the sample (the corpus), as shown in Equation 3.

$$\text{dice}_{(k_i,k_j)} = \frac{2 \times f_{k_i,k_j}}{f_{k_i} + f_{k_j}} \tag{3}$$

where k_i and k_j are tokens of a corpus of size W, f_{k_i} and f_{k_j} are the frequencies of k_i and k_j in the corpus, respectively, and f_{k_i,k_j} is the frequency in which tokens k_i and k_j occur altogether.

Measures that express termhood For the identification of the property denominated termhood, the following statistical measures are usually used:

(a) Term frequency (tf)
Known as term frequency, this measure considers the absolute frequency of a given candidate in a corpus. Equation 4 formally defines this measure.

$$\mathrm{tf}_{t_j} = \sum_{x=1}^{D} f_{d_x,t_j} \tag{4}$$

where f_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document).

(b) Relative frequency (rf)
of a candidate in a corpus and the total frequency of all words in the same corpus, according to Equation 5.

$$\mathrm{rf}_{t_j} = \frac{\mathrm{tf}_{t_j}}{W} \tag{5}$$

where tf_{t_j} is the absolute frequency of t_j (jth candidate) and W is number of words in the same corpus.

(c) Document frequency (df)
Document frequency considers the number of documents where a term appears, according to Equation 6.

$$\mathrm{df}_{t_j} = \sum_{x=1}^{D} (1 | f_{d_x,t_j} \neq 0) \tag{6}$$

where f_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document).

(d) Average term frequency (atf)
Average term frequency corresponds to the ratio of the candidate frequency in a corpus and the document frequency of this same candidate, according to Equation 7.

$$\mathrm{atf}_{t_j} = \frac{\mathrm{tf}_{t_j}}{\mathrm{df}_{t_j}} \tag{7}$$

where tf_{t_j} is the absolute frequency of t_j (jth candidate) and df_{t_j} is document frequency of the candidate in the same corpus.

(e) Residual inverse document frequency (ridf)
Residual inverse document frequency [32] corresponds to the difference between the logs of actual inverse document frequency and inverse document frequency, according to Equation 8.

$$\mathrm{ridf}_{t_j} = \mathrm{idf}_{t_j} - \log_2 \left(\frac{1}{1 - p(0; \lambda_j)} \right) \tag{8}$$

where idf_{t_j} corresponds to $\log_2 \left(\frac{D}{\mathrm{df}_{t_j}} \right)$ and p is the Poisson distribution with parameter $\lambda_j = \frac{\mathrm{cf}_j}{D}$, the averge number of occurences of t_j per document. $1 - p(0; \lambda_j)$ is the Poisson probability of a document with at least one occurence.

(f) Term frequency - inverse document frequency (tf-idf)
Term frequency - inverse document frequency [33] considers the frequency of a candidate term (tf) according to its distribution in the collection of documents, attributing lower weight to those candidates that appear in many documents (idf), as shown in Equation 9.

$$\text{tf-idf}_{t_j} = \underbrace{\mathrm{tf}_{d_x,t_j}}_{\text{tf part}} \times \underbrace{\log \left(\frac{D}{\mathrm{df}_{t_j}} \right)}_{\text{idf part}} \tag{9}$$

where tf_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document) and df_{t_j} is the document frequency of the jth candidate.
There are some investigations that use this definition of $\mathrm{tf} - \mathrm{idf}$ [5-7] and others use different definitions [9,34]. Among the definitions available in the literature, we highlight the definition of Witten et al. [35] since it avoids that the $\mathrm{tf} - \mathrm{idf}$ value drops to 0 if a candidate occurs in all documents of a corpus, as observed in Equation 10.

$$\mathrm{tf} - \mathrm{idf}_{d_x,t_j} = \underbrace{(1 + \log(\mathrm{tf}_{d_x,t_j}))}_{\text{tf part}} \times \underbrace{log \left(1 + \frac{D}{\mathrm{df}_{t_j}} \right)}_{\text{idf part}} \tag{10}$$

where tf_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document) and df_{t_j} is the document frequency of the jth candidate.

(g) Term contribution (tc)
Known as Term contribution [36], this measure considers that the importance of a term corresponds to the contribution of this term to the similarity of the documents. In this regard, the contribution of the term provides higher ranking to those terms that appear in few documents, not considering very rare or very frequent terms in the collection, as shown in Equation 11.

$$\mathrm{tc}_{t_j} = \sum_{x=1}^{D} \sum_{y=1}^{D} f_{d_x,t_j} \times \mathrm{idf}_{t_j} \times f_{d_x,t_j} \times \mathrm{idf}_{t_j} \tag{11}$$

In this equation, there are xth document and yth document in a corpus, where f_{d_x,t_j} is the frequency of the t_j (jth candidate) in the d_x (xth document) and idf_{t_j} is the inverse of the frequency of the documents of the jth candidate.

(h) Term variance (tv)
Known as term variance [37], this measure considers that important terms are those that do no appear with low frequency in the documents and keep a non-uniform distribution in the collection (higher variance). For this, the variance of all terms of the collection is calculated (Equation 12).

$$\text{tv}_{t_j} = \sum_{x=1}^{D} [f_{d_x,t_j} - \bar{f_{t_j}}]^2 \tag{12}$$

where f_{d_x,t_j} is the absolute frequency of t_j (jth candidate) in the d_x (xth document) and $\bar{f_{t_j}}$ is the average of the frequencies of the jth candidate in the documents of the corpus.

(i) Term variance quality (tvq)
Known as term variance quality [37], this measure is an adaptation of the tv measure, but it aims at qualifying the variance of the words. This measure considers that the words with little variation presents little discriminant power, as they occur in a uniform way in the whole collection. Equation 13 formally describes TVQ.

$$\text{tvq}_{t_j} = \sum_{x=1}^{D} f_{d_x,t_j}^2 - \frac{1}{D}\left[\sum_{x=1}^{D} f_{d_x,t_j}\right]^2 \tag{13}$$

where f_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document).

(j) Zone-scored term frequency (zstf)
The zstf [38] measure, formally described in Equation 14, assumes that some parts of the document (such as the abstract and the conclusion) bring higher relevant information about the contents of the document than other parts. Based on this consideration, it attributes higher weights to the words that occur in parts of the document with higher impact or in which higher information related to the content of the document is concentrated.

$$\text{zstf}_{t_j} = \sum_{x=1}^{D}\sum_{z=1}^{Z} f_{d_x,t_j} \times \text{weight}_z \tag{14}$$

where f_{d_x,t_j} is the frequency of t_j (jth candidate) in the d_x (xth document) and weight_z is the weight calculated to the zth zone of the documents of the corpus. This weight should follow the restriction showed in Equation 15.

$$\sum_{z=1}^{Z} \text{weight}_z = 1 | 0 < \text{weight}_z \leq 1 \tag{15}$$

(k) Term domain specificity (tds)
Term domain specificity [39] assumes that a relevant word for a domain is more frequent in the corpus of this domain than in other corpora. Based on this consideration, the authors highlight the relevance of the words in a domain corpus, based on the probability of occurrence of these words and considering the contrasting corpora. This measure is formally described in Equation 16.

$$\text{tds}_{t_j}^{(c)} = \frac{P(t_j{}^{(c)})}{P(t_j{}^{(g)})} = \frac{\frac{\text{tf}_{t_j}^{(c)}}{W^{(c)}}}{\frac{\text{tf}_{t_j}^{(g)}}{W^{(g)}}} = \frac{\text{prob. in domain } c}{\text{prob. in corpus } g} \tag{16}$$

Considering that $P(t_j{}^{(c)})$ corresponds to the occurrence probability of t_j candidate in corpus c, $W^{(c)}$ is the total number of words in corpus c, and g is a contrasting corpus.

(l) Termhood index (thd)
The thd measure [40] assumes that a relevant word for a domain is more frequent in the corpus of that domain than in other corpora. To verify the relevance of the word in the domain, the index ponders the ordination of the word considering the words of the corpus, as shown in Equation 17.

$$\text{thd}_{t_j}^{(c)} = \frac{r_{t_j}^{(c)}}{\left|W^{(c)}\right|} - \frac{r_{t_j}^{(g)}}{\left|W^{(g)}\right|} \tag{17}$$

where $\left|W^{(c)}\right|$ is the amount of words of c, $r_{t_j}^{(c)}$ is the ordination value of candidate t_j in corpus c and, g is a contrasting corpus.

(m) Term frequency, inverse domain frequency (TF-IDF)
Term frequency, inverse domain frequency [41], formally described in Equation 18. In this paper, the acronym of this measure is used in upper-case letters (TF-IDF) in order to avoid confusion with term frequency - inverse document frequency [33], in which acronym of the latter is used in lower-case letters (tf-idf). TF-IDF is based on the term frequency - inverse document frequency (tf-idf), but instead of considering the occurrences of the terms in individual documents, TF-IDF uses contrasting corpora, in addition to the domain corpus, and it considers the term occurrences in each individual corpus.

$$\text{TF-IDF}_{t_j}^{(c)} = \underbrace{\frac{\text{tf}_{t_j}^{(c)}}{W^{(c)}}}_{\text{TF part}} \times \underbrace{\log\left(\frac{|G^*|}{\left|G_{t_j}^*\right|}\right)}_{\text{IDF part}} \tag{18}$$

where $\text{tf}_{t_j}^{(c)}$ is the absolute frequency of the candidate t in the corpus c, G^* is the set of all contrasting corpora and corpus c, and $G_{t_j}^*$ is the subset of G^* in which candidate t_j appears at least once.

(n) Term frequency - disjoint corpora frequency (tf-dcf)

Term frequency - disjoint corpora frequency [34] penalizes a word proportionally to the number of contrasting corpora where it appears and also to the number of occurrences of this word in each of these corpora, according to Equation 19.

$$\text{tf-dcf}_{t_j}^{(c)} = \frac{\text{tf}_{t_j}^{(c)}}{\prod_{\forall g \in \mathcal{G}} 1 + \log\left(1 + \text{tf}_{t_j}^{(g)}\right)} \tag{19}$$

considering t_j as the candidate of the corpus c, g is a contrasting corpus, and $\mathcal{G}$ as the set of contrasting corpora.

(o) Weirdness
Weirdness [42] was initially proposed for document retrieval system. This measure considers that the distribution of candidates in a specific domain corpus is different from the candidate distribution in a general corpus. Equation 20 formally defines this measure.

$$\text{weirdness}_{t_j}^{(c)} = \frac{\text{tf}_{t_j}^{(c)} / W^{(c)}}{\text{tf}_{t_j}^{(g)} / W^{(g)}} \tag{20}$$

where $W^{(c)}$ is the amount of words of domain corpus (c), $\text{tf}_{t_j}^{(c)}$ is the frequency of candidate t_j in corpus c, and g is a contrasting corpus.

The linguistic approach

According to this approach, the candidate terms are identified and extracted from a corpus based on their linguistic characteristics or properties, which may be from different types or levels.

The NLP manuals, such as the one of Jurafsky and Mantin [43], are based on a hierarchy of types of linguistic knowledge, elaborated based on a scale of abstraction and complexity, i.e., the higher the level of this scale, the more complex are the modeling and computational treatment of the knowledge (according to Figure 1). In the lowest level of this scale is the morphological knowledge, followed by the syntactic, semantic, and pragmatic-discursive knowledge.

The morphological level knowledge (i.e., referring to the internal structure of the terms) considers only the Greek-Latin morphemes or the morphemes that are typical of the domain that indicates the occurrence of a possible term. In the extraction of medicine terms and correlated areas, for instance, it is common to identify morphemes, whether radical or affix morphemes, with a Greek or Latin origin, such as stated by Vivaldi and Rodriguez [15] in '*artri/o-*' ('arthr(o)-'), from the Greek '*arthros*', in '*artrite*' ('arthritis'). In ATE, it is possible to identify candidate terms from the domains of the nanoscience and nanotechnology based on the identification of morphemes such as *nano-*, as this term composes several simple terms (for example: '*nanotubo*' - 'nanotube') and complex terms (for example: '*nanotubo de carbono*' - 'carbon nanotube') from this domain [27].

The extraction based on the syntactic knowledge (i.e., in relation to the order and function of the terms in the sentences) normally identifies the syntagmatic structure of the sentences, from which the noun phrases are selected as candidate terms. For example, based on this criterion, '*moléculas*' ('molecules') is identified as a candidate term since it is the head of a noun phrase in sentence '*Moléculas de ácido silícico condensam com formação de água*' ('Silicic acid molecules condensed with the formation of water').

For the candidate term extraction based on the semantic level knowledge (i.e., in relation to the subjacent meaning or concept of the terms), the extractor of the linguistic approach identifies the semantic type (for example: '*mundo*' <concreto.lugar>, in English, 'world' <concrete>), of the candidates or the concept subjacent to them. In the general literature, there are few proposals that use a more abstract knowledge, such as the semantic level (for example: [44]).

Regarding the pragmatical level knowledge, to the best of our knowledge, there are no contributions that identifies candidate terms based on properties related to the use.

In general, the term extraction according to the linguistic approach is frequently based on the morphosyntactic level knowledge [22]. In this case, the goal is to perform ATE by means of (i) the syntactic category of the *n*-grams (for example: verb, noun, adjective, etc.) of the corpus and/or (ii) morphosyntactic patterns (for example: N + Adj and N + PREP + N). Regarding the categories, several contributions base the ATE on the identification of candidates from the category of the names, as the terminologies are composed, mostly, by terms of such category. The morphosyntactic patterns, in turn, are frequently used because the terminologies tend to concentrate a large volume of noun terms that present internal structures as the

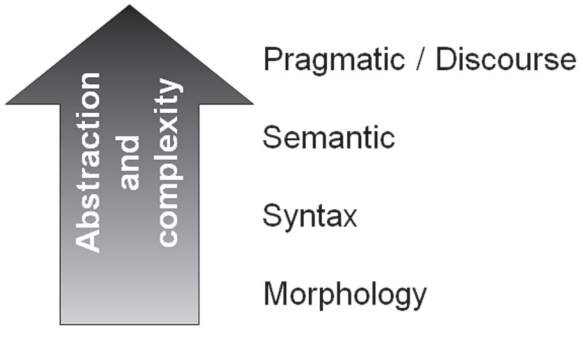

Figure 1 Linguistic levels of knowledge of language. Adapted from Jurafsky and Martin [43].

terms illustrated by patterns N + adj (for example: *'material nanoestruturado* ('nanostructured material')) and N + PREP + N (for example: '*nanotubo de carbono*' ('carbon nanotube')).

Considering the usage of the linguistic knowledge, the pre-processing of corpora for ATE involves other processes, in addition to the sentential delimitation, tokenization, and removal of stopwords. Based on the identification of the syntactic categories and of the morphosyntactic patterns, the systems may perform different processes during the pre-processing stage, such as part-of-speech tagging, which consists of the association of a tag that indicates its syntactic category to each word of the corpus (for example: '*nanotubo_N de_PREP carbono_N*' ('carbon nanotube')) [45] and normalization of text words, which consists of unifying them by reducing their variations. The normalization may be performed by using the techniques of (i) lemmatization, which consists on the reduction of each word of a text to its lemma or canonical form, i.e., non-marked forms, with no flexions [46]; during lemmatization, verbs are reduced to the infinitive (for example: 'casamos → casar' ('to get married')) and nouns and adjectives to the masculine singular (for example: '*latas → lata*' ('can') / '*feias → feio*' ('ugly')); (ii) stemming, which consists on the reduction of the words of a text to their radical [46] (for example: '*casamos → cas*' / '*latas → lat*'); and (iii) nominalization, in which words begin to present a syntactic/semantic behaviour similar to that of a noun[13] [47] (for example: '*casamos → casa*' / '*latas → lata*'). When based on the identification of phrases, the systems usually perform the recognition of the syntactic structure, parsing, of the sentences by attributing syntactic functions to the recognized constituents [48], for example, subject and predicate, noun and verb phrases.

This way, the linguistic extractor may use processing tools of natural language as sentence splitters, tokenizers, taggers, lemmatizers, stemmers, 'nominalizators', and parsers.

Independently of the type of knowledge adopted, the results obtained by such approaches are, in general, better than the results obtained by the statistical approach. However, the linguistic approach is not free of problems either. In this case, the extraction is language-dependent, as the identification of the candidates requires the specification of some type of linguistic knowledge (for example: the syntactic category of the words) that is obtained by tools, such as taggers and parsers, lemmatizers, etc, which generate frequent errors that affect the tasks of identification and extraction of candidate terms. When performed manually, the necessary linguistic specification makes the candidate extraction more costly and slower.

In general, the main issues of the systems developed according to the linguistic approach are related to the 'silence', previously mentioned, and to the 'noise', i.e., to the identification/extraction of a large number of candidates, which are discarded during the evaluation phase. As an example of noise, if we consider that nouns may be terms while adjectives cannot, if an adjective (for example, '*ecológico*' ('ecological') from the ecology domain) is wrongly marked as a noun, it would mistakenly be extracted as a domain term.

The hybrid approach

The hybrid approach considers statistical and linguistic properties for the identification and extraction of candidates.

In this approach, the order of usage of the knowledge may vary. In some systems, the statistical knowledge is used before the linguistic knowledge, while in others, the statistical knowledge is used after the use of the linguistic knowledge. According to Teline [31] and Pazienza et al. [22], the best results are obtained when the statistical measures are applied on a list of previously extracted candidates based on some linguistic property, as the reliability of the statistical measures is higher when applied to linguistically 'justified' candidate terms. One of the reasons for this is that the terms usually follow pre-defined patterns for each domain (nouns, mainly). These patterns are identified during the morphosyntactical analysis of the candidate term; however, the pattern may be different depending on the context in which the candidate appears. For example, the word '*segundo*' ('second') may be a noun, such as '*Alguns **segundos** são suficientes...*' ('Some **seconds** are enough...'), or an ordinal numeral, such as '*O **segundo** ano...*' ('The **second** year...'). The statistical methods usually do not consider such context and this is one of the reasons why it is advised to first identify the linguistic properties of the candidates and, then, apply statistical methods. In addition, statistical methods are more rigid and may eliminate terms with low frequency, but which might be important for the domain.

Examples of hybrid measures[14], i.e., measures that use statistical and linguistic knowledge, include the c-value, nc - value , and the *glossEx*:

(a) c-value
For the c-value [49] measure, the linguistic resource supports the generation of a list of candidate terms[15] according to a linguistic filter based on the search for pre-determined syntactic patterns. Next, the calculation of the potential of each candidate to be a term or not is carried out, and, for that purpose, the length of each candidate is considered, in grams (whether it is a bigram, trigram, etc.), as well as its frequency in the corpus.

$$\text{c-value}_{t_j} = \begin{cases} \log_2 |t_j| \times \text{tf}(t_j), \text{if } t_j \notin \text{a}V; \\ \log_2 |t_j| \left(\text{tf}(t_j) - \frac{1}{P(T_{t_j})} \sum_{b \in T} \text{tf}(b) \right), \\ \text{otherwise.} \end{cases} \quad (21)$$

For the formal description of the c-value (Equation 21), we consider t_j as the jth candidate term (noun phrase), $|t_j|$ as the length in grams of t_j, $\mathrm{tf}(\mathrm{t_j})$ as the frequency of t_j in the corpus, T_{t_j} as the set of candidates with length in grams larger than t_j and which contains t_j, $P(T_{t_j})$ as the number of such candidates (types) including the type of t_j, $\sum \mathrm{tf}(b)$ as the total number of t_j as a sub-string of candidate b so that $|t_j| < |b|$, and V as the set of neighbours of t_j. The c-value measure was initially proposed to express the unithood property; thus, it works with complex expressions. Barrón-Cedeño and his co-workers [50] adapted this measure in order to make it possible to express the termhood and, thus, apply it to unigrams (see Equation 22).

$$\text{c-value}_{t_j} = \begin{cases} \mathrm{c} \times \log_2 |t_j| \times \mathrm{tf}(t_j), \text{if } t_j \notin \mathrm{a}V; \\ \mathrm{c} \times \log_2 |t_j| \left(\mathrm{tf}(t_j) - \frac{1}{P(T_{t_j})} \sum_{b \in T} f(b) \right), \\ \text{otherwise.} \end{cases} \tag{22}$$

where $c = i + \log_2 |t_j|$. The authors state that by using $i = 1$, it is possible to obtain experimentally better results.

(b) nc-value
The nc-value [49] measure expresses both, the unithood and the termhood. This measure assumes that the concept in which the candidates appear is meaningful to determine whether these are terms or not. In this wise, the nc-value considers that the neighbourhood of each of the candidates may favour the quality of such determination. This neighbourhood consists of the words around the candidate, called 'context words'. To identify them, it is necessary to previously define the size of the window and consider only the words that have the grammatical classes of nouns, adjectives, or verbs. In this sense, for each of these words (w), a weight is calculated weight_w (Equation 23).

$$\mathrm{weight}_w = \frac{t(w)}{\mathrm{nc}} \tag{23}$$

where $t(w)$ is the number of candidates where the word w appears and nc is the total number of candidates considered in the corpus.

In the sequence, it is possible to calculate the nc-value measure, which is formally expressed in Equation 24.

$$\text{nc-value}_{t_j} = 0.8\text{c-value}_{t_j} + 0.2 \sum_{b \epsilon C_{t_j}} f_{t_j}(b) \mathrm{weight}_b \tag{24}$$

In Equation 24, t_j is the candidate term, C_{t_j} is the set of words of the context of candidate t_j, b is a context word of candidate $_{t_j}$, $f_{t_j}(b)$ is the occurrence frequency of b as a context word of candidate $_{t_j}$, and weight_b is the calculated weight for b as a context word.

Evaluation measures for term extraction

As previously mentioned, term extraction may be the basis for several tasks. For this reason, there is not a pattern for the evaluation of extracted candidate terms. For instance, in case the terms are used in a taxonomy, the candidates have to be evaluated considering their representativeness in relation to the domain in question, the position of each candidate in the taxonomy, among others.

Vivaldi and Rodríguez [15] highlight an issue related to the evaluation of candidate terms that is difficult to be answered: 'Who determines which are the relevant terms in a given test text?'. In this wise, before evaluating the candidate terms, it is necessary to consider in which task the extracted terms will be used, i.e., if they will be used for the building of taxonomies, information retrieval, etc. Therefore, it is necessary to define the meaning of their 'quality' by analyzing the requirements involved in the final task. Thus, the evaluation must be done to verify such candidate term quality.

According to Almeida and Vale [27], the evaluation may be done in three distinct ways, as follows: the objective analysis of the list of candidate terms; the subjective analysis of experts in the domain in question; and the combination of the former evaluation strategies.

The first evaluation strategy comprehends the comparison of the list of candidate terms against a gold standard, i.e., a list of terms considered as the pattern of the domain in question. Most of the times, the gold standard is elaborated by domain experts, as they have the knowledge of the domain. This task performed by experts may be supported by some automatic processes. In this sense, the creation of this list demands time, effort, and carries their subjectivity. Additionally, there might not be a concordance among the experts whether a given candidate term is a domain term or not. This lack of concordance may be explained by the difficulty in defining a set of measurable properties that contribute to the evaluation of the quality of term extraction [15].

This form of evaluation using gold standards may be performed by means of: precision, recall, noise, silence, and accuracy. The precision measures the degree of correctness of the candidate terms. The recall measures the degree of coverage of the candidate terms. The noise is the complement of the precision. The silence is the complement of the recall.

The formal descriptions of these measures are presented in Figure 2. In this figure, A corresponds to the candidate terms that were correctly extracted as terms; C corresponds to the candidate terms that were mistakenly accepted as terms (the candidates that should not have been extracted); B corresponds to the terms that should

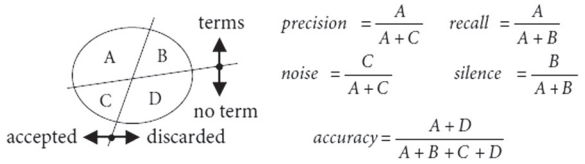

Figure 2 Term evaluation measures [15].

have been extracted, but were not; and D corresponds to the 'non-terms' which were, correctly, not extracted as terms. This way, A and B belong to the gold standard.

In addition to these measures, it is common to calculate the F-measure, which is a harmonic average between recall and precision that penalizes high divergence between the precision and recall. Its formal definition is presented in Equation 25.

$$\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{25}$$

Moreover, this first strategy of evaluation may be carried out by replacing the gold standard by a corpus of the general language. Such substitution allows the evaluation of whether the extracted candidate terms belong to the general language or not. If so, they are probably not candidate terms specific for the domain.

Additionally, it is possible to apply the *context term weight* (ctw) measure [51] to the terms. This measure evaluates the number of times (the frequency of the candidate term in the corpus) in which a candidate term occurs in a given context. In order to apply this measure to the terms, it is possible to consider the gold standard as context. Thus, for the evaluation of the candidates, they are retrieved from the list. The formal description of the ctw measure adapted for the evaluation of terms, and presented in the work of Conrado [52], is presented in Equation 26.

$$\text{ctw}(t_j) = \sum_{i \epsilon T} f_{t_j}(t_i) \tag{26}$$

where t_j is the jth candidate term; T is the set of terms of the gold standard that coincides with the extracted candidate terms; i is the ith term of T; and $f_{t_j}(t_i)$ is the frequency of t_i in the corpus as a candidate term t_j, which is obtained during the extraction of the candidate terms.

The second evaluation strategy corresponds to the submission of the candidates to be subjectively evaluated by domain experts. Counting on the support of such experts in subjective evaluations, which demand considerable manual effort, it is often a costly and quite slow process when compared to the objective evaluation, which presents on a delay for the conclusion of the process.

The last considered evaluation strategy is the sequential use of the two previously mentioned strategies, i.e., first, the list of candidate terms is compared against a gold standard (or against the general language corpus) and, in the sequence, experts analyse the obtained results.

Next, we describe the existing corpora in the Brazilian Portuguese language, focused on the term extraction task.

Corpora for Portuguese

The number of available corpora in Brazilian Portuguese has increased significantly. In general, such corpora may be obtained in the Linguateca repository, in the OntoLP portal [53], both described in subsection 'Projects related to the Brazilian Portuguese term extraction', in the Group of Studies and Research in Terminology (GETerm[16]), in the Group of Natural Language Processing (NLP Group[17]), in the Laboratory of Computational Intelligence (LABIC[18]), and in the Inter-institutional Centre for Research and Development in computational Linguistics (NILC[19]).

Table 2 describes the 17 corpora found until now in the Brazilian Portuguese language that are used in the term extraction task.

The corpus of Centre of Metalworking Information(CIMM[20]) has 3,326 texts of dissertations, news, and thesis related to the metalworking domain.

The CorpusDT[21] [55] corpus was developed at NILC at the University of São Paulo (USP[22]). This corpus has 52 texts originated from thesis and dissertations in the computer science domain. In addition to these categories, the documents are classified into eight different sub-domains, which are databases, computational intelligence, software engineering, hypermedia, digital systems, distributed systems, and concurrent programming, graphical computing and image processing, and high-performance computing. This corpus was created to serve as a base for the study carried out on the structure of scientific texts such as thesis and dissertations in the Portuguese language.

The EaD [56] corpus was developed at GETerm at the Federal University of São Carlos (UFSCar[23]). The corpus has 347 texts in the distance education domain obtained from the Internet. These texts were divided into two subcorpora, in which 307 texts are considered *-Technical* and 40 texts are considered as *+Technical*. The authors classified these sub-corpora *-Technical* in (i) Scientific divulgation, which were texts from the articles of divulgation; (ii) Informative, texts from textbooks and handouts; and (iii) Instructional, texts from news/reporters. The sub-corpora *+Technical* received the tag of Technical-scientific, as it contains texts from thesis, dissertations, research projects, and scientific papers. For the elaboration of the gold standards, only the NPs that satisfied the following conditions were considered as candidate terms: (i) the NPs that presented at least a pre-defined absolute frequency in the EaD corpus; and (ii) the NPs that were not manually excluded for being, for instance, proper

Table 2 Textual bases in the Brazilian Portuguese language

Names	Responsible	Number of texts	Number of classes	Gold standard	Domains
CIMM	CIMM	3.326	3	No	Metalworking
CorpusDT	NILC-USP	52	2	No	Computer science
Corpus.EaD	GETerm-UFSCar	347	2	Yes	Distance education
CSTNews	NILC-USP	140	50	No	General
ECO	NILC-USP	390	-	Yes	Ecology
Embrapa	CNPTIA-EMBRAPA LABIC-USP	2.149	8	No	Agribusiness
Folha-RIcol	PLN-PUCRS	5.090	18	No	General
Geology	PLN-PUCRS	234	-	No	Geology
IFM2	LABIC-USP	134	-	No	Production engineering
JPED	Coulthard [54]	283	-	Yes	Pediatrics
Data mining	PLN-PUCRS	53	-	No	Data mining
Stochastic modelling	PLN-PUCRS	88	-	No	Stochastic modelling
Muniz	NILC-USP	50	5	No	Appliances
Nanoscience and nanotechnology	GETerm-UFSCar	1.057	5	Yes	Nanoscience and nanotechnology
Parallel processing	PLN-PUCRS	62	-	No	Parallel processing
Ceramic coating	NILC-USP	164	-	Yes	Ceramic coating

names or NPs with no terminological value (for example: *século XIX* - "19th century"); and (iii) for the unigrams, the NPs that did not occur in the general language corpus. Thus, an expert decided which of these candidates are terms of the domain, by creating the gold standards, with 59 unigrams, 102 bigrams, 63 trigrams, and 5 tetragrams [57].

CSTNews [58] was developed at NILC. The corpus has 140 news texts divided into four sub-themes: Daily news, Sports, World, and Politics. This corpus was created to serve as basis for the research on multi-document summarization.

The ECO[24] [16] corpus was developed at NILC. The corpus has 390 documents from the ecology domain. Its objective is to support the creation of a knowledge base with ontological information for terms from the ecology domain. The authors built the gold standards considering the terms occurring, at the same time, in two books, two specialized glossaries, one online dictionary, all related to the ecology domain, and the ECO corpus. After the removal of the duplicated terms, the lists totalled 322 unigrams, 136 bigrams, and 62 trigrams.

The EMBRAPA corpus, originated from documents of the National Centre for Technological Research in Informatics for Agriculture (CNPTIA[25]) of the Brazilian Agricultural Research Corporation (EMBRAPA) [52], was developed at LABIC, USP, with a partnership with CNPTIDA, EMBRAPA. The corpus has 2,149 texts from the agribusiness domain referring to eight products: corn, sugarcane, beans, milk, apple, cowpea, eucalyptus, and cashew. This corpus was created for the term extraction task, as, for this task, they had the support of experts in this domain for the evaluation of the extracted terms.

The Folha-RIcol[26] corpus was developed in the NLP group of the Pontifical Catholic University of Rio Grande do Sul (PUCRS) for the evaluation and training of information retrieval systems. This corpus contains 5,090 papers from the NILC corpus, which, in turn, were extracted from the Folha de São Paulo newspaper. Folha-RIcol has 18 researched subjects, which are car accidents, drug traffic, Brazilian music, teaching, soccer games, telephone selling, electoral campaign, nature phenomenons, tropical fruits, airplane travelling, serious sicknesses, pets, salary raises, real-estate renting, international travelling, computer usage, university professors, and other subjects.

The Instituto Fábrica do Milênio (IFM2) [59] corpus was developed at LABIC, USP and has 134 papers on the production engineering domain, which may be divided into 5 classes: WP01, WP02, WP03, WP04, and WP05.

JPED [54] has 283 texts on the pediatrics domain published online in the Pediatrics Journal (Jornal de Pediatria) (JPED)[27]. It was organized to study Portuguese-English translation patterns. The gold standards of this corpus were elaborated in the TEXTQUIM / TEXTECC project, previously mentioned, and have 1,534 bigrams and 2,647 trigrams.

The corpora of Geology, Data mining, Stochastic Modelling, and Parallel Processing [60] contain 221, 94, 53, and 86 texts, respectively. These corpora are composed by thesis, dissertations, and papers of these domains. The creation of these corpora was destined to the computational processing for applications, such as building of glossaries, information retrieval, and ontology building.

The Muniz [61] corpus was developed at NILC. The corpus has 50 texts related to technical manuals from 5 categories of appliances, considered as classes, which are food centrifuges, irons, stoves, hair dryers, and televisions. The total of words of the corpus is 182,000. The creation of this corpus was destined to the term extraction focused on technical manuals.

The Nanoscience and Nanotechnology [62] corpus was developed at GETerm, UFSCar, and has 1,057 texts divided into 5 categories: Scientific, Divulgation-scientific, Informative, Technical-administrative, and Others (research companies and institute prospects, presentation slides, etc.). This corpus was built in the scope of the NANOTERM project [63], which was detailed in 3. For the elaboration of the gold standards, the authors identified candidate terms by statistical methods, after the removal of stopwords. They manually excluded some of these candidates by a linguist and, then, an expert decided which of these candidates are domain terms, creating, this way, the gold standards, that include 1,794 unigrams, 586 bigrams, 590 trigrams, and 151 tetragrams [62,64].

The Ceramic Coating corpus [31] was developed at NILC. The corpus has 164 papers from the Industrial Ceramics Magazine[28], totaling 448,352 words. Each text has on average from 4 to 8 pages (approximately 4,000 words). The gold standards have 264 unigrams, 74 bigrams, and 43 trigrams. The elaboration of this corpus has allowed the evaluation of methods for the automatic term extraction for a Master's research [31].

In this section, we described 17 corpora found until now in the Brazilian Portuguese language. Among these 17 corpora, only 4 corpora (see Table 2) contain gold standards, which means that only with these 4 corpora that objective evaluation of the ATE task is possible.

In next section, we describe the state of the art of term extraction in the Brazilian Portuguese language.

State of the art of term extraction in Brazilian Portuguese

In this section, we highlight the main contributions identified in the literature that are related to the automatic term extraction from the Brazilian Portuguese corpora. Figure 3 presents the organization of such investigations in function of the approach and the type of linguistic knowledge. We observed that most of contributions used the hybrid approach for ATE, considering linguistic properties of the candidate terms in the levels of syntax and morphology and using statistical measures. Only a few contributions considered the semantic properties and other few contributions did not use any statistical measure. In summary, there are investigations that analysed the application of several statistical measures for the ATE in corpora of Brazilian Portuguese [65,66]. Another proposal analysed the use of linguistic knowledge only (morphological, in this case) [27]. Some contributions compared the term extraction according to the statistical and linguistic approaches [67,68] and other contributions explored ATE according to the hybrid approach [5-12,14,16,17,31,34,52,53,61,69-72].

These contributions were classified according to their goals. The first group of contributions (subsection 'Research developed for Brazilian Portuguese term extraction') corresponds to investigations that primarily compared, adapted, or developed investigations for term extraction. More specifically, such contributions are described in function of the approach (linguistic, statistical, and hybrid) in which the ATE is based. Furthermore,

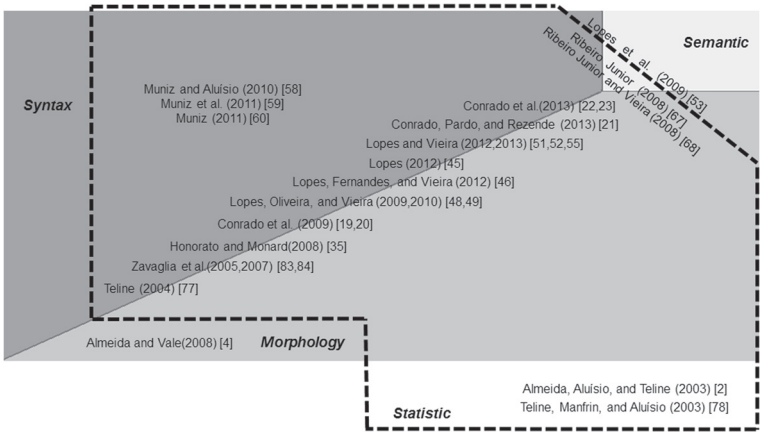

Figure 3 Organization of the contributions related to term extraction for the Brazilian Portuguese.

for those contributions that are based on linguistic knowledge, the level of knowledge employed is stated (i.e., morphological, syntactic, and/or semantic level). The second group of contributions (subsection 'Research related to Brazilian Portuguese term extraction') are investigations that aimed at discussing about term extraction in the Brazilian Portuguese language. The names of the authors of each contribution are rendered in italics in order to highlight their contributions.

Additionally, we describe the main current projects (subsection 'Projects related to the Brazilian Portuguese term extraction') we have found in the literature that are related to term extraction. Finally, Section 'Discussion about the state of the art in term extraction for the Brazilian Portuguese language' includes a general overview of the performance of the ATE work available for processing Brazilian Portuguese, a discussion regarding the state of the art on ATE, as well as a geographical mapping of all contributions directly related to the ATE of Brazilian Portuguese.

Research developed for Brazilian Portuguese term extraction

Teline et al. [66] evaluated the use of statistical measures applied to the corpus of the Industrial Ceramic Magazine, described in Section 'Corpora for Portuguese'. The evaluated measures were, for unigrams, frequency; for bigrams, frequency, mutual information, log likelihood ratio, and Dice's coefficient; and for trigrams, *tf*, *mi*, and *ll*. As results, the authors observed that it was not possible to identify which of the adopted measures are the best to be applied in bigrams of this corpus, because the results were quite similar. Regarding the case of trigrams, the absolute frequency measure presented a better result than the mutual information and log likelihood ratio measures. Based on the results presented by the authors, it was possible to calculate the following values for the F-measure: 26%, 9%, and 0.62% for unigrams, bigrams, and trigrams, respectively.

Zavaglia and her co-workers [16,17] evaluated the term extraction according to the linguistic, statistical, and hybrid approaches. For all three approaches, the authors removed stopwords and used some indicative phrases, for instance '*definido como*' ('defined as') and '*chamado*' ('called'). More specifically, for the linguistic approach, they considered morphosyntactic patterns for the term extraction. For the statistical approach, the authors compared the ATE results obtained using separately different statistical measures (absolute frequency, mutual information, log likelihood ratio, and Dice's coefficient), available in Ngram Statistics Package (NSP) [73]. Regarding the hybrid approach (which corresponded to the use of the aforementioned statistical and linguistic measures together), they combined the knowledge obtained by these adopted measures. For the experiments, the authors used the ECO [16] corpus, which contains 390 text documents in Portuguese from the ecology domain. They observed that the hybrid approach obtained the best results, although the number of extracted candidate terms was median. As results, the best values for the F-measures were for unigrams, 16.48%; for bigrams, 16.88%; for trigrams, 5.77%.

In the work of *Honorato and Monard* [8], the authors developed a framework for the extraction of terminology using the hybrid approach for the medical report. This framework, called 'Term Pattern Discover' (*TP-Discover*), in summary, selects words and phrases that occur with a certain Absolute Frequency (statistical method) and, for that purpose, the lemmatisation technique (linguistic method) is applied using the TreeTagger [74] lemmatiser. Then, the terms that follow predefined morphosyntactic patters are selected (for example: term '*terço distal*' ('distal third') follows the N+Adj pattern). As we did not have access to the exact measures of precision and recall, based in their available results, we assumed that the best F-measure value was 59%.

Ribeiro Junior and Vieira [14,53] performed term extraction using the hybrid approach and the following three stages: (i) selection of semantic sets, (ii) simple term extraction, and (iii) composed term extraction. The selection of semantic sets consists of the removal of stopwords and the use of the semantic information made available by the PALAVRAS parser [75]. These are prototypical information that classify common names in general classes, for example the tag '<*an*>' attributed to the noun '*olho*' ('eye'), indicates that the word belongs to the class '*Anatomia*' ('Anatomy'). In this wise, the nouns tagged with the same tag are grouped in semantic groups. The expert in the domain analizes the list of obtained semantic tags ordered by the relative frequency (rf) of each tag. Then, the expert excludes the semantic groups that he/she considers not to have relation with the domain in question. For the extraction of simple and composed terms, the authors used the relative frequency (rf), tf-idf [33], and nc-value [49] statistical measures. Moreover, for the extraction of the composed terms, they used the ca-value [49] statistical measure. Regarding the extraction of simple terms, they only extracted the candidates that belong to certain grammatical classes defined by the expert, as well as the head of the noun phrases. For the candidates to composed terms, instead, the authors only considered those that consisted of determined morphosyntactic patterns, as well as those that constituted noun phrases. Finally, the authors combined these linguistic and statistical methods, originating hybrid methods. All these methods were used to extract terms from two corpora in Portuguese: Nanoscience and Nanotechnology [62], which contain 1,057 texts from those domains, and JPED, which is composed by 283 text documents from the pediatric domain. For the first

corpus, the precision of the extracted terms were calculated and, for the JPED corpus, the F-measure values were also calculated, and the best results were 22.39%, 10.04%, and 5.46% for unigrams, bigrams, and trigrams, respectively.

Conrado and co-workers [52,69] performed term extraction (unigrams, bigrams, and trigrams) using the hybrid approach. The authors applied word normalization techniques (stemming, lemmatization, and nominalization) in the agribusiness domain. They removed standard stopwords for the Portuguese available at PRETEXT [76], together with the conjugations of the verb TO BE as well as the words composed of only one character. In the sequence, they applied statistical measures, as follow: for unigrams, bigrams, and trigrams, they used document frequency (df $\geq$ 2) and for the bigrams and trigrams, they applied log likelihood ratio. They removed unigrams considering their df values formed a new list of words, denominated 'stoplist of the collection' or 'stoplist of the domain', and this list is incorporated to the standard stopwords and used to form *n*-grams. The extracted terms were evaluated by the authors in an objective manner - using, for instance, the ctw [51] measure - and in a subjective way with the support of domain experts. This term extraction approach was focused to be used in the TopTax project [77], whose explanation is included in this section, although it may be used for other objectives.

There is also the work of *Lopes et al.* [71] that used the OntoLP tool [53] to compare three ways of term extraction (bigrams and trigrams) based on the linguistic approach. The first way considered only *n*-grams, the second one used pre-established morphosyntactic patterns, while the third one only considered the noun phrases. Both, for the identification of morphosyntactic patterns and noun phrases, the authors used the PALAVRAS parser [75]. The authors compared these three strategies among themselves. Moreover, they added information regarding semantic groups to each one of these forms, generating, now, three new ways of ATE, which were compared with the strategies that did not use semantic information. These semantic groups are prototypic information supplied by the PALAVRAS parser and made available by the OntoLP tool. Such information classify common names in general classes. An example given by the authors is the tag '<*an*>' attributed to the noun '*músculo*' ('muscle'), which indicates that the word belongs to the class '*Anatomia*' ('Anatomy'). As a result, the best F-measure values for bigrams were 11.51% and for trigrams were 8.41%, obtained with the JPED corpus, considering the noun phrases, excluding terms by semantic groups.

In the contributions of *Lopes et al.* [67,68], the authors performed a comparative analysis of the extraction of bigrams and trigams using a linguistic and a statistical approach. They extracted these terms from the JPED corpus. The linguistic approach used the EχATO*LP* tool [19] to identify noun phrases from a corpus previously noted by the PALAVRAS parser [75]. The statistical approach used the NSP package to identify terms that contained an absolute frequency superior to a given value. Also for the statistical approach, they removed (i) stopwords; (ii) text structural demarcations, such as '*Introduction*' and '*References*'; and (iii) the candidate terms whose words began with capital letters in order to remove proper nouns, such as '*São Paulo*'. The extracted terms were evaluated with the support of a gold standard of *n*-grams. As a result, the authors state that the statistical approach presents high simplicity in its execution. However, they obtained better results when using the linguistic approach. The values obtained by the tool, for a corpus of the JPED corpus, have the F-measure = 34.48% for bigrams and the F-measure = 38.37% for trigrams [68].

Lopes et al. [72] performed term extraction (bigrams and trigrams) from the JPED corpus. This extraction used the OntoLP tool [53], that considers three different ways to extract terms: (i) the more frequently used *n*-grams, (ii) candidates that follow some morphosyntactic patterns; and (iii) noun phrases. Additionally, the authors tested different cut-off points. The best F-measure values were, for bigrams, 56.84% when using POS and, for trigrams, 52.11% when using the frequency of *n*-grams, both values were achieved considering the thresholds of 5E-6 and 6E-6 for absolute cut-off points.

In the contributions of *Muniz and collaborators* [11,12,61], the authors presented the NorMan Extractor tool[29], which extracts terms from instruction manuals using the hybrid approach, such as manuals of appliances. The term extraction is based on specific relations existing in the genre in question. That is, the instruction manuals have two basic procedural relations: relation '*gera*' ('generation'), when an action *A* automatically generates an action *B*, and the relation '*habilita*' ('enablement'), when the realization of an action *A* allows the realization of action *B*. The steps taken for the term extraction were the following: Firstly, the user selects an instruction manual to be used. There is also the possibility of the user to submit a corpus on the domain so it may be used in the calculation of the c-value measure [49]. This measure is used for the extraction of composed candidate terms. Then, the instruction manual is noted by the PALAVRAS parser [75]. From this notation, it is possible to extract the terms using the '*gera*' and '*habilita*' relations. Lastly, the lists of unigrams, bigrams, and *n*-grams extracted are presented to the user, allowing the user to perform a cut using the offered values by the c-value measure for each extracted term. Considering there

is not a gold standard for instruction manuals, the authors did not present results using the F-measure. However, the results were compared with other methods of extraction focused on scientific papers. Additionally, the authors also used a statistical measure, *Kappa* [78], which indicates the concordance among annotators at the same time that it discounts the concordance by chance.

Lopes and Vieira [10] extracted terms using the linguistic knowledge. For this, they only considered the noun phrases that fit one of the 11 proposed linguistic heuristics. An example of these heuristics is the removal of the NPs that begin with an adverb. The best values for F-measure in the experiments carried out using the JPED corpus were 64% for bigrams and 50% for trigrams.

Lopes and co-workers [9,34] extracted bigrams and trigrams based on the same linguistic methods used by Lopes and Vieira [10] and ordered them using the numerical values obtained by the application of the following statistical measures: tf, tf-idf, tds, thd, TF-IDF (refered in this work in upper-case letters to differentiate it from the tf-idf measure). Lopes has also proposed and used the tf-dcf. That measure, according to the author, considers the absolute frequency of the term as a primary indication of the relevance of a term, and penalizes the terms that occur in the contrasting corpora of other domains dividing the term absolute frequency in the corpus of the domain by the geometric composition of the absolute frequency in each of the contrasting corpora. After the ordering of the terms by each of these measures, cut points were chosen and applied to the ordered lists of terms. For the experiments, they used the JPED corpus and four other contrasting corpora [79], which are Stochastic modelling, Data mining, Parallel processing, and Geology. The precision of bigrams and trigrams extracted from the JPED corpus were evaluated in the following scenarios: (i) comparison of the linguistic heuristics adopted for the selection or removal of NPs, while it is possible to show that the use of the proposed heuristics significantly improve results; (ii) comparison of the statistical measures used, while, for this corpus, the precision rates are higher when the tf-dcf is used; and (iii) comparison of the variation of the contrasting corpora using the tf-dcf measure, that made it possible to show that when the four contrasting corpora are used together, better results are obtained. As results and considering cuts in the number of candidate terms, the author obtained the F-measure values equal to 81% for bigrams and 84% for trigrams.

Conrado, Pardo, and Rezende [5] presented a term extraction approach (unigrams) in which inductors classify the words in terms or non-terms. This classification is based on a set of 19 characteristics identified for each word. These characteristics use linguistic knowledge (such as noun phrases and POS), statistical knowledge (such as tf and tf-idf), and hybrid knowledge (such as the frequency of words of a corpus in the general language and the analysis of the context of the words). For the experiments, three corpora, from different domains, were used: Ecology (ECO), Distance education (DE), and Nanoscience and Nanotechnology (N&N). The authors tested two different cutoffs (*C1* and *C2*). In *C1*, only unigrams that occur in at least two documents in the corpus were preserved. In *C2*, considering the candidates of *C1*, the authors preserved only the unigrams that occur in noun or prepositional phrases and also follow some of these POS: nouns, proper nouns, verbs, and adjectives. The best F-measure values were 24.26%, 17.58%, and 54.04% for ECO, EaD, and N&N, respectively. Among the identified characteristics for the candidate terms, tf-idf [33] was the one that better supported the term extraction, followed by *N_Noun* (created by the authors, this characteristic counts how many nouns originated the candidate when it was normalized), and TVQ [37] (characteristic that considers that the terms do not have low frequency and at the same time keep a non-uniform distribution through the corpus).

In the work of *Conrado et al.* [6], the authors used the same characteristics and corpora of those used in [5]. However, in [6], the authors followed the *C2* cutoff, described in [5]. The best F-measure values were 23.40%, 18.39%, and 48.30% for ECO, EaD, and N&N, respectively. Additionally, in [6], they discussed how the characteristics of different levels of knowledge help classifying terms and gave examples of extracted candidates correctly and incorrectly.

There is also the work of *Conrado et al.* [7] that proposed the use of transductive learning to ATE. Transductive learning performs the classification spreading the labels from labeled to unlabeled data in a corpus. The advantage of this learning is that it needs only a small number of labeled examples (candidates) to perform the classification. The authors extracted terms based on a set of 25 characteristics identified for each unigram. These characteristics use linguistic knowledge (such as noun phrases and POS), statistical knowledge (such as tf and tf-idf), and hybrid knowledge (such as the behavior of a candidate in a corpus of general language and the analysis of the context of the words). The experiments used a corpus of the ecology domain (ECO) and achieved 27% of the F-measure while the best F-measure for the same corpus when using inductive learning with 19 characteristics was 24% [5].

The previously described contributions used different corpora and, in some cases, the authors used different evaluation measures for the results or evaluated only a part of the list of candidate terms. Even considering such differences, aiming at providing a general overview of the results of the contributions on term extraction for the

Portuguese language, Table 3 presents a summary of the results and their best F-measure values.

Research related to Brazilian Portuguese term extraction

Some investigations discuss about term extraction in the Brazilian Portuguese language.

Almeida et al. [65] used the corpus of the Industrial Ceramics Magazine to discuss the manual and automatic process of term extraction. The authors stated that the manual extraction carried out by domain experts, in which the experts indicated the terms of the domain, is considered as a semantic criterion. In addition, the authors analysed the candidate terms obtained in the extraction considering the three cases: with and without stopword removal and with the correction of possible errors in the used corpus. For the extraction of unigrams, the authors used the frequency measure and, for bigrams, they compared mutual information, log likelihood ratio, and frequency.

In the work of *Teline* [31], the author carried out a bibliographical review on the weak and strong points of the term extraction methods in these three approaches, statistical, linguistic, and hybrid. Regarding the statistical approach, the author compared the frequency, mutual information, log likelihood ratio, and Dice's coefficient measures. For the linguistic approach, Teline removed stopwords and used some indicative phrases, such as '*definido(a)(s) como*' ('defined as'), '*caracterizado(a)*' ('described as'), '*conhecido(a)(s) como*' ('known as'), '*significa(m)*' ('mean(s)'), as well morphosyntactic patterns. In the hybrid approach, the author combined the linguistic approach with the frequency measure separately for the extraction of unigrams, bigrams, and trigrams. Also, for the extraction of bigrams and trigrams, separately, the author combined the linguistic approach with frequency with the mutual information measure. Each one of these extractions was evaluated in the domain of ceramic coating. The hybrid approach (linguistic part with frequency) presented better F-measure values, which were 11%, 17%, and 33% for unigrams, bigrams, and trigrams, respectively. Considering a manual selection (a linguist and the author) carried out in the candidate terms, it has reached the F-measure values for the unigrams, also ordered by the frequency measure of 58% and for trigrams using Dice's coefficient, obtained at 26%.

Almeida and Vale [27] discussed about the specific morphological patterns that occur in three domains: ceramic coating, physiotherapy, and nanoscience and nanotechnology. As results, for the ceramic coating domain, they obtained a high frequency of combinations, such as '*argila refratária aluminosa*' ('alumina refractory clay') and '*análise granulométrica por peneiramento*' ('granulometric analysis by sieving'), and of simple words followed by morphemes that may be useful as term identifiers, such as derivational suffixes *-agem*, *-ção*, for instance '*secagem*' ('drying') and '*moagem*' ('grind'). For the physiotherapy domain, there are many erudite formations, of Greek or Latin origin, due to the fact that such terminology has many terms from medicine, such as '*arthr(o)-*' ('arthr(o)-') that may form, for instance, the terms '*artralgia*' ('arthrodesis') and '*artrite*' ('arthritis'). Regarding the nanoscience and nanotechnology domain, the most remarking characteristic is the high absolute frequency of the *nano-* prefix, which may originate, for instance, the terms '*nanocristais*' ('nanocrystals') and '*nanossistema biológico*' ('biological nanosystem').

The most common way to extract terms is to attribute a value for each candidate term according to some measure. Therefore, the candidates are ranked using their values and it is necessary to know how to perform the cutoff, i.e., to know until which value/candidate should be considered as good candidate. *Lopes and Vieira* [70] discussed and compared three different forms of candidate cutoffs, which were (i) absolute cutoff, the authors ranked the candidates using the tf-dcf measure and performed cutoffs considering intervals between 100 and 3,500 first candidates; (ii) threshold cutoff, they carried out cutoffs in the Pediatrics domain considering the frequency of the candidate in the corpus (0 until 15); and (iii) relative cutoff, they also used tf-dcf and removed percentage of candidates (1% until 30%). Finally, they analyzed these three cutoffs and proposed the combination of threshold and relative cutoffs, in which mantained candidates that have tf-dcf $>$ 2 and correspond up to 15% of the ranked candidades.

Projects related to the Brazilian Portuguese term extraction

In this section, we presented some of the main projects related to the term extraction in the Brazilian Portuguese language, namely: NANOTERM, TEXTQUIM/TEXTECC, Bio-C, E-TERMOS, and TermiNet. For each project, we highlighted where they applied term extraction by using italicized words. We also described the OntoLP portal and the Linguateca repository, in which researchers may have found resources to perform term extraction. A summary of these projects, portal, and repository is presented in Table 4 and their main characteristics are highlighted.

The NANOTERM Project

The project named 'Terminology in the Portuguese language of Nanoscience and Nanotechnology: Systematisation of Vocabular Repertory and Creation of a Pilot Dictionary' (NANOTERM) [63] was developed between 2006 and 2008 in GETerm of the Federal University of São Carlos with the collaboration of NILC of the University of São Paulo.

Table 3 Summary of the contributions on term extraction

Contributors	Domains	F-measure values		
		Unigrams	Bigrams	Trigrams
Almeida, Aluísio, and Teline [65]	Ceramic coating	-	-	-
Almeida and Vale [27]	Ceramic coating, physiotherapy, and nanoscience and nanotechnology	-	-	-
Conrado [52] and Conrado et al. [69]	Agribusiness	-	-	-
Conrado, Pardo, and Rezende [5]	Ecology, distance education, and nanoscience and nanotechnology	54.04%	-	-
Conrado et al. [6]	Ecology	23.40%	-	-
	Distance education, and	18.39%	-	-
	nanoscience and nanotechnology	48.30%	-	-
Conrado et al. [7]	Ecology	27.00%	-	-
Honorato and Monard [8]	Medicine (medical reports)		59.00%	
Lopes [9] and Lopes and Vieira [70]	Pediatrics	-	81.00%	84.00%
Lopes, Fernandes, and Vieira [34]	Pediatrics	-	-	-
Lopes, Oliveira, and Vieira [67,68]	Pediatrics	-	51.42%	41.26%
Lopes and Vieira [10]	Pediatrics	-	64.00%	50.00%
Lopes and Vieira [70]	Pediatrics	-	81.00%	84.00%
Lopes et al. [71]	Pediatrics	-	11.50%	8.40%
Lopes et al. [72]	Pediatrics	-	56.84%	52.11%
Muniz and Aluísio [11], Muniz et al. [12] and Muniz [61]	Appliances	-	-	-
Ribeiro Junior [53] e Ribeiro Junior and Vieira [14]	Pediatrics	22.39%	10.04%	5.46%
Teline [31]	Ceramic coating	11.00%	17.00%	46.00%
Teline, Manfrin, and Aluísio [66]	Ceramic coating	26.00%	9.00%	0.62%
Zavaglia et al. [16,17]	Ecology	16.48%	16.88%	5.77%

The objectives of this project were (i) the constitution of a corpus in the Portuguese language of nanoscience and nanotechnology; (ii) the search for equivalents in Portuguese (input language) from a nomenclature in English (output language); (iii) creation of an ontology in the Portuguese language of the domain of nanoscience and nanotechnology, and (iv) the elaboration of the first pilot-dictionary of nanoscience and nanotechnology in the mother language.

The *semi-automatic term extraction* in this project is related to the obtainment of the terminological set that will compose the nomenclature of the dictionary or glossary and it is done a semi-automatic manner, as in this task, the role of the linguist is always foreseen, in addition to the automatic work carried out with the NSP package. Nomenclature is understood as the set of lexical units[30] that will constitute the inputs of the glossary or dictionary. For the term extraction, the E-TERMOS computational environment is used, which is described afterwards. At last, the extracted terms are inserted in the ontology in the Portuguese language in the domain of nanoscience and nanotechnology.

The TEXTQUIM/TEXTECC Project

The project named Texts of Chemistry (TEXTQUIM)[31], which began in 2003 and is developed by the Federal University of Rio Grande do Sul (UFRGS), is becoming project Technical and Scientific Texts (TEXTECC)[32] because it will also comprise the domains of Chemistry, Physics, Pediatrics, Cardiology, Nursing, and Veterinary.

The objective of this project is to develop a dictionary to support translation students, initially in the domain of pediatrics. For the purpose of studying patterns of the Portuguese-English translation, Coulthard [54] built a corpus, namely JPED, which is composed of 283 texts (785,448 words) in the Portuguese language extracted

Table 4 Summary of the projects, portal, and repository related to the term extraction

Names	Responsible	Periods	Objectives	Domains
Bio-C	GETerm and NILC	2007 to 2009	Generate the systematized terminology of the biofuel domain	Biofuel
E-TERMOS	EMBRAPA, GETerm, and NILC	2009 to Today	Terminological management	Several
NANOTERM	NILC and IFSC	2006 to 2008	Constitute corpus, build ontology, and elaborate pilot-dictionary	Nanoscience and nanotechnology
TermiNet	GETerm and NILC	2009 to Today	Build ontologies, develop terminological textual bases, and build a WordNet	Several
TEXTQUIM / TEXTECC	UFRGS	2003 to Today	Develop dictionary for translation	Chemistry, Physics Pediatrics, Cardiology Nursing, and Veterinary
TOPTAX	LABIC and EMBRAPA	2005 to Today	Organize and keep information on specific domains	Several
OntoLP	PUCRS	2008	Divulge tools and resources	Several
Linguateca	IST-UTL, UC, and PUC-Rio	1998	Maintain linguistic resources	Several

from the Journal of Pediatrics. In the scope of project TEXTQUIM-TEXTECC, it was carried out a manual term extraction (without linguistic notation) from this corpus considering only *n*-grams that occurred more than four times in this corpus. In the sequence, it was carried out a filtering based on heuristics that resulted in a new list of *n*-grams considered as possibly relevant to integrate the glossary. These *n*-grams were evaluated in relation to their relevance and manually refined by translation students with knowledge of the domain. With this process, the gold standards of the JPED corpus were originated. These gold standards have been used in experiments of *composed term extractions* and concept candidates, as the case of the OntoLP project [53].

The Bio-C Project

The project named 'Biofuel Terminology: morphological and semantic description aiming at systematisation' (Bio-C)[33] was developed between 2007 and 2009, by GETerm of the Federal University of São Carlos together with the support of NILC of the University of São Paulo.

The purpose of this project is to generate the systematized terminology of the biofuel domain, including the fundamental terms of the aforementioned domain, which includes the sub-domains of the ethanol and bio-diesel, in order to support the creation, *a posteriori*, of the first glossary of this knowledge domain in the Brazilian Portuguese language.

For the *semi-automatic term extraction* in this project, the NSP package was used to generate lists of candidate terms (unigrams, bigrams, trigrams, quadrigrams, and pentagrams). Stopwords were removed to reduce the excess of noise in the lists of candidate terms. In the sequence, such lists of candidate terms were manually cleaned by a linguist, as well as *a posteriori* validation of the candidates by a domain expert. As a result, it is expected to obtain validated terms that will integrate the area glossary.

The E-TERMOS *Project*

The 'Electronic terms' (E-TERMOS[34]) project [80] originated from the transformation of the TermEx project to E-TERMOS. It is a free WEB collaborative computational environment with free access dedicated to the terminological management. This project was developed at NILC of the University of São Paulo with the collaboration of

GETerm of the Federal University of São Carlos and of the Brazilian Agricultural Research Corporation (EMBRAPA).

The main objective of E-TERMOS is to make the creation of terminological products possible, whether they are for academic research or promotion purposes, by means of the (semi) automation of the stages of the terminological work. The goal of the *automatic term extraction* in this project is to obtain candidate terms from the corpora of the specificity in question. In order to perform the extraction, firstly, it is possible to choose the size of the gram to be used, which may be from 2 to 7. Then, it is possible to remove stopwords with the use of a list of provided stopwords, which is a result of the work of Teline [31]. After the removal of the stopwords, the terms are extracted with the support of the statistical and/or linguistic knowledge. According to the author, the incorporation of statistical measures (log likelihood ratio, mutual information, and Dice's coefficient) from the NSP package, linguistic (to be defined) and hybrid (union of statistical and linguistic knowledge) are to be included. Nowadays, the simple frequency statistical measure is available.

For the edition of the conceptual map of term categorization, in E-TERMOS, the creation, edition, and visualization of the conceptual maps and computational resources for the insertion and evaluation of the terms by experts is allowed.

Therefore, the management of the terminological database is obtained, in which the terminological record is created and filled and the definitional base is elaborated, with the support of tools that manage the terminological database.

Finally, in the stage of interchange and diffusion of terms, the entries are edited and the diffusion, interchange, and query of the terminological products may be performed with the help of applying terminological data exporting tools, making it possible for the users to query the entries.

The TermiNet Project

The TermiNet project (Terminological WordNet) [81] is under development, since 2009, at the laboratory of GETerm of the Federal University of São Carlos with the collaboration of NILC of the University of São Paulo.

This project has two main objectives. The first one is to develop a generic semi-automatic methodology, based on corpus, for the building of lexical databases in the WordNet format. The second objective is to validate this methodology with the help of the building of a TermiNet.

The *candidate term extraction* of TermiNet uses the linguistic approach with the help of the EχATO*LP* [67] and OntoLP [14] tools, as well as the statistical approach with the use of the NSP [73] package. Experts in the domain in question carry out a manual validation of the candidate terms. The candidate terms are also compared to a list of lexical units from a contrasting corpora.

The TOPTAX *Methodology*

The Topic Taxonomy Environment (TOPTAX)[35] [77] methodology aims at organizing and maintaining information of specific domains. This is possible due to the creation of a topic taxonomy on the domain knowledge represented by the collection of texts. The considered taxonomy is a hierarchical topic organization extracted from a collection of texts, in which the upper topics are *parents* of the lower topics, i.e., the lower topics are specializations of the upper topics. In addition, it is possible to associate resources of the textual base at each level of the taxonomy, referring to its domain, thus, facilitating the organization of the information under this taxonomy.

TOPTAX, in order to achieve its objectives, follows the stages of the Text Mining process [82], which are problem identification, pre-processing, pattern extraction, post-processing, and use of the knowledge.

The stage of problem identification must delimit the problem to be tackled by selecting and retrieving the documents that form a textual collection to be worked with.

In the pre-processing stage, the documents of the obtained textual base are prepared to serve as input for the tools that will be used. In this stage, the documents are converted to the form of plain text without formatting. Afterwards, the words of the documents are normalized using one of the word normalization techniques (stemming, lemmatization, or nominalization) and stopwords are removed. Next, *terms are extracted*, and therefore, they are used to describe the text base, as detailed in Conrado [52]. To reduce the amount of terms to be worked with, a term selection is performed by using, e.g., the Luhn, Salton, and term variance, methods, which are detailed in the work of Nogueira [38].

In the pattern extraction stage, the document hierarchical clustering is performed in order to build a topic taxonomy. With the hierarchy, the obtained clusters keep topics or sub-topics to which the documents refer to. In the sequence, as described in Moura et al. [77], the descriptors for each group found by obtaining the most significant terms are identified, while it is possible to add resources of topic information to each node, such as documents, videos, and associated images.

In the post-processing stage, this obtained hierarchy is visualized and validated. The knowledge regarding the domain at hand represented in this hierarchy is then used to support the decision and organization of the information contained there (stage of Knowledge Use).

The OntoLP Portal

The 'Portal de Ontologia' (OntoLP)[36] [53] is developed by the Group of Natural Language Processing of the University of Rio Grande do Sul (PUCRS) and has the objective of divulging the available ontologies in the Portuguese language, as well as terminological bases, controlled vocabularies, and even more complex ontologies of the OWL-DL (Web Ontology Language-Description Logics) type, and tools and resources related to the research in the area.

The Linguateca Repository

The Linguateca Repository[37] is formally named *Centro de Recursos – distribuído – para a língua portuguesa* and was officially created in 2002, but the initial contributions related to it started in 1998. Linguateca consists of a repository of linguistic resources focused on the Portuguese language. The responsibility on Linguateca, since its start (in 1998) used to be passed from pole to pole in several colleges. From 2009 on, it was established that the responsibility on it would be given only to the Oslo operational pole. The people in charge of this pole are four researchers (Diana Santos, Cristina Mota, Rosário Silva, and Fernando Ribeiro) of *Instituto Superior Técnico, Universidade Técnica de Lisboa* (IST-UTL)[38] and *Universidade de Coimbra* (UC)[39], two Ph.D. students in Portugal (Nuno Cardoso and Hugo Oliveira) of IST-UTL, all in Portugal, as well as a Brazilian researcher (Maria Cláudia de Freitas) of *Pontifícia Universidade Católica do Rio de Janeiro* (PUC-Rio[40]).

Next, we discuss about the state of the art of term extraction in the Brazilian Portuguese language.

Discussion about the state of the art in term extraction for the Brazilian Portuguese language

As observed in Figure 3 (Section 'State of the art of term extraction in Brazilian Portuguese'), 77% of the presented contributions only used the statistical approach or combined it with some not very sophisticated linguistic information, as noun phrases and morphosyntactic patterns. Up to date, it was possible to find only two papers that used linguistic knowledge of the semantic level, [71] and a dissertation [53] published in [14]. However, we found no contribution that used knowledge at the pragmatic level. The higher use of the statistical knowledge in regard to the linguistic is due to the lack of efficient and advance resources of NLP for the processing of the Portuguese language. In other languages, such as English, Spanish, and French, the term extraction task uses advanced resources, such as WordNet, specific domain ontologies, thesaurus, and different parsers and disambiguation resources. Some of the resources for the Portuguese language are available at the Ontology Portal (OntoLP) and at Linguateca, both described in subection 'Projects related to the Brazilian Portuguese term extraction'.

Among the 25 contributions related to the term extraction in the Portuguese language described in subsections 'Research developed for Brazilian Portuguese term extraction' and 'Research related to Brazilian Portuguese term extraction', we highlight the work of Conrado, Pardo and Rezende [5], which obtained a higher F-measure value for unigrams, 54.04%, in the nanoscience and nanotechnology domain, and the work of Lopes and Vieira [9,70], which obtained a higher F-measure value for bigrams, 81%, and for trigrams, 84%, both in the pediatrics domain. It is important to notice that this comparison was performed not considering existing differences among the contributions, for example, the use of different corpora and the evaluation performed using only part of the list of candidate terms. The performed comparison aimed at providing a general overview of the results of the contributions. It is clear that, in order to state which extraction methods among the methods used by the authors are better or worse, it is necessary to carry out a comparison of all contributions using the same scenario, i.e., the same corpora and the same measures and evaluation conditions. Consequently, this comparison with the same scenario is the existing gap in the Portuguese language, which makes it difficult to choose a method as a baseline.

Additionally, it is possible to make observations related to the corpora described in Section 'Corpora for Portuguese' when considering the contributions that present the F-measure scores or values that allow the F-measure calculation.

The first observation is that these contributions normally used only one corpus to evaluate their results, while in the Portuguese language, we found five corpora that could be used since they have gold standards (see Table 2). Furthermore, such contributions, as a whole, focus only on three domains of corpora: ceramic coating, ecology, and pediatrics.

The second observation is that 95% of the 25 contributions (subsections 'Research developed for Brazilian Portuguese term extraction' and 'Research related to Brazilian Portuguese term extraction') performed extraction of bigrams and trigrams, but only 64% of them also (or only) extracted the unigrams. We may conclude that for Portuguese, the term extraction is limited to unigrams, bigrams, and trigrams since, generally, they are the most common terms in the documents and the terms that are in the gold standards.

The third observation is related to the knowledge used by each of these contributions. Frequency is the most used measure to obtain statistical knowledge, while 45% of the 25 contributions applied it to extract unigrams,

38% for bigrams, and 47% for trigrams. The other most used measures were: for unigrams, 24% used the df measure, for bigrams, 30% used mi and log likelihood ratio, and for trigrams 35% applied log likelihood ratio. Regarding the acquisition of linguistic knowledge, the information on morphosyntactic patterns is the most used one, occurring in 45% of the contributions that extracted unigrams and in 28% of the contributions that extracted bigrams and trigrams. Considering the contributions that extracted unigrams, 19% of the contributions used indicative phrases and 9% considered noun phrases for ATE; for the extracting of bigrams and trigrams, 23% used noun phrases. Finally, 19% of the contributions for the acquisition of hybrid knowledge (statistical and linguistic) used the c-value measure.

It is observed that there is a considerable gap related to the use of hybrid resources, in addition to the lack of use of WordNets and Wikipedia to support the extraction, while these resources are already used in other languages, such as Spanish, French, and English. Another existing gap, also present in the mentioned languages, is the evaluation of terms. The problem related to the evaluation is that the gold standard that is used is built in a subjective way and, consequently, this subjectivity will continue in the evaluation.

By analysing the resources used in the contributions, it may be observed that there is a tendency in the recent contributions to consider knowledge from domains, whether from the same domain in which the term extraction is being carried out, which is the case of the domain stoplists, or contrasting domains regarding the one used in the extraction, which is the case of the tds, tf-dcf, TF-IDF, and thd measures.

Figure 4 shows a mapping of the contributions, projects, and corpora related to the term extraction in the Brazilian Portuguese language. For such mapping, we considered only the state in which the first author was bound.

In this figure, we observed that, to the best of our knowledge, most of the cited contributions, projects, and corpora are concentrated in the states of Sao Paulo and Rio Grande do Sul. We found two projects and one corpus created by researchers of the Santa Catarina state. It is important to state that this fact does not indicate that there are no contributions related to the term extraction in the other states of Brazil, as some existing contributions might have not been published in conferences yet or, also, research partnerships with the aforementioned authors might exist.

Conclusions

In this paper, we present an updated survey of the state of the art of the term extraction focused on the Brazilian Portuguese language. There are five main issues observed with this survey of the state of the art.

The first issue is the diversity in which the term extraction task is inserted. We observed that the extraction has high importance for tasks of different areas. When the term extraction is carried out, there is, usually, a specific objective, such as the building of taxonomies and ontologies, the elaboration of dictionaries, the translation, the organization, or the retrieval of textual data. We notice this fact when observing the projects that use ATE for reaching other objectives, according described in subsection 'Projects related to the Brazilian Portuguese term extraction'.

The second issue is that the term extraction task may be used in different areas. We observed by the fact that there are contributions that do not perform the extraction directly, but they used terms to reach their final objectives. An example of that is the work of Tagnin [83], which used the terms to demonstrate how the corpus linguistic area may support the translator to find equivalents for technical terms in several areas. As this work did not perform the term extraction, but only used the terms, the same was not considered as a contribution of term extraction, specifically. For the same reason, we did not mention, in this paper, other contributions that performed the term extraction to reach specific objectives, for example, the clustering of texts and textual classification.

The second issue focuses on the way the aforementioned contributions extracted terms and how these contributions were organized in accordance to the knowledge used to extract terms. We must consider here the work of Lima et al. [84], in which it is stated that until 2007 the hybrid approaches generally worked more with the lexical, morphosyntactic, and syntactic levels considering the Brazilian Portuguese language. For languages whose investigations were more developed, such as English, we may find some contributions using the syntactic and semantic levels. In general, there is an indication of the lack of deeper work in these areas, mainly in the Portuguese language. However, until this day, we found in the literature neither an organization and a mapping of the contributions in the Brazilian Portuguese language, nor a synthesis of all the contributions. In this sense, in this paper, we identified, described, and mapped the available corpora to perform the term extraction task for the Brazilian Portuguese, the contributions on ATE, and the projects related to the ATE task.

Also, in relation to the approaches adopted for the extraction, the use of statistical methods for this task aggregates a large advantage in the extraction system, namely the independence of the target language. However, when the experts extract terms of certain domains, they, (sub)consciously, use their knowledge of the domain and of the linguistic to indicate the words or collocations that are really terms. The former is the explanation for the fact that the use of linguistic knowledge increases the quality

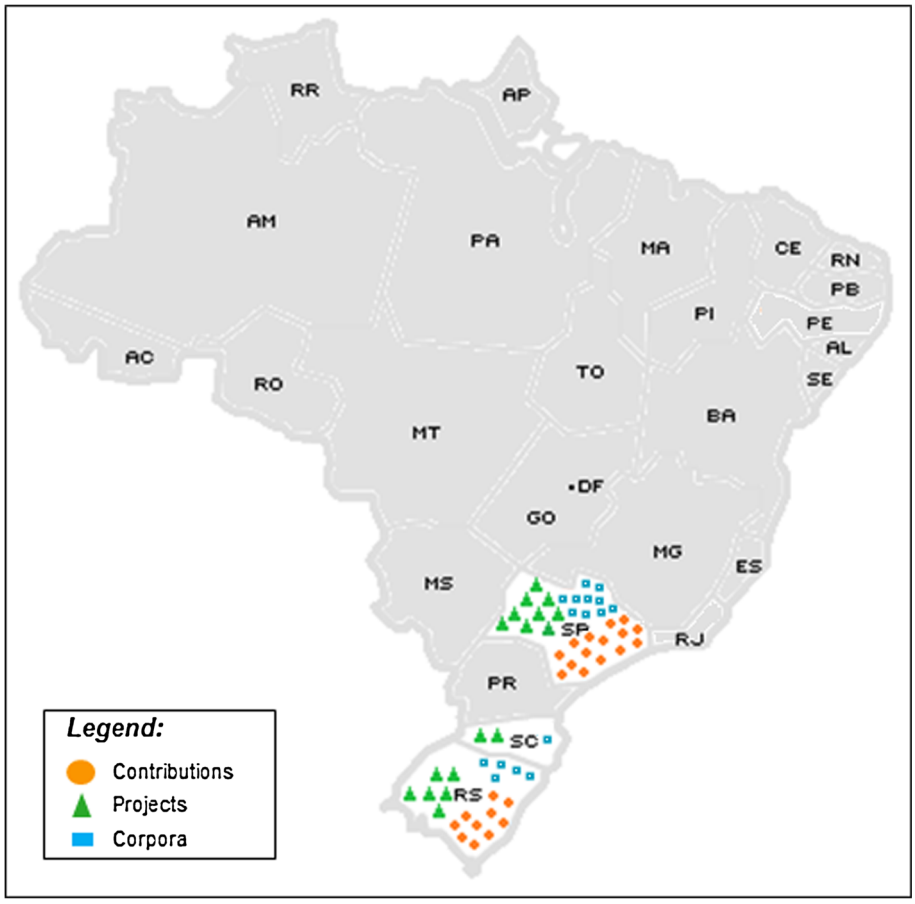

Figure 4 Mapping of contributions, projects, and corpora related to term extraction for Brazilian Portuguese language.

of the extraction. In this wise, the best way is to use statistical methods together with linguistic methods, having, therefore, a hybrid method.

The third observed issue refers to the advances obtained in the term extraction task. By analysing the contributions focused on the Brazilian Portuguese language, we may observe they are divided in relation to the resources they use. Normally, they use methods that contain two linguistic levels of knowledge: morphological and syntactic, while some few contributions only use the morphological level or purely statistical resources. This way, we may conclude that, for Portuguese, there is still a shortage in relation to the other linguistic levels of knowledge. This future advance will allow the improvement of the quality of the extracted terms. In addition to that, we may observe that there is a recent advance in corpora with gold standards, which also supports such improvement.

The fourth issue is the existence of a tendency in the recent contributions in considering knowledge from domains, whether from the same domain in which the term extraction is being carried out, which is the case of the domain stoplists, or from contrasting domains, in relation to the domain used in the extraction, which is the case of the tds, tf-idf, and tf-dcf measures.

As a fifth issue, we observed the increase of a practical comparison of the existing contributions for the extraction focused on the Brazilian Portuguese language.

Finally, with this survey, those who are interested in the systematization of terminologies by means of the automatic term extraction may find a detailed description of the main extraction paradigms (statistical, linguistic, and hybrid) and their linguistic-computational resources/tools, of the types of knowledge (statistical and linguistic) that may be used in ATE, in addition to an organization and mapping of the available corpora to perform the term extraction task and the main contributions and projects related to ATE for the Brazilian Portuguese language.

Endnotes

[1]According to the translation of the work of Sanchez [85], a corpus is defined as follows:

> 'A set of linguistic data (belonging to the oral or written use of the language, or to both) systematized according to certain criteria, which are sufficiently extensive in amplitude and depth, in such a way that they are representative of the totality of the linguistic usage or of some of their scope, disposed in such a way that they may be processed by a computer, aiming at providing useful and varied results for the description and analysis'.

Souza and Felippo [56], in the scope of the TermiNet project [81], summarize the criteria that define 'corpus' in representativeness, sampling, size, authenticity, diversity, and balancing.

[2]In this paper, we present all given examples in Portuguese and their English translations between brackets.

[3]Examples extracted from the work of Souza and Di Felippo [56].

[4]Examples extracted from the work of Almeida and Vale [27].

[5]Examples extracted from the work of Almeida et al. [86].

[6]Almeida et al. (2011, p. 32).

[7]Almeida et al. (2011, p. 27).

[8]'*Nano-*' (Greek prefix), which remits to '*nánnos*' ('of excessive minuteness') or '*nânos*' ('dwarf'), which is equivalent to a 10^{-9} multiplier of the indicated unit. In this regard, a '*nanômetro*' ('nanometer') corresponds to 10^{-9} m (1 nm = 10^{-9} m) [87].

[9] Definition taken from http://oxforddictionaries.com/definition/english/.

[10]Tokens are sequences of characters separated by blank spaces. In this regard, a token may represent a word, number, or punctuation mark.

[11]Stopwords, which constitute a stoplist, are basically functional words (for instance: prepositions, articles, conjunctions, etc), which present high frequency in the corpora and no terminological value.

[12]In the statistical approach, an *n*-gram is a sequence of *n* tokens (for example, unigram, bigram, trigram, etc).

[13]Text on traditional grammar and lexical categorization - www.dacex.ct.utfpr.edu.br/paulo3.htm.

[14]Tds, thd, TF-IDF, tf-dcf, and weirdness were classified as statistic measures in subsection 'Measures that express termhood'. Nonetheless, if we consider the use of contrastive corpora as linguistic resource, they may be classified as hybrid measures.

[15]For these measures, only noun phrases are considered as candidate terms.

[16]GETerm - http://www.geterm.ufscar.br/geterm2/.

[17]NLP Group - www.inf.pucrs.br/~linatural.

[18]LABIC - www.labic.icmc.usp.br/.

[19]NILC - www.nilc.icmc.usp.br.

[20]CIMM - www.cimm.com.br.

[21]CorpusDT - www.nilc.icmc.usp.br/nilc/projects/scipo.htm.

[22]USP-ICMC - www.icmc.usp.br/.

[23]UFSCar - www2.ufscar.br/.

[24]ECO - www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm.

[25]CNPTIA-EMBRAPA - www.cnptia.embrapa.br/.

[26]Folha-RIcol - www.linguateca.pt/Repositorio/Folha-RIcol/.

[27]JPED - www.jped.com.br.

[28]The Industrial Ceramic Magazine - www.ceramicaindustrial.org.br/.

[29]NorMan - http://nilc.icmc.usp.br/norman/extractor/.

[30]In the area of linguistics, the concept of 'lexical unit' is used in a more specific way. In this work, 'lexical unit' is used generically as a synonym of 'lexical item'.

[31]TEXTQUIM - http://www.ufrgs.br/textecc/textquim/.

[32]TEXTECC - www6.ufrgs.br/textecc/.

[33]Bio-C - www.geterm.ufscar.br/geterm2/?page_id=111.

[34]E-TERMOS - www.etermos.cnptia.embrapa.br/.

[35]TOPTAX - http://sites.labic.icmc.usp.br/toptax/.

[36]OntoLP - www.inf.pucrs.br/~ontolp/.

[37]Linguateca - www.linguateca.pt.

[38]IST-UTL - www.ist.utl.pt/.

[39]UC - www.uc.pt/.

[40]PUC-Rio - www.puc-rio.br/.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MSC performed the literature review, analysed the results, and drafted the manuscript. ADF, TASP, and SOR contributed in the literature review and were involved in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgments
This work was supported by Grants 2009/16142-3, 2011/19850-9, 2012/03071-3, and 2012/09375-4, from the Sao Paulo Research Foundation (FAPESP), and the National Counsel of Technological and Scientific Development (CNPq), Brazil.

Author details
[1]Laboratory of Computational Intelligence (LABIC), Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), P.O. Box 668, 13561-970 São Carlos, SP, Brazil. [2]Interinstitutional Center for Research and Development in Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo (USP), P.O. Box 668, 13561-970 São Carlos, SP, Brazil. [3]Interinstitutional Center for Research and Development in Computational Linguistics (NILC), Research Group of Terminology (GETerm), Federal University of São Carlos (UFSCar), Rodovia Washington Luís, km 235 - SP-310, CP 676, 13565-905 São Carlos, SP, Brazil.

Received: 15 May 2013 Accepted: 9 March 2014
Published: 30 May 2014

References

1. Cabré MT, Vivaldi REJ (2001) Automatic term detection: a review of current systems. In: Bourigault D, Jacquemin C, L'Homme MC (eds) Recent Advances in Computational Terminology. John Benjamins, Amsterdam, Philadelphia, pp 53–88
2. David S, Plante P (1991) Le progiciel termino: De la necessite d'une analyse morphosyntaxique pour le Wipouillentent terminologique des textes In: Proceedings of the Montreal Colloquium Les Industries de la Langue, Volume 1, Montreal, Canada, pp 21–24
3. Almeida GMdB, Oliveira LHM, Aluísio SM (2006) A terminologia na era da informática. Ciência e Cultura 58: 42–45
4. Estopá R, Martí J, Burgos D, Fernández S, Jara C, Monserrat S, Montané A, Munoz P, Quispe W, Rivadeneira M, Rojas E, Sabater M, Salazar H, Samara A, Santis R, Seghezzi N, Souto M (2005) La identificación de unidades terminológicas en contexto: de la teoría a la práctica. In: Cabré T, Bach C, Martí J (eds) Terminología y derecho: complejidad de la comunicación multilingüe. Cabré, T. and Bach, C. and Martí, J., Barcelona: Institut Universitari de Linguística Aplicada, pp 1–21
5. Conrado MS, Pardo TAS, Rezende SO (2013) A machine learning approach to automatic term extraction using a rich feature set In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) - Student Research Workshop (SRW). Atlanta, USA
6. Conrado MS, Rossi RG, Pardo TAS, Rezende SO (2013) Exploration of a rich feature set for automatic term extraction. In: Castro F, Gelbukh A, González M (eds) Advances in Artificial Intelligence and its Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp 342–354. [http://dx.doi.org/10.1007/978-3-642-45114-0_28]
7. Conrado MS, Rossi RG, Pardo TAS, Rezende SO (2013) Applying transductive learning for automatic term extraction: the case of the ecology domain In: Proceedings of IEEE the Second International Conference on Informatics & Applications (ICIA), Volume 1. Lodz, Poland, pp 264–269
8. Honorato DF, Monard MC (2008) Metodologia para mapeamento de informações não estruturadas descritas em laudos médicos para uma representação atributo-valor. Master's thesis, Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP). São Carlos, SP, Brazil
9. Lopes L (2012) Extração automática de conceitos a partir de textos em língua portugesa. PhD thesis, Porto Alegre, RS. Pontifícia Universidade do Rio Grande do Sul (PUCRS)
10. Lopes L, Vieira R (2012) Improving Portuguese term extraction. In: Caseli H, Villavicencio A, Teixeira ao P (eds) Proceedings of the Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR), Volume 7243 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp 85–92
11. Muniz F, Aluísio SM (2010) NorMan extractor: automatic term extraction from technical manuals (Resumo e Demonstração de ferramenta) In: the Ninth International Conference on Computational Processing of the Portuguese Language (PROPOR), Porto Alegre, RS, Brazil, pp 1–2. [Publicado em CD-ROM]
12. Muniz F, Watanabe WM, Scarton CE, Aluísio SM (2011) Extração de Termos de Manuais Técnicos de Produtos Tecnológicos: uma Aplicação em Sistemas de Adaptação Textual In: XXXVIII Seminário Integrado de Software e Hardware Volume 1. Natal, RN, Brazil, pp 1293–1306
13. Nazar R (2011) A statistical approach to term extraction. Int J Eng Stud 11(2): 159–182
14. Ribeiro Junior LC, Vieira R (2008) OntoLP: Engenharia de Ontologias em Língua Portuguesa In: Seminário Integrado de Software e Hardware - Anais do XXVIII Congresso da Sociedade Brasileira de Computação (SEMISH). Belém, PA: SBC, pp 1–15
15. Vivaldi J, Rodríguez H (2007) Evaluation of terms and term extraction systems: a practical approach. Terminology: Int J Theor Appl Issues Specialized Commun 13(2): 225–248
16. Zavaglia C, Aluísio SM, Nunes MGV, Oliveira LHM (2007) Estrutura Ontológica e Unidades Lexicais: uma aplicação computacional no domínio da Ecologia In: Proceedings of the Fifth Workshop em Tecnologia da Informação e da Linguagem Humana (TIL) - Anais do XXVII Congresso da Sociedade Brasileira da Computação (SBC), Rio de Janeiro, Brazil, pp 1575–1584
17. Zavaglia C, Oliveira L, Nunes M, Teline M, Aluísio S (2005) Avaliação de métodos de Extração Automática de Termos para a Construção de Ontologias. Tech. Rep. 248, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos. SP, Brazil
18. Korkontzelos I, Klapaftis IP, Manandhar S (2008) Reviewing and evaluating automatic term recognition techniques. In: Nordström B, Ranta A (eds) Proceedings of the 6th International Conference on Advances in Natural Language Processing, Volume 5221 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg, pp 248–259
19. Lopes L, Fernandes P, Vieira R, Fedrizzi G (2009) EχATO$_{LP}$ - An automatic tool for term extraction from Portuguese language corpora In: Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC), Faculty of Mathematics and Computer Science of Adam Mickiewicz University, pp 427–431
20. Ahrenberg L (2009) Term extraction: a review. [http://vir.liu.se/~lah/Publications/tereview_v2.pdf[10/10/2013]]. [Draft Version 091221]
21. Kageura K, Umino B (1996) Methods of automatic term recognition - a review. Terminology 3(2): 1–23
22. Pazienza MT, Pennacchiotti M, Zanzotto FM (2005) Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis S (ed) Knowledge Mining Series: Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, Berlin, pp 255–279. [http://dx.doi.org/10.1007/3-540-32394-5_20]
23. Barros LA (2004) Curso básico de terminologia. Acadêmica (São Paulo, Brazil), Edusp
24. Sager JC (1990) A practical course in terminology processing. J. Benjamins Publishing Company, Manchester, UK
25. Batista RP (2011) Caracter Ãsticas de terminologia empresarial: um estudo de caso. Master's thesis, Programa de Pós-Graduação em Linguística Aplicada - Universidade do Vale do Rio dos Sinos (UNISINOS), RS, Brazil
26. Gianoti AC (2012) Descrição sintático-semântica dos termos do domínio da Educação à Distância em português do Brasil. [Trabalho de Conclusão de Curso (Graduação em Bacharelado em Linguística) - Universidade Federal de São Carlos]
27. Almeida GMB, Vale OA (2008) Do texto ao termo: interação entre Terminologia, Morfologia e Linguística de Corpus na extração semi-automática de termos. In: Isquerdo AN, Finatto MJB (eds) As Ciências do Léxico: Lexicologia, Lexicografia e Terminologia, Volume IV, 1 edition, Campo Grande, MS, Brazil, UFMS, pp 483–499
28. Manning CD, Schütze H (2001) Foundations of statistical natural language processing. MIT Press, Cambridge, Massachusetts
29. Church KW, Hanks P (1989) Word association norms, mutual information, and lexicography In: Proceedings of the 27th annual meeting on Association for Computational Linguistics (ACL). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 76–83
30. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3): 297–302
31. Teline MF (2004) Avaliação de métodos para a extração Automática de terminologia de textos em português. Master's thesis, São Carlos, SP, Brazil
32. Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, MA
33. Salton G, Buckley C (1987) Term weighting approaches in automatic text retrieval. Tech. rep., Ithaca, NY, USA. [http://ecommons.library.cornell.edu/bitstream/1813/6721/1/87-881.pdf(10/10/2008)]
34. Lopes L, Fernandes P, Vieira R (2012) Domain term relevance through *tf-dcf* In: Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI). CSREA Press, Las Vegas, USA, pp 1001–1007
35. Witten I, Moffat A, Bell T (1999) Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann, San Francisco, CA, USA
36. Liu T, Liu S, Chen Z (2003) An evaluation on feature selection for text clustering In: Proceedings of the 10th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, USA, pp 488–495
37. Liu L, Kang J, Yu J, Wang Z (2005) A comparative study on unsupervised feature selection methods for text clustering In: Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), Wuhan, China, pp 597–601
38. Nogueira BM (2009) Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos. Master's thesis, Instituto

de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP), São Carlos. SP, Brazil
39. Park Y, Patwardhan S, Visweswariah K, Gates SC (2008) An empirical analysis of word error rate and keyword error rate In: 9th Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, Brisbane, Australia, pp 2070–2073
40. Kit C, Liu X (2008) Measuring mono-word termhood by rank difference via corpus comparison. Terminology 14(2): 204–229
41. Kim SN, Baldwin T, Kan MY (2009) Extracting domain-specific words - a statistical approach In: Proceedings of the Australasian Language Technology Association Workshop, Sydney, Australia, pp 94–98
42. Ahmad K, Gillam L, Tostevin L (1999) University of Surrey participation in TREC8: weirdness indexing for logical document extrapolation and retrieval (WILDER) In: TREC, Gaithersburg, US, pp 1–8
43. Jurafsky D, Martin JH (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (Prentice Hall Series in Artificial Intelligence). Prentice Hall, 1 edition. [Neue Auflage kommt im Frahjahr 2008], Upper Saddle River, USA
44. Vivaldi J (2001) Extracción de candidatos a término mediante la combinación de estrategias heterogéneas. PhD thesis, Universidad Politécnica de Catalunya, Departament de Llenguatges i Sistemes Informàtics, Spain
45. Voutilainen A (2004) Part-of-speech tagging. In: MITKOV R (ed) The Oxford handbook of computational linguistics. Oxford University Express, pp 219–232
46. Sparck-Jones K, Willett P (eds.) (1997) Readings in information retrieval. San Francisco, CA: Morgan Kaufmann Publishers Inc, USA
47. Gonzalez MAI, de Lima VLS, de Lima JV (2006) Tools for nominalization: an alternative for lexical normalization In: Proceedings of the VII Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR) vol. 3960, Itatiaia, Brazil, pp 100–109
48. Carroll J (2003) Parsing. In: Mitkov R (ed) The Oxford Handbook of Computational Linguistics. Oxford University Press, pp 233–248
49. Frantzi KT, Ananiadou S, Tsujii JI (1998) The C-value/NC-value method of automatic recognition for multi-word terms In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL). Springer-Verlag, London, UK, pp 585–604
50. Barrón-Cedeño A, Sierra G, Drouin P, Ananiadou S (2009) An improved automatic term recognition method for spanish In: Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). Springer-Verlag, Berlin, Heidelberg, pp 125–136
51. Maynard D, Ananiadou S (1999) Identifying contextual information for multi-word term extraction In: Proceedings of Terminology and Knowledge Engineering Conference (TKE). Innsbruck, Austria, pp 212–221
52. Conrado MS (2009) O Efeito do uso de Diferentes Formas de Extração de Termos na Compreensibilidade e Representatividade dos Termos em Coleções Textuais na Língua Portuguesa. Master's thesis, Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP), São Carlos, SP, Brazil
53. Junior, Ribeiro LC (2008) OntoLP: construção semi-Automática de ontologias a partir de textos da Língua Portuguesa. Master's thesis, Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brazil
54. Coulthard RJ (2005) The application of corpus methodology to translation: the JPED parallel corpus and the Pediatrics comparable corpus. Master's thesis, Programa de Pós-Graduação em Estudos da Tradução (PPGET) - Universidade Federal de Santa Catarina (UFSC), SC, Brazil
55. Feltrim VD, Nunes MGV, Aluísio SM (2001) Um Corpus de Textos Científicos em português para a análise da Estrutura esquemática. Tech. Rep. NILC-TR-01-4, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, SP, Brazil, [http://www.nilc.icmc.usp.br/nilc/index.php/publications#technical_reports]
56. Souza JWC, Felippo AD (2010) Um exercício em lingüistica de Corpus no âmbito do Projeto TermiNet. Tech. Rep. NILC-TR-10-08, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos. SP, Brazil. [www.letras.ufscar.br/pdf/NILC-TR-10-08-SouzaDiFelippo.pdf]
57. Gianoti AC, Felippo AD (2011) Extração de conhecimento terminológico no projeto TermiNet. Tech. Rep. NILC-TR-11-01, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP. São Carlos SP, Brazil. [http://www.ufscar.br/~letras/pdf/NILC-TR-11-01_GianotiDiFelippo.pdf]
58. Cardoso P, Maziero E, Jorge M, Seno E, Di Felippo A, Rino L, Nunes M, Pardo T (2011) CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese In: Proceedings of the 3rd RST Brazilian Meeting Volume 1. Cuiabá, MT, Brazil, pp 88–105
59. Oliveira JFG, Rozenfeld H, Amaral DC (2008) II Assembléia Geral do IFM. Book & CD. [http://www.numa.org.br/]
60. Lopes L, Vieira R (2013) Building domain specific parsed corpora in Portuguese language In: Proceedings of the 10th National Meeting on Artificial and Computational Intelligence (ENIAC), Fortaleza, CE, Brazil, pp 1–12
61. Muniz FAM (2011) Extração de termos de manuais técnicos de produtos tecnológicos: uma aplicação em Sistemas de Adaptação Textual. Master's thesis, Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP), São Carlos, SP, Brazil
62. Coleti JS, Mattos DF, Genoves Junior LC, Candido Junior A, Di Felippo A, Almeida GMB, Aluísio SM, Oliveira Junior ON (2008) Compilação de Corpus em Língua Portuguesa na área de Nanociência/Nanotecnologia: Problemas e soluções, Volume 1. São Paulo, SP, Brazil: Stella E. O.Tagnin; Oto Araújo Vale. (Org.), 192 edition
63. Aluísio SM (2005) Desenvolvimento de uma estrutura conceitual (ontologia) para a área de nanociência e nanotecnologia. Tech. Rep 276. Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, SP, Brazil, [www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_276.pdf(26/06/2010)]
64. Coleti JS, de Mattos DF, de Barcellos, Almeida GM (2009) Primeiro dicionário em Língua Portuguesa de Nanociência e Nanotecnologia In: Caderno de Resumos do II Encontro Acadêmico de Letras (EALE), São Carlos, SP, Brazil, pp 1–10
65. Almeida GMB, Aluísio SM, Teline MF (2003) Extração manual e Automática de terminologia: comparando abordagens e critérios In: Workshop em Tecnologia da Informação e da Linguagem Humana (TIL) - Sixteenth Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAP) vol. 1, pp 1–12
66. Teline MF, Manfrin AMP, Aluísio SM (2003) Extração Automática de Termos de Textos em português: Aplicação e Avaliação de Medidas estatísticas de Associação de Palavras. Tech. Rep. 216, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos. SP, Brazil
67. Lopes L, Oliveira LHM, Vieira R (2009) Análise comparativa de métodos de extração de termos utilizando abordagens lingüistica e estatística. Tech. Rep 053. Pontifícia Universidade Católica do Rio Grande do Sul (PUC) Porto Alegre, [www3.pucrs.br/pucrs/files/uni/poa/facin/pos/relatoriostec/tr053.pdf(26/06/2010)]
68. Lopes L, Oliveira LHM, Vieira R (2010) Portuguese term extraction methods: comparing linguistic and statistical approaches (Poster) In: Additional Proceedings of 9th International Conference on Computational Processing of the Portuguese Language (PROPOR). Springer, Porto Alegre, RS, Brazil, pp 1–6
69. Conrado MS, Moura MF, Marcacini RM, Rezende SO (2009) Avaliando Diferentes Formas de Geração de Termos a partir de Coleções Textuais. Tech. Rep 334. Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, SP, Brazil, [www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_334.pdf(26/02/2010)]
70. Lopes L, Vieira R (2013) Aplicando pontos de corte para listas de termos extraídos In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL). SBC, Fortaleza, CE, Brazil, pp 79–87
71. Lopes L, Vieira R, Finatto MJ, Martins D, Zanette A, Ribeiro Junior LC (2009) Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde In: Revista Eletrônica de Comunicação, Informação & Inovação em Saúde (RECIIS) vol. 3, Rio de Janeiro, RJ, pp 76–88
72. Lopes L, Vieira R, Finatto MJB, Martins D (2010) Extracting compound terms from domain corpora. J Braz Comput Soc (JBCS) 16(4): 247–259
73. Banerjee S, Pedersen T (2003) The design, implementation, and use of the ngram statistic package In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Mexico, pp 370–381

74. Schmid H (1994) Probabilistic part-of-speech tagging using decision trees In: Proceedings of International Conference on New Methods in Language Processing (NEMLAP), Manchester, UK, pp 44–49
75. Bick E (2000) The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Arhus
76. Soares MV, Prati RC, Monard MC (2008) PreTexT II: Descrição da Reestruturação da Ferramenta de Pré-processamento de Textos. Tech. Rep. 333. Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, SP, Brazil, [www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_333.pdf(14/01/2009)]
77. Moura MF, Marcacini RM, Nogueira BM, Conrado MS, Rezende SO (2008) Uma abordagem completa para a Construção de taxonomias de Tópicos em um domínio. Tech. Rep. 329. Instituto de Ciências Matemáticas e de Computação (ICMC) - USP, São Carlos, SP, Brazil, [www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_329.pdf(10/04/2009)]
78. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Computat Linguist 22(2): 249–254
79. Lopes L, Vieira R (2010) Building domain specific corpora in Portuguese language. Tech. Rep 062. Pontifícia Universidade Católica do Rio Grande do Sul (PUC) Porto Alegre, [http://www3.pucrs.br/pucrs/files/uni/poa/facin/pos/relatoriostec/tr062.pdf(11/10/2013)]
80. Oliveira LHM (2009) e-Termos: Um ambiente colaborativo web de gestão terminológica. PhD thesis, Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP), São Carlos, SP, Brazil
81. Di Felippo A (2010) The TermiNet project: an overview In: Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) - Young Investigators Workshop on Computational Approaches to Languages of the Americas, Los Angeles, CA, USA, Association for Computational Linguistics, pp 92–99
82. Ebecken NFF, Lopes MCS, Aragão MCd (2003) Mineração de Textos. In: Rezende SO (ed) Sistemas Inteligentes: Fundamentos e Aplicações, 1 edition, Manole, pp 337–364
83. Tagnin S (2007) A identificação de equivalentes tradutórios em corpora comparáveis In: Anais do I Congresso Internacional da Associação Brasileira de Professores Universitários de Inglês (ABRAPUI). ABRAPUI, Belo Horizonte, MG, Brazil, pp 1–13
84. Lima VL, Nunes M, Vieira R (2007) Desafios do Processamento de Línguas Naturais In: Anais do XXVII Congresso da Sociedade Brasileira da Computação (SBC). SBC, SBC, Rio de Janeiro, Brazil, pp 2202–2216
85. Sanchez AE (1995) CUMBRE: Corpus Linguistico del Español Contemporaneo: Fundamentos, Metodologia, y Aplicaciones. SGEL, Madrid, Spain
86. Almeida GMB, Kamikawachi DSL, Manfrim AMP, Souza IP, Izumida FH, Di Felippo A, Zauberas RT, Melchiades FG, Boschi AO (2011) Glossário de Revestimento Cerâmico. In: Humanitas/Faculdade de Filosofia (ed) Cadernos de Terminologia, Volume 4, 1 edition. Alves, I. M. (Org.), Letras e Ciências Humanas (FFLCH) - USP, São Paulo, SP, Brazil, pp 03–56
87. Houaiss A, Villar MS (2001) Dicionário eletrônico Houaiss da língua portuguesa. Rio de Janeiro, Brazil, Editora Objetiva

doi:10.1186/1678-4804-20-12
Cite this article as: Conrado *et al.*: **A survey of automatic term extraction for Brazilian Portuguese.** *Journal of the Brazilian Computer Society* 2014 **20**:12.