Termout: a tool for the semi-automatic creation of term databases

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Nicolás Acosta
Facultad de Filosofía y Letras
Universidad Nacional de Cuyo
niacosta@ms.uncu.edu.ar

Abstract

We propose a tool for the semi-automatic production of terminological databases, divided in the steps of corpus processing, terminology extraction, database population and management. With this tool it is possible to obtain a draft macrostructure (a lemma-list) and data for the microstructural level, such as grammatical (morphosyntactic patterns, gender, formation process) and semantic information (hypernyms, equivalence in another language, definitions and synonyms). In this paper we offer an overall description of the software and an evaluation of its performance, for which we used a linguistics corpus in English and Spanish.

1 Introduction

Terminology-related software has been available for more than sixty years (Hutchins, 1998), first promoted by the Vienna School (Wüster, 1979; Felber, 1984), but later gravitating towards computational linguistics (Sager, 1990; Kageura, 2012). Currently, the field of computer assisted terminology consists of a large variety of tools and methods, not only for term management (Steurs et al., 2015), but also for terminology extraction (Kageura and Umino, 1996; Rigouts Terryn et al., 2022), bilingual terminology alignment (Simões and Almeida, 2008; Filippova et al., 2021) and information extraction (Pearson, 1998; Meyer, 2001), among other related areas.

Despite all the efforts, there is still ample room for improvement not only in each of the individual areas but in the field as a whole. There is, in fact, no tool yet available that can offer an integral solution for all the different problems terminologists face up to when creating terminological databases. In this context, we present Termout[1], a tool for automatising, at least partially, many of those tasks.

[1] http://www.termout.org

The tool we present allows the user to process a specialised corpus and extract a draft macrostructure (a lemma-list) as well as data for the microstructural level, such as grammatical and semantic information. The possibilities of this software are very diverse and there is potential to benefit different professionals, foremost terminologists and lexicographers. Users are able to generate raw material which they can later improve manually by adding or correcting data. If the raw material is of some quality, it is undoubtedly better to build from it than starting from scratch. It is hoped that, with the help of this system, larger databases will be possible, saving time otherwise spent in tedious mechanical tasks.

The current implementation of the software is a web-based prototype that can perform the tasks of corpus processing (file uploading, conversion to plain-text format, language detection, POS-tagging and indexing), terminology extraction (with optional human supervision), information extraction (hypernymy, definitions, equivalence in another language, term variation, etc.) and database management (editing, storage, retrieval and import/export options in HTML, CSV and TBX).

In this paper we focus on the evaluation of the results of the main functions of the software: terminology and information extraction. To this end, we experimented with a linguistics corpus in English and Spanish. As the evaluation shows, in its current state the software can already be useful for terminology processing.

The structure of the paper is as follows. Section 2 offers a brief overview of computational terminology techniques with emphasis in terminology extraction. In Section 3 we present a description of the proposed method. Finally, in Section 4 we discuss about the advantages and disadvantages of the method as well as the challenges ahead.

Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, pages 9–18
Varna, Sep 7, 2023
https://doi.org/10.26615/978-954-452-090-8_002

2 Related work

As mentioned in the introduction, the first efforts in initiating the computational treatment of terminology were by members of the Vienna School, but then the field took a turn towards empiricism and began to import methods from computational linguistics (Sager, 1990; Kageura, 2012). This change was accompanied by the emergence of new schools and theories, since data analysis lead to the admission of previously unrecognised phenomena, such as polysemy and term variation, which are less evident when relying only on introspection (Humbley, 2022).

Automatic terminology extraction (ATE), i.e. the separation of terms from the general vocabulary of a corpus (Kageura and Umino, 1996), was an early and strong force of change in practical terminology. The topic attracted the attention of many researchers and a wide variety of ideas were proposed. In the early years, some systems used statistical measures to detect multi-word terms (Daille, 1994; Frantzi et al., 2000). Others incorporated syntactic knowledge (Justeson and Katz, 1995; Bourigault et al., 1996). Others used statistics to calculate keywordness or weirdness, which means exploiting reference corpora by comparing the frequency of a term in a specialised corpus versus a corpus of general language (Ahmad et al., 1999; Drouin, 2003; Baisa et al., 2017).

The most recent tendency in the literature is the application of machine learning techniques, especially deep neural networks (Hazem et al., 2020; Lang et al., 2021; Rigouts Terryn et al., 2022; Tran et al., 2023). A drawback is however that their complexity makes them difficult to use, to interpret their results and, as Rigouts Terryn et al. (2020) point out, their behaviour is often unpredictable.

Aside from terminology extraction, other relevant subfields must be commented upon. One of those is bilingual terminology alignment using parallel, comparable or unrelated corpora (Simões and Almeida, 2008; Lefever et al., 2009; Aker et al., 2013; Haque et al., 2018; Filippova et al., 2021). Another subfield consists of the application of text mining techniques to obtain information about the terms from the corpus, which can be definitions (Pearson, 1998; Meyer, 2001; Anke et al., 2016); hypernymy relations (Hearst, 1992; Weeds and Weir, 2003; Bordea et al., 2015; Shwartz et al., 2017) and term variants (synonyms) (Ville-Ometz et al., 2007; Cram and Daille, 2016). The work by Wachowiak et al. (2021) is a recent example of a combination of term and term-relation extraction.

The number of relevant and recent publications to the different subareas of computer assisted terminology is on the thousands and still rising. However, the tendency seems to be analytical, i.e., to specialise in the different individual problems. As a consequence, not may proposals exist for the comprehensive solutions needed in practical terminology. There are some terminology extraction services (e.g. OneClick Terms[2] or MultiTerm[3]), but no software exists, commercial or public, that can accompany the user in the different steps of a terminology project. The software Terminus[4] (Cabré and Nazar, 2011) was a first attempt in that direction, but it was not further developed.

The present year 2023 is, of course, one of unprecedented changes in the field of A.I., and it is likely that new proposals for terminology processing will come from that side. In fact, some lexicographers (de Schryver and Joffe, 2023) have already started using ChatGPT (OpenAI, 2023) to automate different lexicographic tasks. Again, the results of neural network algorithms using large language models are promising although unpredictable, as they occasionally "hallucinate"[5].

In this juncture, we still think that there is room for experimentation with alternative methods, especially if they do not entail great complexity, require massive computing power and are not own by large private corporations.

3 Description and evaluation of the prototype

3.1 Overview

The described tool is designed to help the terminologist in every step of a project. The routines for the development of a terminology database are the compilation and processing of a specialised corpus (3.2); terminology extraction (3.3), information extraction (3.4) and database management (3.5). In order to evaluate the different functions we compiled a corpus of research articles from 15 open access scientific journals in English and Spanish in

[2] https://terms.sketchengine.eu/

[3] https://www.trados.com/products/multiterm-desktop/

[4] http://terminus.iula.upf.edu/

[5] According to the technical report, "care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important" (OpenAI, 2023, p. 2).

the field of general linguistics[6]. The sample consists of 3680 PDF files with a total extension of ca. 35 million word tokens.

3.2 Corpus preprocessing

With this tool, a terminology project starts with a corpus, which at the moment must be provided by the user[7]. A specialised corpus (Pavel and Nolet, 2002; Steurs et al., 2015) must cover a single topic or domain, must have some authority in the field and, most importantly, it must be very large. The latter is especially important in our case, as results deteriorate considerably with corpus of less than 200 documents.

The corpus can be uploaded as a ZIP file. It will be uncompressed and each input document will be submitted to the following processes:

Format detection and conversion: The program will guess the type of file (ZIP, TXT, PDF, PS, DOC, DOCX, ODT, HTML, XML, etc.) and convert it to UTF-8 Unix plain text format.

Language detection: It detects the main language of each document and also fragments of text inside that are in a different language. This is based on text similarity measures using samples of text in different languages. The text samples were downloaded from the Wortschatz Project[8] (Goldhahn et al., 2012). The program will only accept text in the supported languages (for now, only English and Spanish).

POS-tagging: Once with the documents separated by language, the corpus is submitted to a POS-tagging procedure. This is done with UDPipe (Straka and Straková, 2017), an external tool.

Indexing: As the program makes intensive use of concordance extraction for various functions, speed is thus critical, and for this a corpus indexing is needed as part of the pre-processing. We developed an indexing method consisting of a table with the positions of each word type in the corpus.

[6] We downloaded papers published in the last 15 years in the following journals: Alfal (ISSN 2079-312X); Anuario de letras (2448-8224); Boletín de Lingüística (0798-9709); Colombian Applied Linguistics Journal (0123-4641); Cuadernos de Lingüística Hispánica (0121-053X); Forma y Función (0120-338X); Íkala (0123-3432); Lenguaje (0120-3479); Letras (0459-1283); Lexis (0254-9239); Lingüística (2079-312X); Literatura y lingüística (0716-5811); Logos (0716-7520); Núcleo (0798-9784); Signos (0718-0934) and RLA (0718-4883).

[7] New functions for automatic corpus compilation are now in development, as explained in Section 4.

[8] https://wortschatz.uni-leipzig.de

3.3 Terminology extraction

As explained in Section 2, terminology extraction is a categorisation problem in which, for every term candidate, a system will produce as a result a score which will lead to the acceptance or rejection of the candidate. In this respect, this system does not depart from traditional approaches, but the method to score the term candidates is original.

The proposed terminology extraction method has a battery of filters arranged in increasing order of computational complexity, finishing in a combination of statistical measures. The initial exclusion rules are computationally inexpensive because they are based on stoplists and morphosyntactic patterns. The core of the method is the later application of a series of statistical measures such as term frequency, dispersion (based on document frequency) and co-occurrence (the analysis of other words sharing the same sentences with the candidate).

The first step of the terminology extraction procedure is the creation of lists of word n-grams (with n defined by the user, ranging from 1 to 5 by default). Each n-gram is treated as a potential term and submitted to the following battery of measures:

Stoplist: This is a set of simple exclusion rules to eliminate $n$grams that begin or end with a member of a list of function words (grammemes such as prepositions, articles, conjunctions, some adverbs, etc.). These function words are however admitted inside the candidate, as it may occur with some $n$grams with $n > 2$ (e.g., the linguistics term *part of speech*).

Morphosyntactic patterns: In this project we have opted to limit the number of term candidates to those which can be parsed as noun phrases. Candidates including other grammatical categories or patterns, such as verbs or adverbs, are excluded[9].

Term frequency: For any candidate x that survives the previous filters, we calculate its term frequency: $f(x)$. This measure might not be useful in isolation or while analysing a single document, as most terms in a text will be hapax legomena or dis legomena, but it can be a useful indicator if used in conjunction with other statistical measures and when analysing a large collection of specialised documents.

Dispersion: This measure is defined as a combination of term frequency and document frequency,

[9] This is certainly a limitation for users interested in specialised predicates, but these units may require a different methodology.

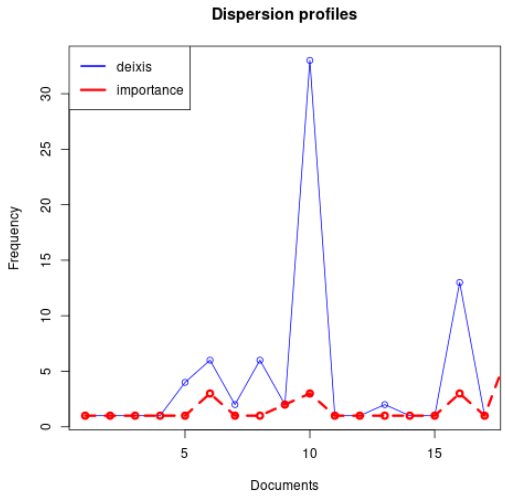

Figure 1: Dispersion of units *deixis*, a linguistics term, and *importance*, a non-term, in a sample of documents

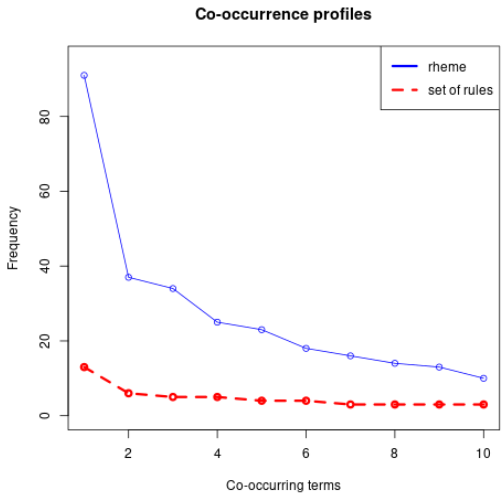

Figure 2: Co-occurrence profile of units *rheme*, a linguistics term, and *set of rules*, a non-term

$df(x)$, i.e. the number of documents in which a term occurs. When one observes how a candidate is distributed within a corpus, useful patterns begin to emerge, which can be exploited to make a prediction. Drawing inspiration from Spärck Jones (1972), we used coefficient (1) to measure the dispersion of a candidate. It can be described as a simplified derivative of tf-idf, less costly to compute. The variable $h(x)$ in (2) is the number of documents in a collection D in which term x has frequency 1.

$$d(x) = 1 - \frac{h(x)}{df(x)} \tag{1}$$

$$h(x) = \sum_{i=1}^{|D|} \begin{cases} 1 & f(x, D_i) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Figure 1 shows the dispersion of two units in the corpus. The blue, continuous line corresponds to the term *deixis*, a genuine linguistics term, and the red, dashed line to *importance*, a non-term. Rough curves with sharp spikes appear to be associated with higher information, because they show that when a term occurs in a document, it is also likely that it will be used more than once. On the contrary, smoother curves mean that the expression is often used once per document, a pattern associated with non-terminological units.

Co-occurrence: As shown in previous work dating back from Harris (1954) and Firth (1957), one can know about a word by looking at the company it keeps. In this case, this means that terminological units are often revealed by their co-occurrence patterns, and this can be used as a robust predictor of the specialised value of a candidate. Terms show a tendency to co-occur with a reduced number of other terms which conform their semantic field. For instance, Figure 2 shows the case of a pair of units, *rheme*, a linguistics term (blue, continuous line), and *set of rules*, a sequence of words with no terminological value (red, dashed line). As expected, the term shows a tendency to appear in the same sentences with other related terms such as *theme, clause, progression, sentence*, etc. The other one, however, does not show a strong association with any other word despite being 20 times more frequent than the first. We used a co-occurrence measure (3) to exploit this phenomenon.

$$c(x) = \frac{\log_2 \sum_{i=1}^{k} R_{x,i}}{\log_2 f(x)} \tag{3}$$

In Equation 3, x is a term candidate; R_x the set of (single) words co-occurring with x; $f(x)$ is, again, the frequency of x and $R_{x,i}$ the frequency of the ith most frequent co-occurring word in the contexts of occurrence of x. The value k is an arbitrary parameter[10].

Extras: With variable $e(x)$ we denote an additional value for x when it is found in the title of bibliographic references in the corpus and/or when definitional patterns are found in the immediate vicinity of a term (the program includes a module

[10] In our experiments, $k = 20$. Larger ks mean longer processing times, but not necessarily better results. Users will have to experiment and adjust this parameter themselves to find the best compromise.

for the extraction of definitions from the corpus, explained later in Subsection 3.4). Appearing in titles and being defined are both taken as indicators of the significance of a term.

Final score: The above mentioned statistical measures, frequency, dispersion, co-occurrence and extras, defined as set A (4), are combined to produce a final score $s(x)$ (5). A threshold for this score is defined by the user.

$$A = \{\sqrt{f(x)}, d(x), c(x), e(x)\} \tag{4}$$

$$s(x) = \prod_{i=1}^{|A|}(1 + A_i) \tag{5}$$

After the calculations, the system also classifies the term candidates by language, which is done by inspecting the language of their contexts of occurrence, using the same mechanism described in 3.2. It also displays tables of rejected candidates that scored close to the cutting threshold, so users can manually rescue eventual false negatives. There is also the possibility of eliminating all candidates that include any arbitrary component.

As an alternative, the program also offers the user the possibility of uploading a list of terms to be used as examples. In this way, users may obtain more refined results, as the program will promote those candidates that tend to co-occur with those presented as examples. In particular, this last function may benefit those users who need terms of a very specific topic but only have a general corpus of the discipline (e.g., those interested only in phonology terms but having a general linguistics corpus or interested in PTSD terms but having a general psychiatry corpus, etc.).

For the purpose of evaluation, we extracted terms from the corpus restricting the minimum frequency to 10, a conservative parameter that favours precision over recall. This way we obtained a total of 1882 term candidates, automatically separated by language: 618 in English and 1264 in Spanish. The separation by language was almost perfect (we found only four errors). Regarding the term/non term separation, there were 104 false positives in English and 190 in Spanish. That makes a total precision of 84%.

Some examples of correct terms in English are the following: *argument structure; bilingualism; evidentiality; universal grammar*, etc. Among the errors we find some proper nouns (*Alarcos Llorach; Berkeley Linguistics Society; Prentice Hall*; etc.), some subject-verb pairs (*students work; teachers need*, etc.) among other cases (*assistant professor; Chinese student*, etc.). Figure 3 shows a screenshot of the program's interface with a fragment of the list of extracted candidates.

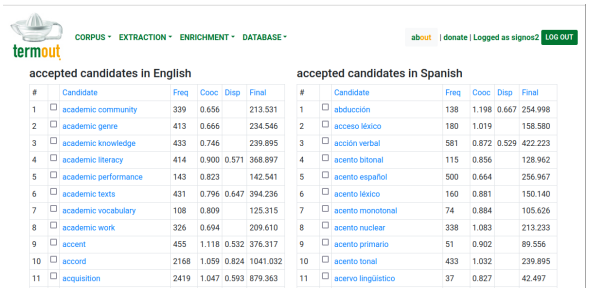

Figure 3: A screenshot of the results of the terminology extraction function

3.4 Information extraction

Once a list a terms has been obtained and, ideally, manually revised, the program then offers a battery of functions to populate the terminology database with a number of fields. Aside from fields such as inflection, grammatical gender and part of speech, the following functions provide further database enrichment:

Semantic categorisation: This function produces full hypernymy chains for each extracted term in each language, with progressive levels of abstraction and a graphic depiction of the conceptual hierarchies. The algorithm that produces this result combines co-occurrence statistics and morphosyntactic patterns (Nazar et al., 2021). Co-occurrence statistics tend to be asymmetric in the case of hyponym-hypernym pairs, in such a way that hyponyms show a tendency to co-occur with hypernyms in a non-reciprocal relation. This is combined with rules of morphosyntactic patterns *à la* Hearst (1992), which are used to triangulate information and reinforce a suspicion of hypernymy between pairs of terms. The main difference with respect to previous research using such type of patterns is that our algorithm only uses them to gather information about one term at a time. That is, it first collects all the contexts of occurrence of a given term and then computes statistics on the number of patterns found among those contexts.

An example of a correct result for the case of the term *articulatory phonetics* is the following hypernymy chain: *phonetics* → *linguistics* → *social science* → *science* → *study* → *abstract entity*

→ *entity*. After the evaluation, we found that in 99% of the cases there was a result and 64% of those were correct. The main cause of errors in the assignment of hypernym chains were cases of polysemy, in particular regular polysemy. An example of this type of error is the following: *narrative text → document → artefact → physical object*. The problem here is that the term *narrative text* in our corpus actually refers to the abstract content of the text, not the physical object.

Semantic clustering: This function produces clusters of terms that are semantically related. Here, a semantic relation is operationalised as co-occurrence associations computed as in Subsection 3.3, and the clustering is done with a co-occurrence-graph algorithm[11]. Specialised terms have a semantic field consisting of a set of other terms. Consequently, terms sharing a similar co-occurrence profile are placed in the same cluster.

For the evaluation of this function, a total of 65 clusters were produced, of which 83% presented internal consistency. For instance, one cluster presents discourse-related terms, another presents corpus linguistics terms, and so on. Figure 4 shows a fragment of a cluster the system creates with 35 terms in this case related to phonology in Spanish. For the visualisation of these graphs we used the GraphViz library (Gansner and North, 2000).

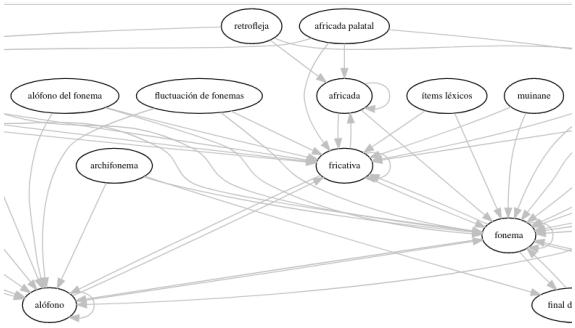

Figure 4: A fragment of a co-occurrence-graph cluster for phonology terms in Spanish

Spurious clusters were invariably cases with weakly interconnected nodes, and this could be exploited to further develop the method.

Definitions: With this function, users can obtain definitions of the terms from the corpus. For this we manually compiled a large list of definitional patterns in English and Spanish. Similarly as with the extraction of hypernymy chains, we first scan all the contexts of occurrence of a given term and extract the concordances that match a definitional pattern. These concordances are then sorted according the type of pattern found and its proximity to the analysed term.

For the evaluation we considered a correct result one in which for a term at least one context (out of max. 5) provides enough data for a definition. Consider, for instance, the following result for the case of *language planning*: "language planning refers to deliberate efforts to influence the behaviour of others with respect to the acquisition, structure, or functional allocation of language"[12]). Considering only the definitions extracted for the genuine terms, we found that 53% of the proposed definitions were acceptable; 12% of the cases produced no result and the rest were errors.

Bilingual alignment: Users can obtain a bilingual alignment of the extracted terms. This is achieved by applying a combination of dispersion and co-occurrence association measures, including also an orthographic similarity coefficient for the cognates. To calculate dispersion and co-occurrence we followed a similar principle as in Subsection 3.3. In the case of co-occurrence, the only difference is that in this case the interest is to find the intensity of the association between two terms i and j, for which we used coefficient 6. As in Subsection 3.3, this measures co-occurrence in the same sentences, irrespective of the order and distance between the two terms. In the case of dispersion and orthographic similarity, we used coefficient 7, in one case to measure how many documents two terms i and j have in common and in the other case how many character bigrams (sequences of two letters) they share.

$$coo(i,j) = \frac{f(i,j)}{\sqrt{f(i)}.\sqrt{f(j)}} \quad (6)$$

$$sim(i,j) = \frac{2|i \cap j|}{|i| + |j|} \quad (7)$$

Regarding the evaluation, from the sample of extracted terms we obtained 466 alignments (75% of the 618 English terms). Among these, we found a total of 108 errors (ca. 77% precision). Some example of correct alignments are the following: *academic genre = género académico; action verbs = verbos de acción; phonological system = sistema*

[11] We developed a clustering method based on co-occurrence graphs to avoid the quadratic complexity of to classical agglomerative clustering algorithms.

[12] The fragment is attributed to Cooper, R. L. (1989). Language Planning and Social Change. Cambridge University Press.

fonológico, etc. Typical errors are alignments of terms that are semantically related but not equivalent (e.g. *conceptual metaphor* ≠ *dominio fuente; critical language awareness* ≠ *conciencia crítica; foreign language learners* ≠ *lengua extranjera*). Figure 5 shows a moment of the bilingual alignment process.

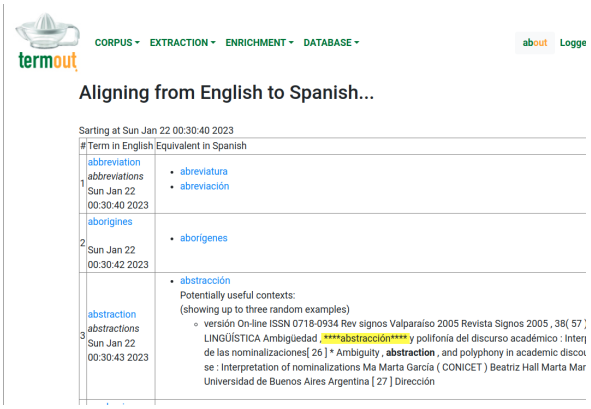

Figure 5: Examples of bilingual alignment

Term variants: The final function in this Section consists of extracting term variants, i.e. terms in the same language which have different forms but the same meaning. In our case, the proposal to address this problem is based on the bilingual alignment conducted in the previous function, and it follows a simple intuition: two terms in the same language i and j can be considered term variants if they consistently share the same equivalences in the other language. For instance, *analyser* and *parser* are considered specialised synonyms because they share the same equivalence in Spanish (*analizador*), and the same occurs for other pairs such as *semantic field* ∼ *semantic space; coefficient* ∼ *ratio; discourse* ∼ *speech; poll* ∼ *survey; phrase* ∼ *sentence; core* ∼ *nucleus; meaning* ∼ *significance*; etc. Examples in Spanish are similar: *alfabetismo crítico* ∼ *alfabetización crítica; debate* ∼ *discusión; aplicación* ∼ *implementación*, and so on.

From the dataset of 1884 terms, a total of 105 pairs or groups of variant terms were obtained. From those, 60 cases we confirmed to be genuine synonyms (57%). Typical errors consist of pairs of words that are semantically related but are not synonyms (e.g. *learning* ≠ *pupil; apprenticeship* ≠ *learning; classroom* ≠ *teaching*, etc.).

Task	Precision
Term extraction	84%
Semantic categorisation	64%
Semantic clustering	83%
Definition extraction	53%
Bilingual alignment	77%
Term variant extraction	57%

Table 1: Summary of evaluation figures per task

3.5 Term management

In addition to the term extraction and information extraction functions, the tool also offers the possibility of manually editing the database in order to correct false information, to complete term records with missing data, or to delete and/or create new term records.

The system also offers the standard functions for querying the database with a search form that allows to retrieve information by any field or a combination of fields. As usual in this type of systems, a user may, for instance, retrieve all the terms that have a certain component (word or segment of word), or a certain term as a hypernym, or as equivalent, as synonym, etc.

When satisfied with the result, users can export the database in CSV, TBX or HTML formats. They can also import databases in CSV or in an industry standard such as TBX (Melby, 2015). The latter can be convenient for users already having a terminology database that needs to be completed, expanded or edited.

3.6 Summary of evaluation figures

Table 1 offers a summary of the evaluation figures obtained in this section, indicated in all cases as precision rates. Evaluation of recall for all functions would be harder to estimate in most cases. It would be possible to approximate a figure of recall in the case of term extraction by manually annotating some documents. But in the case of other functions it would be more challenging. Consider, for instance, the case of semantic clustering or bilingual alignment. It is difficult to determine how many clusters or alignments are in the corpus. We therefore leave the evaluation of recall for a future paper.

Here we also have to mention that some researchers have proposed annotated corpora to evaluate term extraction systems. Among them, we find the ACL RD-TEC 2.0 (QasemiZadeh and Schu-

mann, 2016) and the TermEval 2020 (Rigouts Terryn et al., 2020). We did not use these materials, however, for different reasons. In the first case, because it is intended for systems that operate on the sentence-level, and thus they only include small fragments of text (abstracts). In contrast, our system is designed to work with natural, integral texts. In the second case, because we seem to have a different definition of what constitutes a term. As already mentioned, we only include sequences that can be parsed as noun phrases. We exclude predicate-argument structures as terminological units (e.g., for us, to *combat corruption* or *fight corruption* are not multi-word terms, but a combination of a verb and its complement).

4 Conclusions and future work

In this paper we presented a software for terminology processing that integrates a variety of tools for the creation of a terminology database, and we reported on a series of tests to evaluate its performance. As a first take after our assessment, we believe that despite some limitations, it could be useful for professional lexicographers and terminologists. In addition, we see also a possible application of the tool in the teaching of terminology, as students may use it to learn from practical experience in term database creation.

As pointed out in the introduction, there is today no single software product that can provide solutions for the different tasks involved in term-database creation. The software products now available for terminology and lexicography processing are too time-consuming. We believe, thus, that a tool such as the one we propose is useful not only for the convenience of automation but also because a technical glossary should be created using specialised corpora as input. Another advantage of the proposal is that it is based on simple algorithms, compared to those using neural networks. Dispersion and co-occurrence statistics can be performed in relatively cheap hardware, although it is still necessary to improve computational efficiency to reduce processing times.

The current implementation of Termout is freely available and has no restrictions of any kind. This might become a problem if the number of users increases significantly, since we lack the necessary infrastructure (manpower, servers, etc.). If confronted with such scenario, we would be forced to explore alternatives for sustainability.

It is also worth pointing out that the system uses no information external to the user's own corpus. We are, however, considering the possibility of changing this in future versions, in order to include the optional use of Wikipedia or other external knowledge sources.

Another point to mention is that, currently, the system only operates with English and Spanish text. However, the method is fundamentally based on statistical and language-agnostic algorithms, apart from the POS-tagger and the lexical patterns used in the extraction of hypernyms and definitions. We are, indeed, already attempting to adapt the system to different European languages.

We are also exploring new ways to let the users acquire corpora, and this function will soon be available. One alternative is to provide the program with a URL that contains links to other documents, and let the program decide which links are relevant. The other possibility is to upload a single document that the program will use to automatically extract, using text-similarity measures, a subset of similar documents from a larger general corpus such as the TenTen corpora collection (Jakubíček et al., 2013). This will offer the user the possibility of having the most laborious tasks of a terminology project fully automated.

Acknowledgments

This research received funding from a grant by the Chilean Government (Proyecto Fondecyt Regular 1231594, directed by Irene Renau). We would also like to thank the reviewers for their work.

References

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Ahmet Aker, Monica Paramita, and Rob Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411, Sofia, Bulgaria. Association for Computational Linguistics.

Luis Espinosa Anke, Roberto Carlini, Horacio Saggion, and Francesco Ronzano. 2016. Defext: A semi supervised definition extraction tool. *CoRR*, abs/1606.02514.

Vít Baisa, Jan Michelfeit, and Ondřej Matuška. 2017. Simplifying terminology extraction: Oneclick terms. In *Proceedings of the 9th International Corpus Linguistics Conference*.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910, Denver, Colorado. Association for Computational Linguistics.

Didier Bourigault, Isabelle Gonzalez-Mullier, and Cécile Gros. 1996. Lexter, a natural language processing tool for terminology extraction. In *Proceedings of the 7th EURALEX International Congress*, pages 771–779, Göteborg, Sweden. Novum Grafiska AB.

Teresa Cabré and Rogelio Nazar. 2011. Terminus: a workstation for terminology and corpus management. In *Proceedings TOTh 2011*, pages 73–74. Presses Universitaires Savoie Mont Blanc.

Damien Cram and Béatrice Daille. 2016. Termsuite: Terminology extraction with term variant detection. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*, pages 13–18.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Ph.D. thesis. Paris 7.

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

Helmut Felber. 1984. *Terminology manual*. United Nations Educational, Scientific and Cultural Organization : International Information Centre for Terminology, Paris.

Darya Filippova, Burcu Can, and Gloria Corpas Pastor. 2021. Bilingual terminology extraction using neural word embeddings on comparable corpora. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 58–64.

John Firth. 1957. A synopsis of linguistic theory, 1930-55. In *Studies in Linguistic Analysis*, pages 1–31. Blackwell, Oxford.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Emden R. Gansner and Stephen C. North. 2000. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.*, 30(11):1203–1233.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Rejwanul Haque, Sergio Penkale, and Andy Way. 2018. Termfinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52(2):365–400.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille. 2020. TermEval 2020: TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100, Marseille, France. European Language Resources Association.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

John Humbley. 2022. The reception of Wüster's general theory of terminology. In Pamela Faber and Marie-Claude L'Homme, editors, *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialized knowledge*, pages 15–36. John Benjamins, Amsterdam.

John Hutchins. 1998. The origins of the translator's workstation. *Machine Translation*, 13(4):287–307.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127, Lancaster.

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

Kyo Kageura. 2012. *The Quantitative Analysis of the Dynamics and Structure of Terminologies*. John Benjamins.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(1):259–289.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620. Association for Computational Linguistics.

Els Lefever, Lieve Macken, and Veronique Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 496–504.

Alan Melby. 2015. TBX: A terminology exchange format for the translation and localization industry. In Hendrik Kockaert et al., editor, *Handbook of Terminology*, pages 393–424. John Benjamins, Amsterdam.

Ingrid Meyer. 2001. Extracting a knowledge-rich contexts for terminography: A conceptual and methodological framework. In D. Bourigault, C. Jaquemin, and M.C. L'Homme, editors, *Recent Advances in Computational Terminology*, pages 279–302. John Benjamins, Amsterdam.

Rogelio Nazar, Antonio Balvet, Gabriela Ferraro, Rafael Marín, and Irene Renau. 2021. Pruning and repopulating a lexical taxonomy: experiments in spanish, english and french. *Journal of Intelligent Systems*, 30(1):376–394.

OpenAI. 2023. Gpt-4 technical report.

Silvia Pavel and Diane Nolet. 2002. *Manual de Terminología*. Translation Bureau. Public Works and Government Services, Québec.

Jennifer Pearson. 1998. *Terms in Context*. John Benjamins, Amsterdam.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

Ayla Rigouts Terryn, Veronique Hoste, and Els Lefever. 2022. D-terminer : online demo for monolingual and bilingual automatic term extraction. In *Proceedings of the Workshop on Terminology in the 21st century*, pages 33–40. European Language Resources Association (ELRA).

Juan C. Sager. 1990. *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam.

Gilles-Maurice de Schryver and David Joffe. 2023. The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks. https://www.youtube.com/watch?v=mEorw0yefAs. 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75. Association for Computational Linguistics.

Alberto Simões and José João Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, (41):281–288.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Frieda Steurs, Ken De Wachter, and Evy De Malsche. 2015. Terminology tools. In Hendrik J. and Kockaert et al., editors, *Handbooks of Linguistics and Communication Science*, pages 222–249. John Benjamins, Amsterdam.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. 2023. The recent advances in automatic term extraction: A survey.

Fabienne Ville-Ometz, Jean Royauté, and Alain Zasadzinski. 2007. Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 13(1):35–59.

Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann. 2021. Towards Learning Terminological Concept Systems from Multilingual Natural Language Text. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 22:1–22:18, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Eugen Wüster. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Springer, Wien.