Author's final version

Periñán-Pascual, Carlos and Ricardo Mairal Usón (2018) "A framework of analysis for the evaluation of automatic term extractors". VIAL - Vigo International Journal of Applied Linguistics 15, pp. 105-125.

A framework of analysis for the evaluation of automatic term extractors

Carlos Periñán-Pascual

Universitat Politècnica de València

Paranimf, 1

46730 Gandia (Valencia), Spain

jopepas3@upv.es

Ricardo Mairal-Usón

UNED

Senda del Rey, 7

28040 Madrid, Spain

rmairal@flog.uned.es

Abstract

Departing from previous research on automatic term extraction, the primary aim of this paper is to propose a more robust and consistent framework of analysis for the comparative evaluation of term extractors. Within the different views for software quality outlined in ISO standards, our proposal focuses on the criterion of external quality and in particular on the characteristics of functionality, usuability and efficiency together with the subcharacteristics of suitability, precision, operability and time behavior. The evaluation phase is completed by comparing four online open-access

automatic term extractors: TermoStat, GaleXtract, BioTex and DEXTER. This latter resource forms part of the virtual functional laboratory for natural language processing (FUNK Lab) developed by our research group. Furthermore, the results obtained from the comparative analysis are discussed.

Resumen

A partir de investigaciones anteriores sobre extracción automática de términos, este artículo tiene como objetivo fundamental desarrollar una propuesta más consistente y robusta para la evaluación de los extractores terminológicos. De los criterios expuestos en los estándares ISO para la evaluación de la calidad del software, nos centramos en el criterio de calidad externa y, más concretamente, en las características de funcionalidad, usabilidad y eficiencia así como en las subcaracterísticas de adecuación, precisión, operabilidad y comportamiento de tiempos. Aplicamos este marco de análisis para evaluar los siguientes extractores automáticos de términos que son de acceso abierto: TermoStat, GaleXtract, BioTex y DEXTER. Este último recurso forma parte del laboratorio virtual para el procesamiento computacional del lenguaje desde un paradigma funcional (FUNK Lab) desarrollado por nuestro equipo de investigación. Finalmente, presentamos los resultados que hemos obtenido para cada uno de los indicadores.

1. *Introduction*

One of the key areas of interest in our latest research has been the development of a virtual computational laboratory based on functionally-oriented linguistic premises. In particular, we have developed a number of computational resources that are inspired in the analytical tools of Role and Reference Grammar (RRG), a functional linguistic

theory (cf. Van Valin, 2005; Mairal, Guerrero and González 2012, etc.). One of the strengths of RRG is unequivocally its typological adequacy, that is, its potential to articulate analytical tools that are valid in a multilingual scenario, a feature which makes it particular attractive to be implemented computationally.[1] Within this context, we began to work on the computational adequacy of RRG and developed an inventory of different natural language processing resources and tools. At this stage, the following applications have thus far been developed:

a) Navigator: this tool allows the user to retrieve data from the lexical entries in the English Lexicon (e.g. morphosyntactic, pragmatic and collocational information) and from the conceptual entries in the Core Ontology (e.g. thematic frame, meaning postulate etc), as developed within the framework of the FunGramKB Project.

b) **A**utomatically **R**epresenting **TE**xt **M**eaning via an **I**nterlingua-based **S**ystem (ARTEMIS): this computational resource is currently a proof-of-concept laboratory which allows the automatic generation of a conceptual logical structure (CLS), that is, a fully specified semantic representation of an input text, on the basis of a reduced sample of sentences (cf. Periñán, 2013; Cortés and Mairal, 2016).

c) RONDA (**R**ec**O**gnizi**N**g **D**omains with I**A**TE): this tool is used to categorize a text or a collection of documents in different specialized domains as specified in the IATE database.

d) **CA**tegory- and **S**entiment-based **P**roblem Find**ER** (CASPER): this resource analyses micro-texts (e.g. tweets) for the automatic detection of user-defined problems by following a symbolic approach to topic categorization and sentiment analysis

[1]Indeed, a few researchers have recently devoted their work to applying RRG in different computational models, e.g. Diedrichsen, (2013), Guest, (2009), Nolan and Periñán-Pascual (2014), Nolan and Salem (2011), Salem et al. (2008), or Van Valin and Mairal (2014).

e) **DA**ta **MI**ning **EN**countered (DAMIEN): it is a workbench that allows researchers to do text analytics by integrating corpus-based processing with statistical analysis and machine-learning models for data mining tasks.

f) **D**iscovering and **EX**tracting **TER**minology (DEXTER): this tool has been developed as an online multilingual workbench which is provided with a suite of tools for (a) the compilation and management of small- and medium-sized corpora, (b) the indexation and retrieval of documents, (c) the elaboration of queries by means of regular expressions, (d) the exploration of the corpus, and (e) the identification and extraction of term candidates (i.e. unigrams, bigrams and trigrams) (Periñán-Pascual 2015).[2]

This paper is concerned with DEXTER and in particular with the potential of this computational resource for automatic term extraction (ATE) from Spanish texts. In so doing, DEXTER is evaluated by comparing it to the following automatic term extractors: TermoStat (Drouin 2003),[3] GaleXtract (Barcala, Domínguez-Noya, Gamallo, López, Moscoso, Rojo, Santalla and Sotelo 2007),[4] and BioTex (Lossio-Ventura, Jonquet, Roche and Teisseire 2014a)[5].

The organization of this paper goes as follows: Section 2 provides a critical description of the frameworks used for the evaluation of comparative extractors; Section 3 offers a description of the characteristics and subcharacteristics of our framework of analysis for the comparative evaluation of ATE software; Section 4 discusses the results obtained for each of the computational tools in terms of their suitability, precision, operability and time behavior.

[2] http://www.fungramkb.com/nlp.aspx
[3] http://termostat.ling.umontreal.ca

[4] http://gramatica.usc.es/~gamallo/php/gale-extra/gale-extra2.1/index.php

[5] http://tubo.lirmm.fr/biotex/

2. *Evaluation of term extractors*

To the best of our knowledge, Sauron (2002) and Zielinski and Safar (2005) can be considered as the two most outstanding studies whose primarily aim was to develop a comprehensive framework for the comparative evaluation of term extractors, going further than the testing of the metric performance.

Sauron (2002) applied an evaluation methodology based on ISO standards and the work of the EAGLES Evaluation Working Group (1999). Her intention was "the development of a standardised methodology for the evaluation of such tools" (Sauron 2002: 1), where she examined four characteristics (i.e. functionality, usability, reliability and efficiency), which were broken down into seven subcharacteristics (e.g. accuracy, interoperability, learnability, recoverability, suitability, time response and understandability). However, both the attributes to evaluate the systems and the scoring rules to rate every attribute should have been further refined. On the one hand, most of the attributes were inaccurately formulated. For example, Sauron (2002: 7) stated that if the vocabulary used to describe the different functions of the system is "badly incoherent and inconsistent" throughout the documentation, then the score is 0, but if there is "one or more inconsistencies in the terminology used", then the score is 2.5. Hence, it follows from the above that "consistency with the documentation language" is not evaluated as a gradual attribute but as a polar one; this surprisingly implies that, for example, two or twenty inconsistencies make any documentation equally inconsistent. Sauron (2002: 11) also suggested that if the software is user-friendly, then the score is 5; but if it is "not very user friendly", then the score is 2.5. Here *very* is a vague word, and, because this adverb is subject to different interpretations, the attribute "user-friendliness" cannot be objectively measured with this wording. On the other hand,

every attribute is rated as good, acceptable or unacceptable, where every rating is assigned a particular score depending on the attribute that has been selected. In terms of good research practice, a five-point scale would have been more appropriate, since changing the number of response categories from three to five increases reliability in Likert-type rating scales (Preston and Colman 2000; Lee and Paek 2014).

Zielinski and Safar (2005) presented an online survey of term extractors in which over 400 professional translators, terminologists and interpreters took part. This survey, which was divided into sections such as Personal Information, Working Environment, Translation, Terminology Management and Terminology Extraction, was intended to "(…) investigate the relationship between research and practice in the area of terminology extraction and evaluate if there is any need to reconcile both" (Zielinski and Safar 2005: 1).

With these studies in mind, we designed what we believe to be a more robust and consistent framework of analysis for the comparative evaluation of term extractors. Following the ISO evaluation framework, which is applicable to any kind of software, Sauron's research (2002) marked the starting point in the selection of some of the quantifiable attributes of the new evaluation model. Moreover, Zielinski and Safar's study (2005) helped to provide new insights into the way to interpret the results of the evaluation on the basis of the functionalities required to fit the different user profiles. Finally, ISO/IEC 9126-1 (2001), which distinguished three different views of software quality (i.e. internal quality, external quality, and quality in use),[6] helped us determine the perspective of this research, which is only concerned with the external quality of the software, since this is the most relevant view from which language researchers can

[6] Although ISO/IEC 25010 (2011) was released to replace ISO/IEC 9126 (2001), the latter is still the most commonly used quality standard.

decide what tool best suits their needs; more particularly, this paper focuses on three characteristics (i.e. functionality, usability and efficiency) that were broken down into four subcharacteristics: suitability, precision, operability and time behaviour. In other words, our evaluation is based on those parts of the software the user gets directly into contact with (i.e. black-box evaluation). This evaluation phase is carried out by comparing the four online open-access term extractors mentioned above.

3. *Developing the framework of analysis*

3.1. Selecting characteristics and subcharacteristics

According to ISO/IEC 9126-1 (2001), there are two main elements in the external quality model: characteristics that are refined into subcharacteristics. First, we determined which subcharacteristics are the most relevant for each of the characteristics chosen to evaluate the term extractors, as well as determining the weight of every subcharacteristic, in such a way that:

(1)

$$\sum_{i=1}^{m} w_{s'_i} = 1, \text{where } s' \in c'$$

where *c'* represents the characteristic, s'_i is a subcharacteristic, *w* is the corresponding weight, and *m* is the number of subcharacteristics of a given characteristic. The remainder of this section gives a brief account of the three characteristics and four subcharacteristics examined in this research.

On the one hand, the characteristic of functionality is defined as "the capability of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions" (ISO/IEC 9126-1 2001: 7). We are only concerned with two subcharacteristics: suitability and accuracy. Suitability is defined as

"the capability of the software product to provide an appropriate set of functions for specified tasks and user objectives" (ISO/IEC 9126-1 2001: 8), and accuracy is defined as "the capability of the software product to provide the right or agreed results or effects with the needed degree of precision" (ISO/IEC 9126-1 2001: 8). "Accuracy" and "precision" actually refer to different evaluation metrics; therefore, to avoid any misunderstanding, we employ only "precision" to refer to positive predictive values:

(2)

$$\text{precision} = \frac{\text{true positives}}{\text{extracted candidates (true positives+false positives)}}$$

On the other hand, the characteristic of usability is defined as "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions" (ISO/IEC 9126-1 2001: 9), where the most relevant subcharacteristic of term extractors is operability, which is defined as "the capability of the software product to enable the user to operate and control it" (ISO/IEC 9126-1 2001: 9). Finally, the characteristic of efficiency is defined as "the capability of the software product to provide appropriate performance, relative to the amount of resources used, under stated conditions" (ISO/IEC 9126-1 2001: 10), where time behaviour is the most outstanding subcharacteristic, i.e. "the capability of the software product to provide appropriate response and processing times and throughout rates when performing its function, under stated conditions" (ISO/IEC 9126-1 2001: 10).

3.2. Identifying attributes and features

On the basis of the main software capabilities required to support terminology and terminography research, we compiled a list of significant attributes for every

subcharacteristic, where each attribute was in turn analyzed as a set of features, in such a way that:

(3)

$$\sum_{j=1}^{n} w_{f_j} = 1, \text{where } f' \in s'$$

where *f'* is a feature and *n* is the number of features for a given subcharacteristic. Every feature took the form of an item in the survey, consisting of a question, a set of response options, and a scoring scheme (or measurement method). Appendix A shows the twenty questions that were created from the sixteen attributes derived from the four subcharacteristics. It should be highlighted that these questions resulted from the analysis of Sauron (2002) and Zielinski and Safar (2005) as representing those issues that are considered relevant for most of the users of this type of tools. This inventory of questions is by no means intended to be exhaustive, but tries to illustrate relevant features of software aimed to support terminology and terminography research, e.g. the construction of specialized glossaries. In fact, adding new questions to the current survey would not invalidate this framework of analysis, which would actually help new questions be organized more adequately.

3.3. Calculating the weight of subcharacteristics and features

Finally, we determined the maximum weight of each main component of the analysis (i.e. features and subcharacteristics), where:

(4)

$$\sum_{i=1}^{m} \left(w_{s'_i} \sum_{j=1}^{n} w_{f_j} \right) = 1$$

Attributes did not take part in the weighting procedure because they only played an organizational role. Therefore, we obtained (a) the weight of each subcharacteristic of a

given characteristic and (b) the weight of each feature of a given subcharacteristic, approaching these two tasks as Multiple-Criteria Decision-Making (MCDM) problems. In a nutshell, MCDM guides the model to select the best weight for a given choice by taking into account all the available alternatives. In the remainder of this section, we describe the main concepts of one of the most widely used methods in MCDM, i.e. the classical Analytic Hierarchy Process (AHP), which was employed in this research.[7] To illustrate, this method is described with the task (b).

For any subcharacteristic with n features, the process started by comparing the n features pairwise. The ratio scale displayed in Table 1 was used to compare the importance weight between attributes. For the estimation of the relative importance of subcharacteristics and features, the judgments of three expert terminologists were taken into account.

Value of f_{ij}	Interpretation
1	i and j are equally important
3	i is slightly more important than j
5	i is more important than j
7	i is strongly more important than j
9	i is absolutely more important than j

Table 1. Ratio scale in the AHP.

The comparison of each feature i with each feature j yielded the values f_{ij}, which were placed in a square matrix of dimension n called the pairwise-comparison matrix, i.e. F = (f_{ij}), which is positive and reciprocal. Thus, a matrix such as the following was obtained: (5)

[7] A thorough description of this method can be found in Saaty (1977, 1980). Moreover, the state of the art in the main types of MCDM methods is described in Tzeng and Huang (2011).

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{pmatrix}$$

where $f_{ij} = \frac{1}{f_{ji}}$ and $f_{ii} = 1$. Another property of F is that it satisfies the condition $f_{jk} = \frac{f_{ik}}{f_{ij}}$. In the next step, the consistency of the pairwise-comparison matrix was verified, i.e. for all i and j, $f_{ij} = \frac{w_i}{w_j}$. Some errors might arise by the subjective perception of judgments, so the eigenvector method was introduced to estimate the weights when errors in judgment occurred. In particular, having a vector ω of order n such that $F\omega = \lambda\omega$, ω is said to be an eigenvector and λ is an eigenvalue. With a positive reciprocal matrix such as F, which involves human judgments, the dominant eigenvalue (λ_{max}) is equal to n if and only if F is a consistent matrix, that is, when the judgments are consistent. However, λ_{max} is always greater than n when the judgments are inconsistent to a greater or lesser degree. Thus, $\lambda_{max} - n$ provided a useful measure of the degree of inconsistency of the matrix. By normalizing this measure by the size of the matrix, the consistency index (CI) was obtained:

(6)

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

Moreover, the consistency ratio (CR) was calculated as follows, where RI refers to the random index:

(7)

$$CR = \frac{CI}{RI}$$

RI was computed for each size of matrix *n* by generating randomly-filled reciprocal matrices and their mean CI value. These RI values are displayed in Table 2.

n	RI
2	0
3	0.58
4	0.90

n	RI
5	1.12
6	1.24
7	1.32

n	RI
8	1.41
9	1.45

Table 2. The RI for the AHP.

Thus, the CR provided a way of measuring how many errors were created with the judgments. If the CR was below 0.1, then the number of errors was fairly small, and the final estimate was accepted. However, larger values of the CR required revising the judgments in order to reduce the inconsistencies. As a result, Appendix B shows the AHP-based scoring scheme of every item in the survey.

It should be recalled that the judgments on the importance of subcharacteristics and features have to be made with a target task in mind. In this research, the task was the construction of domain-specific glossaries. If the task had become different, it would have been necessary to recalculate the weight of the components of the analysis with respect to their relative importance in the new task. However, the whole methodology would have remained the same.

4. ***Survey: results and discussion***

This framework of analysis was used for the comparative evaluation of TermoStat, GaleXtract, BioTex and DEXTER, which was conducted with the web browser Google Chrome 56.0.2924.87 installed in a Windows Vista laptop computer, Intel Core i7 CPU M 640 at 2.80GHz (4 processors). The Internet-connection speed was about 50Mbps

downstream and 2Mbps upstream. Table 3 shows the scores for all the subcharacteristics.

Subcharacteristic	TermoStat	GaleXtract	BioTex	DEXTER
Suitability	0.706	0.491	0.818	0.832
Precision	0.543	0.513	0.540	0.853
Operability	0.106	0.000	0.701	0.930
Time behaviour	0.774	0.858	0.508	0.823

Table 3. Comparative evaluation of term extractors.

In the following sections, we explore the most relevant results obtained from this research.

4.1 Suitability

It can be concluded that DEXTER is more suitable for the task of terminology research than TermoStat, GaleXtract and BioTex. First, DEXTER is not restricted to term extraction and term weighting but consists of a suite of tools that can integrate these two tasks into a corpus manager. In fact, DEXTER is provided with a range of functionalities that the other term extractors do not have. For example:

- Every document in the collection can be manually tagged with a content descriptor. This feature is of particular interest when we intend to find out in what type of texts a given term frequently appears.
- Regex-based queries can be formulated during corpus exploration.

Second, the hybrid model of evaluation in BioTex and DEXTER, which interface with the term databases MeSH/UMLS[8] and IATE[9] respectively, is certainly relevant not only

[8] MeSH (Medical Subject Headings) is a medical thesaurus that was devised for indexing scientific literature in databases such as MEDLINE/PubMed. UMLS (Unified Medical Language System) is a repository of over 150 biomedical vocabularies with the aim of developing computer systems that process biomedical language. Both of them are published by the US National Library of Medicine.

for term recognition but also for term validation, where the term database helps to relieve the burden of such a time-consuming task. Indeed, this is a critical factor in a scenario where "reducing the time needed for validation seems a necessary prerequisite for the acceptance of TETs [Terminology Extraction Tools]" (Zielinski and Safar 2005: 25). Third, TermoStat (English, French, Italian, Portuguese and Spanish), GaleXtract (English, French, Galician, Portuguese and Spanish) and BioTex (English, French and Spanish) and DEXTER (English, French, Italian and Spanish) can be described as monolingual term-extractors in multilingual systems. In this respect, it is important to note the difference between monolingual term extraction in a multilingual system, where a given term-extraction method is suited to work with several languages, and multilingual term extraction, which is intended to produce a multilingual terminological lexicon from aligned parallel corpora. While the first case outputs a monolingual inventory of terms at one time, the second case aims to create bilingual or multilingual resources. Although some research in bilingual term extraction has been carried out (cf. Fan, Shimizu and Nakagawa 2009; Lefever, Macken and Hoste 2009; Lee, Aw, Zhang and Li 2010; Bouamor, Semmar and Zweigenbaum 2012; Gaizauskas, Paramita, Barker, Pinnis, Aker and Pahisa Solé 2015), most of the work is monolingual, since "for terminologists the percentage of monolingual terminology work is significantly higher than in the case of translators and interpreters" (Zielinski and Safar 2005: 15). This is probably due to the profile of terminologists, who do not aim for translation but for the management and standardization of terminology. Fourth, TermoStat, BioTex and DEXTER can discover simple and complex terms, whereas GaleXtract recognizes just complex terms. Finally, “applications for domain-specific glossaries range from those

[9] IATE (InterActive Terminology for Europe), which has about 8.5 million terms in all 24 official EU languages, results from the compilation of all the terms used in many subject matters (e.g. politics, finance, education, applied sciences, humanities, among many others) by the translators of the various language services of the EU institutions.

that support direct human use to those that address the needs of computers" (Park *et al*. 2002: 1). In this sense, as well as being provided with a GUI, BioTex has been released as a Java library and DEXTER as a web service.

4.2. Precision

DEXTER clearly gets the best result in precision. A corpus of 100 Spanish texts (273,476 tokens) about odontostomatology was used to assess the precision of the 100 top-ranked unigrams, bigrams and trigrams. The documents were obtained from the scientific journal *Avances en Odontoestomatología*.[10] Preprocessing was required during corpus compilation, where the English abstract and the list of bibliographical references were removed in each document. With respect to the metrics, TermoStat and GaleXtract employ popular association measures: χ^2 (Nagao *et al*. 1976), log likelihood (Dunning 1994) and log odds ratio (Everitt 1992) in the former, and χ^2, log likelihood, mutual information (Church and Hanks 1990) and symmetric conditional probability (Silva and Lopes 1999) in the latter. In BioTex, a system for biomedical term extraction, Lossio-Ventura, Jonquet, Roche and Teisseire (2014b, 2014c) proposed the measures LIDF-value, F-OCapi and F-TFIDF-C, where the two latter combine C-value with Okapi and TF-IDF respectively to extract both single- and multi-word terms. DEXTER makes use of SRC (Periñán-Pascual, 2015), a parameterized metric for term ranking that relies on the theoretical principles of (a) salience, which measures the prevalence of terms in the document collection, (b) relevance, which measures the tendency in the usage of terms between a domain-specific corpus and a general-purpose one, and (c) cohesion, which measures the degree of stability of multi-word terms. Table 4 presents the results of precision after manual validation by three terminology researchers.

[10] http://scielo.isciii.es/scielo.php?script=sci_serial&pid=0213-1285&lng=es&nrm=iso

metric	unigrams	bigrams	trigrams
χ^2 [TermoStat, GaleXtract]	0.72	0.39	0.43
Log likelihood [TermoStat, GaleXtract]	0.58	0.39	0.40
Log odds ratio [TermoStat]	0.88	0.41	0.34
Mutual information [GaleXtract]	-	0.26	0.38
SCP [GaleXtract]	-	0.31	0.41
F-Ocapi [BioTex]	0.78	0.46	0.38
F-TFIDF-C [BioTex]	0.75	0.48	0.37
LIDF-value [BioTex]	0.65	0.46	0.40
SRC [DEXTER]	0.93	0.88	0.75

Table 4. Precision with the 100 top-ranked unigrams, bigrams and trigrams.

In line with mainstream ATE research, TermoStat, GaleXtract and BioTex primarily focused on nouns and noun phrases, under the assumption that they make up the bulk of the terminological inventory. However, it is also true that "verbs and adjectives, though they have received much less attention, can also be domain-specific" (Ahrenberg 2009), as manifestly shown by DEXTER with the extraction of terms such as *birradicular, bruñir*, *dentinario, estomatológico, gingival, hemostático, malar, mesiodistal*, *periodontal* or *suturar*.

4.3 Operability

The operability of DEXTER was rated significantly better because it is the only one of the four that can really manage a document collection as a corpus. Indeed, TermoStat and GaleXtract can only extract the ngrams from a single document. Moreover, BioTex and DEXTER can be tuned for a better performance of the system. In BioTex, the user can change the number of linguistic patterns used to filter term candidates, as well as the

function (i.e. average, maximum or sum) in the metrics F-Ocapi and F-TFIDF-C. In this research, BioTex was configured with the default number of linguistic patterns (i.e. 200) and with the maximum function, since Lossio-Ventura et al. (2014b) demonstrated that this function has the best behaviour for the first 300 terms after manual validation. In DEXTER, the performance of the SRC metric is conditioned by the true and false domains selected by the user. Whereas "true domains" correspond to the most relevant field(s) of specialized knowledge described in the corpus, "false domains" serve to discard term candidates that are commonly found in many scientific disciplines. Therefore, true and false domains play an important role not only in term recognition but also in term weighting. In this research, the true domains were Health [2841], Health care profession [2841001], Health policy [2841002], Illness [2841003], Medical science [2841004], Nutrition [2841005], Pharmaceutical industry [2841006] and Life sciences [3606003], and the false domains were Science [36], Natural and applied sciences [3606] and Applied sciences [3606001].

4.4 Time behaviour

Table 5 shows the results derived from the evaluation of response time in term weighting with the Spanish corpus (1.57MB).

System	Candidates	Time	Score
TermoStat	6,889	5m 52s	0.774
GaleXtract	1,807	14s	0.858
BioTex	1,200	9m 34s	0.508
DEXTER	3,137	1m 5s	0.823

Table 5. Response times.

It can be concluded that time behaviour is primarily affected by two factors, that is, by the approach to candidate extraction and, to a lesser extent, by the complexity of term weighting. On the one hand, TermoStat, GaleXtract and BioTex adopt a hybrid approach by employing TreeTagger for a POS-based selection of term candidates before statistical weighting, but DEXTER applies shallow lexical filters rather than elaborate morphosyntactic patterns. On the other hand, BioTex and DEXTER combine multiple metrics for term ranking—most of them on the basis of TF-IDF, whereas TermoStat and GaleXtract rely on conventional lexical association measures.

By way of a summary, Figure 1 graphically represents the evaluation of the external quality of TermoStat, GaleXtract, BioTex and DEXTER.

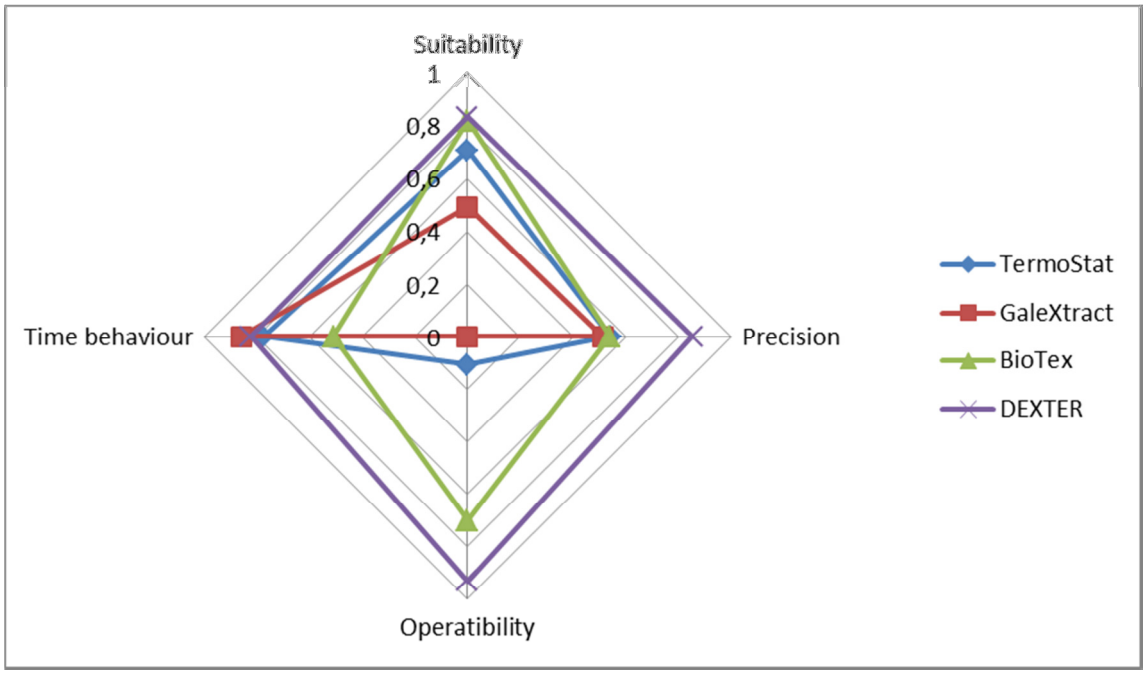

Figure 1. Comparative evaluation of ATE systems.

5. *Conclusions*

Within the FUNK Lab project, we have developed a virtual laboratory for natural language processing using analytical tools inspired in RRG, a functionally-oriented

linguistic paradigm. As part of this laboratory, a number of computational resources have been built, e.g. NAVIGATOR, DAMIEN, RONDA, CASPER, ARTEMIS and DEXTER. This paper offers an evaluation of the latter, which is an ATE system.

Indeed, the main goal of this research is to provide a more comprehensive framework for the evaluation of term extractors by enhancing previous proposals like Sauron (2002) and Zielinski and Safar (2005). Within this new framework, we perform a comparative analysis of DEXTER with the following three online open-access term extractors: TermoStat, GaleXtract and BioTex. The results obtained in terms of features such as suitability, operability, precision and time behavior conclude that DEXTER offers much better results than the other three, which are widely used within the linguistic community.

Acknowledgements

Financial support for this research has been provided by the Spanish Ministry of Economy, Competitiveness and Science, grant FFI2014-53788-C3-1-P.

References

Ahrenberg, L. 2009. Term extraction: A review. Retrieved from http://www.ida.liu.se/~lah/Publications/tereview_v2.pdf

Barcala, M., Domínguez-Noya, E., Gamallo, P., López, M., Moscoso, E., Rojo, G., Santalla, P., and Sotelo, S. 2007. A corpus and lexical resources for multi-word terminology extraction in the field of economy. In *Proceedings of the 3rd Language and Technology Conference*, Poznan, pp. 355-359.

Bouamor, D., Semmar, N., and Zweigenbaum, P. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association, pp. 674-679.

Carrión Delgado, M. G. 2012. Extracción y análisis de unidades léxico-conceptuales del dominio jurídico: un acercamiento metodológico desde FunGramKB. *RaeL* 11: 25-39.

Church, K.W., and Hanks, P. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics* 6 (1): 22-29.

Cortés, F. J. and R. Mairal 2016. "Building an RRG computational grammar" *Onomazein* (34), pp. 86-117.

Diedrichsen, E. 2014. "A Role and Reference Grammar Parser for German" in Brian Nolan and Carlos Periñán-Pascual (eds): *Language Processing and Grammars*. Amsterdam/Philadelphia: John Benjamins, 105-142.

Drouin, P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9 (1): 99-117.

Dunning, T. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1): 61-74.

EAGLES Work Group. 1999. *EAGLES evaluation of natural language processing systems*. Technical Report. Center for Sproktechnologi, Copenhague.

Everitt, B. 1992. *The Analysis of Contingency Tables*. London: Chapman & Hall/CRC.

Fan, X., Shimizu, N., and Nakagawa, H. 2009. Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, pp. 41-45.

Felices Lago, A., and Ureña Gómez-Moreno, P. 2012. Fundamentos metodológicos de la creación subontológica en FunGramKB. *Onomázein* 26: 49-67.

Gaizauskas, R., Paramita, M.L., Barker, E., Pinnis, M., Aker, A., and Pahisa Solé, M. 2015. Extracting bilingual terms from the Web. *Terminology* 21 (2): 205-236.

Guest, E. 2009. "Parsing using the Role and Reference Grammar paradigm". [http://eprints.leedsbeckett.ac.uk/778/6/Parsing%20Using%20the%20Role%20and%20Reference%20Grammar%20Paradigm.pdf, accessed 19 February 2016].

ISO/IEC 25010. 2011. *Systems and Software Engineering – Systems and Software Quality Requirements and Evaluation (SQuaRE) – System and Software Quality Models*. Geneva: International Organization for Standardization International Electrotechnical Commission.

ISO/IEC 9126-1. 2001. *Information Technology – Software Product Quality. Part 1: Quality Model*. Geneva: International Organization for Standardization International Electrotechnical Commission.

Lee, L., Aw, A., Zhang, M., and Li, H. 2010. EM-based hybrid model for bilingual terminology extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp.639-646.

Lee, J., and Paek, I. 2014. In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment* 32: 663-673.

Lefever, E., Macken, L., and Hoste, V. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, pp. 496-504.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., and Teisseire M. 2014a. BioTex: a system for biomedical terminology extraction, ranking and validation. In *Proceedings of the 13th International Semantic Web Conference*, pp. 157-160.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., and Teisseire M. 2014b. Towards a mixed approach to extract biomedical terms from text corpus. *International Journal of Knowledge Discovery in Bioinformatics* 4 (1): 1-15.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., and Teisseire M. 2014c. Yet another ranking function to automatic multi-word term extraction. In *Proceedings of the 9th International Conference on Natural Language Processing*, Warsaw.

Mairal-Usón, R., Lilian Guerrero and Carlos González (eds.), 2012. *El funcionalismo en la teoría lingüística. La Gramática del Papel y la Referencia. Introducción, avances y aplicaciones*. Madrid: Akal.

Mairal-Usón, R., and Periñán-Pascual, C. 2009. The anatomy of the lexicon component within the framework of a conceptual knowledge base. *Revista Española de Lingüística Aplicada* 22: 217-244.

Nagao, M., Mizutani, M., and Ikeda, H. 1976. An automated method of the extraction of important words from Japanese scientific documents. *Transactions of the Information Processing Society of Japan* 17 (2): 110-117.

Nolan, B. and C. Periñán-Pascual (eds.), 2014 *Language Processing and Grammars*. Amsterdam: John Benjamins.

Nolan, B. and Y. Salem, 2011. "UniArab: RRG Arabic-to-English machine translation" in Wataru Nakamura (ed.): *New Perspectives in Role and Reference Grammar.* Newcastle upon Tyne: Cambridge Scholars, 312-346.

Park, Y., Byrd, R. J., and Boguraev, B. 2002. Automatic glossary extraction: beyond terminology identification. In *Proceedings of the 19th International Conference*

on Computational Linguistics. Taipei: Howard International House and Academia Sinica, pp. 1-7.

Periñán-Pascual, C. 2013. A knowledge-engineering approach to the cognitive categorization of lexical meaning. *VIAL: Vigo International Journal of Applied Linguistics* 10: 85-104.

Periñán-Pascual, C. 2015. The underpinnings of a composite measure for automatic term extraction: the case of SRC. *Terminology* 21 (2): 151-179.

Periñán-Pascual, C., and Arcas Túnez, F. 2004. Meaning postulates in a lexico-conceptual knowledge base. In *Proceedings of the 15th International Workshop on Databases and Expert Systems Applications*. Los Alamitos: Institute of Electrical and Electronics Engineers, pp. 38-42.

Periñán-Pascual, C., and Arcas Túnez, F. 2005. Microconceptual-Knowledge Spreading in FunGramKB. In *Proceedings of the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*. Anaheim-Calgary-Zurich: ACTA Press, pp. 239-244.

Periñán-Pascual, C., and Arcas Túnez, F. 2007. Cognitive modules of an NLP knowledge base for language understanding. *Procesamiento del Lenguaje Natural* 39: 197-204.

Periñán-Pascual, C., and Arcas Túnez, F. 2010. Ontological commitments in FunGramKB. *Procesamiento del Lenguaje Natural* 44: 27-34.

Periñán-Pascual, C., and Arcas Túnez, F. 2014. La ingeniería del conocimiento en el dominio legal: La construcción de una Ontología Satélite en FunGramKB. *Revista Signos: Estudios de Lingüística* 47 (84): 113-139.

Preston, C.C., and Colman, A.M. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104: 1-15.

Saaty, T. L. 1977. A scaling method for priorities in a hierarchichal structure. *Journal of Mathematical Psychology* 15: 234-281.

Saaty, T. L. 1980. *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Salem, Y., A. Hensman and B. Nolan, 2008. "Towards Arabic to English machine translation", *ITB Journal* 17, 20–31.

Sauron, V. 2002. Tearing out the terms: evaluating term extractors. In *Proceedings of the Aslib Conference Translating and the Computer 24*. London: The Association for Information Management, pp. 1-18.

Silva, J.F., and Lopes, G.P. 1999. A local maxima method and a fair dispersion normalization for extracting multiword units. In *Proceedings of the 6th Meeting on the Mathematics of Language*, Orlando, pp. 369-381.

Tzeng, G. H., and Huang, J. J. 2011. *Multiple Attribute Decision Making: Methods and Applications*. Boca Raton: CRC Press.

Van Valin, Robert D. Jr. 2005: *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Van Valin, R.D. Jr and R. Mairal Usón 2014. "Interfacing the Lexicon and an Ontology in a Linking Algorithm" In M. Ángeles Gómez, F. Ruiz de Mendoza y F. Gonzálvez-García (eds.) *Theory and Practice in Functional-Cognitive Space.* Amsterdam: John Benjamins, pp. 205-228

Zielinski, D., and Safar, Y. R. 2005. T-survey 2005: An online survey on terminology extraction and terminology management. In *Proceedings of the Aslib Conference*

Translating and the Computer 27. London: The Association for Information Management, pp. 1-27.

Appendix A. Questions in the survey.

	Attribute	Question
SU	A1- Corpus language	Q1- How many languages can the term extractor process?
SU	A2- Corpus size	Q2- Is there a maximum size of the corpus?
SU	A3- Input file format	Q3- Which is the format of input files (i.e. corpus documents)?
SU	A4- Input specification	Q4- When compiling the corpus, can the user record some information about every document?
SU	A5- Output file format	Q5- Which is the format of output files (i.e. list of terms)?
SU	A6- Output specification	Q6- Together with the terms, can the user also obtain their weights?
SU	A7- Output format	Q7- Which is the format of term candidates?
SU	A8- Ngram type	Q8- Which type of ngrams do term candidates take the form of?
SU	A9- Functionality interface	Q9- Which type of interface is used?
SU	A10- Term validation	Q10- Can the term extractor help the user validate term candidates (e.g. with a reference list)?
SU	A11- Term search	Q11- Can the user retrieve the context of a given term? (If No, skip Q12)
SU	A11- Term search	Q12- Can the user build regex-based queries?
PR	A12- Precision	Q13- What is the precision of the term extractor?
OP	A13- Input recovery	Q14- Can the user recover the input (i.e. corpus documents)?

OP	A14- Input management	Q15- In a given terminological project, can the user update the corpus? (If No, skip Q16-Q18)
OP	A14- Input management	Q16- Can the user add new documents?
OP	A14- Input management	Q17- Can the user delete documents?
OP	A14- Input management	Q18- Can the user modify existing documents?
OP	A15- Metric adaptability	Q19- Can the term-extraction metric be configured for a better performance of the system?
TB	A16- Response time	Q20- Once the corpus has been uploaded, how long does it take to extract term candidates from that corpus?

SU = suitability; PR = precision; OP = operability; TB = time behaviour.

Appendix B. Scoring schemes in the survey.

1- one = 0.063; two = 0.127; three = 0.190; four or more = 0.254

2- yes = 0; no = 0.072

3- options: doc, html, odt, pdf, ps, rtf, txt, wp, xml; n × 0.001, where n = number of selected options

4- yes = 0.013; no = 0

5- options: txt/csv, html, json, xml; n × 0.009

6- yes = 0.047; no = 0

7- options: words, stems, lemmas; n × 0.020

8- options: unigram, bigram, trigram, tetragram or longer; n × 0.052

9- options: GUI, API/web service; n × 0.047

10- yes = 0.168; no = 0

11- yes = 0; no = 0

12- yes = 0.039; no = 0.019

13- the average of the precision values corresponding to the 100 top-ranked unigrams, bigrams and trigrams returned by the best metric in the term extractor (Table 4)

14- yes = 0.106; no = 0

15- yes = 0; no = 0

16- yes = 0.123; no = 0

17- yes = 0.050; no = 0

18- yes = 0.020; no = 0

19- yes = 0.701; no = 0

20- $1-\frac{1}{\log_2\left(2+\frac{k}{t}\right)}$, where k = number of term candidates, and t = processing time in seconds for the best metric in the term extractor (Table 4)