ISSN 1870-4069

Genex+, a Semantic-based Automatic Extractor of Examples Applied to Bilingual Terms

Jorge Lázaro[1], Juan-Manuel Torres-Moreno[2,3], Gerardo Sierra[4], Teresa Cabré[5], Andrés Torres Rivera[2,5]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Filosofía y Letras, Puebla, Mexico

[2] Université d'Avignon et des Pays de Vaucluse, Laboratoire Informatique d'Avignon, Avignon, France

[3] Ecole Polytechnique de Montréal, Département de Génie Informatique et Génie Logiciel, Quebec, Canada

[4] Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Mexico

[5] Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, Barcelona, Spain

jorge.lazaro@correo.buap.mx, juan-manuel.torres@univ-avignon.fr, gsierram@iingen.unam.mx, teresa.cabre@upf.edu, andres.torres01@estudiant.upf.edu

Abstract. In this paper we present *Genex+* (Genex plus), an improved implementation of the Genex system (Générateur d'Exemples) [17], by using the semantic closeness between certain fragments of texts. This process was based on the combination of the successive extraction of concordances and collocations with the determination of the semantic closeness by the Mutual Information method [24] and Cosine calculus [3,22]. Results have proven that a good example is almost always associated to the definition of the word to which it is making reference, and that it can be extracted automatically by the consecutive restriction of semantic fields applied to fragments of general language corpora. This has the advantage of presenting a conceptual reformulation of what is said in the definition by using highly informative textual fragments, but with a less formal register. For this second version, we implemented changes required for a bilingual entry.

Keywords. Terminology exemplification, computational lexicography, lexicometrical density, semantic saturation.

1 Background

1.1 Exemplification in Bilingual Dictionaries

When bilingual dictionaries are being planned, the design of correct examples has been a constant concern because of the pedagogical and supportive functions they perform

[1,6,8,10,18,19,20,21]. Nevertheless, the selection of good examples [2] from corpora represents a difficult task that needs techniques and tools to overcome the selection of candidates that could potentially be good examples.

Regarding examples in foreign language (L2) dictionaries, we have based this article on Philippe Humble's proposals[12]. In his study, Humble emphasizes that an example will not be unique for all users, rather that the words surrounding the term to be exemplified will depend on the training level of the users, which will be divided into two groups. The first group includes beginning learners who need the example to comprise very frequent words with very frequent meanings. While the second group includes advanced learners who need the example to comprise less frequent words or less frequent meanings of frequent words. Furthermore, the author mentions that the first group would use *made-up* examples, while the second group would use authentic examples (those retrieved from corpora). Similarly, Humble mentions that these authentic examples can be used to illustrate the use of intermediate frequency words.

However, even with these accurate notes, we must consider that there is no consensus on the exact source of examples. This is, generally the source that tends to be general corpora, this is not always a decisive criterion. This is important, because given that the exact source of an example is unknown, then there is no way of knowing if the speaker of the phrase to be used as an example or the text from where it is extracted differs from the definition of the term. By this, we mean that definitions and examples are drawn from different sources. One might think that the specialized dictionaries would be different because they need more detail to convey specialized concepts, however, both practices are very similar and the origin of examples and definitions is definitely not the same[17]. Furthermore, there are studies which mention that terminology or specialized units should be treated equally when referring to the example associated with them[25], as if examples for a word of general use and examples for a term followed the same process of association both with the term to which they refer, and with the reality they are a part of, in order to reflect the use of the term in the language.

In this sense, neither studies relating to examples in terminology (which are very few), nor studies relating to examples in lexicography, mention the source from which these fragments should be extracted, even when all the features they should have are mentioned. However, a quick scan shows that, because of the information or their function[11], their source is most relatable to a general language corpus.

In the literature devoted to the study of the subject at hand, a point where all perspectives converge has not been found and an exact point where the example could be ideally located and cannot be indicated. Previous research[17] has concluded that the example does not seem to be there *a priori*. To wit, the selected fragments are always related to a set of functions of complementary nature that must comply with regards to the discourse, and that adopts a better understanding of their reference, whether is a term or a word. Therefore, it is difficult to speak of a *location* in the strict sense of the word, because selected examples for dictionaries are most likely to be adaptations of textual fragments found in reference corpora or are creations based on the frequency of use of the words surrounding the term to exemplify.

However, without being there *a priori*, the example can be *identified* – and this is the key word – thanks to the conceptual complementarity needs, it covers regarding the term

and how a receiver or person consulting a dictionary apprehends it. "Complementarity needs" refer to the set of communication situations where the term is used and which readers need to associate with their every day lexicon. This is, to complement the meaning of a term (either very known or not) means locating the word in linguistic contexts where other surrounding words imply semantic relations which may lead the apprehension process of the meaning of the word to exemplify. Therefore, we consider that an example is not only a context where a term or word can be found, rather it is a context where the term or word is enunciative. In this sense, it is therefore basic to consider that the example has certain formal and semantic features that allow it to fulfill its role of *revealing* these semantic relations. The functions it performs, regarding the clarification of the meaning of a word, are based on the functions that each of the surrounding words in the exemplifying fragment establishes.

From the literature mentioned above, it is clear that current work on example identification and extraction from corpora makes use of the criteria we have mentioned, but without systematizing them, let alone naming them, meaning, it is done intuitively. A lexicographist or terminologist looks at how functional and potentially adaptable the documented fragments are (giving them their witness category) in order to be associated with a word, but does not consider the associations or relations between the word, its definition, and the missing information so that any reader can associate the word with a specific meaning. In this proposal we will focus only on examples associated to terms. We will highlight the identification and extraction of examples in specialized dictionaries, and we will consider that current examples have been designed as an adaptation of something existing – or documented –, that is, that they necessarily come from a verifiable source, such as corpora. We will also acknowledge that we have designed them as a pragmatic necessity, but mostly as a communicative necessity, because selecting the ideal textual fragment to exemplify a term must be cohesive internally (all its elements must be related), as well as externally (these elements must also relate to the definition of the exemplified term).

Therefore, based on the identification of examples associated to terms and their communication implications, we will be working specifically based on the Communicative Theory of Terminology [4], to try to explain the behavior of examples in terminology dictionaries. First, we will review the theoretical principals that will allow us to outline the notion of examples in terminology, and then we will show how we apply this proposal in the design of a tool that automatically extracts examples based on their semantic and syntactic features. Furthermore, we will show that a translator module, which allows users to generate examples in a source language based on terminological equivalents in a target language, has been added to this extractor.

1.2 Automatic Extraction of Examples

Many papers on the selection and extraction of examples for dictionaries show that human monitoring especially for linguistic issues (the most common problem being cohesion and coherence in phrases) is still needed; although human influence is trying to be kept to the minimum. For the development of the tool we are now presenting, we have relied on one of the most outstanding works in this field, GDEX or *Good Example*

Extraction[13]. This system tries to be a model of semi-automatic extraction of examples from large textual corpora. In former models, examples mainly followed the criteria of different lexicography manuals and treaties on dictionaries, which are summarized in Atkins and Rundell's work[2]. With this classic methodology as working principal, it is not unusual that some researchers designed tools to support the semi-automatic extraction of examples, consisting in obtaining concordances for collocations that were stored and analyzed by human agents, one-by-one, in order to find a good example. GDEX's method, however, retrieves classified concordances and presents them in order of importance. In order to classify them this way, it assigns weights obtained from a list that students have previously rated (based on their linguistic knowledge) which is then evaluated in a way that the system selects the same options as students, creating rules to select the best collocation[1].

The results suggest an interesting attempt, but the contextualization of the word to exemplify remained a difficult problem to overcome. In other words, GDEX extracts relevant fragments, but their relevance remains as a classification problem given that it cannot be completely done automatically because ambiguities often need to be solved. Research on other languages has obtained similar results which are satisfactory, but not always appropriate. Some cases worth mentioning are Slovenian[14] and Swedish[23]. In the case of German[7], the authors resorted to works previously done for GDEX and adapted the notions of *frequency*, *length*, and *instance of the word "matrix"*, which they summarized in two criteria: readability and complexity[2]. However, it is worth mentioning that this system works acceptably when used for lexicographical purposes, but results are not as satisfactory when used for terms or specialized words.

These approaches show that a certain type of pattern can be found, and that, by the combination of *concordances + collocations* a series of reductions could be used to determine the instance of certain lexical elements that complement the information of a word. These are complex concordances that show specific contexts, of which we will speak later.

1.3 Genex: Générateur d'Exemples

The Genex system works on the basis of an association measure called *lexicometrical density* [17]. This measure is a direct product of the notion of *semantic saturation*[3]. Semantic saturation is a theoretical framework which basically states that ideally a person knows a term in all its instances, its definitions, its contexts, and all the conceptual variants that it may have within a specialized discourse. If we include the above within the Communicative Theory of Terminology[4], which we are currently working with, this would be the same as saying that a person knows enough to cover all the sides of

[1] "Once the features have been identified, the question arises: how should they be weighted? Which features are most important, and by how much? With this in mind, we asked two students to select good examples for 1000 collocations. We then used those "known good" examples to set the weights by automatically finding the combination of weights that would give the "known good" examples the highest average rank. The first two features, sentence length and word frequencies, were given greatest weight"[13]

[2] `http://www.natcorp.ox.ac.uk/`

[3] To observe the specific use of this measure in exemplification refer directly to (Lázaro, 2015)

the polyhedron which represents the concept of a term, that is, the person would ideally fulfill the *Polyhedral principle*[5]. To measure *semantic saturation*, it is essential to acknowledge that the example is closely related to the definition of the term, and that the term's elements should appear ideally in the former, but not in a strict word order or under the same lexical rules. Meaning, there may be variation. Precisely this relation measures the *lexicometrical density*. The items that are measured in the example are those words present in the definition excluding defining verbs or function words. The term, in this theory, is an essential element.

The instance in the example of words from the definition addresses two main issues. On the one hand, the inclusion of the term in the example gives the fragment a conceptual nucleus around the elements of the definition revolve, which will ultimately be satellites and limits to the significance of that term. Seen in this way, the combination of the words of a definition in a new non-defining structure allows the term's activation context[15] to continue operating. A term's activation context is a theoretical concept that allows seeing any given word within a context where it acquires a specialized value. This is, a context where it is already being used as a term. This kind of context is, thus, a fragment which makes a given word a member of a particular semantic field. These fragments, however, are usually simple collocations of the given word. From this perspective, applying another filter is needed in order to find complex collocations.

Not every activation context is a sentence. To give cohesion to a new fragment which could potentially work as an example, only those that contain a verb semantically close to term will be chosen. This verb can exist *a priori* in the definition (a verb other than definitional) or may be the result of a search made within the working corpus through the *Mutual Information measure*[24]. This filter allows finding complex collocations of the terms to exemplify.

As it can be seen, the first part of the *lexicometrical density* calculates how much information a textual fragment has and how much semantic proximity it has with another fragment, so that the former can complement the latter at the conceptual level. In our case, we consider a definition to be a vector formed by its conforming elements, all the words. The second vector, the closest one, would be a potential example. This association is made through the cosine measure[3,22]. The cosine determines that a sentence is semantically close by the amount of information it contains, but does not take into account its length. Thus it is not difficult to infer that many of the fragments extracted with this methodology were too long, even more than the definitions to which they were associated. Therefore, we decided to choose only those fragments with high scores, but with fewer tokens. Having determined this restriction, we can consider that *lexicometrical density* is the product of the cosine value of a sentence (any with words from the definition) regarding another one (the same definition) by the inverse of the logarithm of its length. That is, the second part of the algorithm associated with this formula extracts the shorter sentences but which contain as much information as possible. Graphically, Formula 1 was used to determine the aforementioned:

$$score_i = \cos(sentence_i, q) \times \frac{1}{\log |sentence_i|}. \tag{1}$$

The Genex system uses four basic resources:

1. A general language dictionary.
2. A set of corpora.
3. A morphological tagger (TreeTagger).
4. A program to calculate Mutual Information.

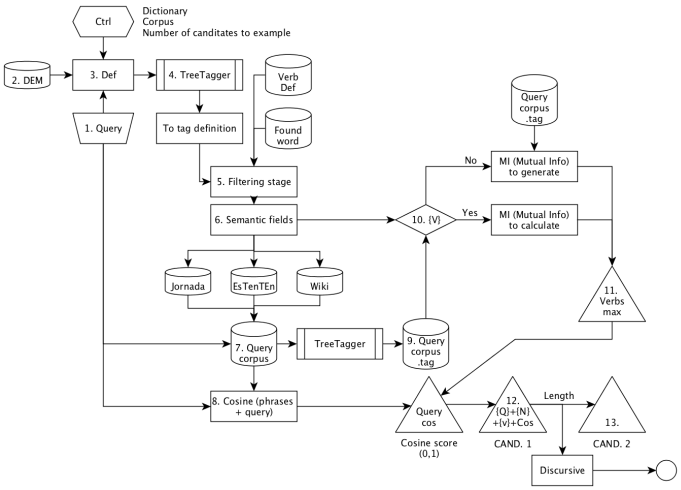

Fig. 1. Genex's Architecture.

As it can be seen, the program basically operates in the following way:

- The user inserts a term (1).
- The program asks to choose a meaning (2 and 3) from a pre-loaded dictionary.
- The program creates the nominal context subcorpus from lexical elements.
- the definition (4-9) from a general corpus.
- The program creates the verbal context with the verbal elements of the definition that do not belong to an exclusion list consisting of defining verbs in Spanish (10).
- If there are no verbs to continue the search, the system determines which verb is closer to term within a subcorpus including all the phrases containing one or more elements of the nominal context. In order to make this determination, it uses the measure Mutual Information (10 and 11).
- Once the nominal and verbal contexts are determined, it chooses those phrases that meet the restrictions: they unfailingly include the term, they contain one or more words of the nominal context, and they contain at least one element of the verbal context.
- It evaluates the chosen phrases using the *cosine measure*[4] to determine which one is closer to the definition.

[4] "...examples are ordered in respect to their goodness with the help of some soft criteria which are listed in the order of their importance: a) including words should be among the 17,000 most frequent words of our balanced corpus; b) including words should be no longer than 15 characters; c) finally, the keyword should be within the matrix clause."

– Among these, it chooses those that, in addition to meeting all the criteria, have fewer tokens.

All these criteria under which Genex works, have allowed us to obtain a methodology to extract those complex concordances of specific contexts of which we spoke earlier. This type of concordances, or phrases as they are called throughout the study, will be known as a *candidate for example*. This list of candidates fairly reliable from the lexical-semantic perspective, from which lexicographers or terminologists can choose an example suitable for the type of dictionary they are creating.

2 Genex+ and its operation

In this new stage, Genex's operation was modified to be used with several languages. Basically, a term translator was added so that equivalences of a target language could be found in a source language. For the translator to work, dictionaries in both target and source languages were needed. Finally, once the system was able to determine which and how many equivalents were relevant for each term, it proceeded to work as usual. This optimized language-independent system (it can work with several languages by only changing the dictionaries) is called *Genex+* (*Genex plus*). To test the effectiveness of Genex+, an experiment was performed with terms from the field area of oncology in two languages, French and Spanish.

2.1 A Term Translator Module

Machine translation (MT) is one of the central issues of Natural Language Processing (NLP). Although several MT issues remain unresolved, research in NLP has lead to the development of MT systems with an increasingly better performance.

A few years ago, MT systems used rules to translate from one language to another. This approach has several limitations, such as the difficulty to formulate rules for every language, or the need for experts who can write and code such rules in appropriate programs. Currently, statistical machine translation (SMT) has evolved with the aid of algorithms which use the large body of available corpora (both aligned and unaligned) to calculate the translation probability of phrases, fragments of phrases, or words.

The challenges of MT also extend to the semantic level, which remains the most difficult to overcome. For example, with the sentence *"Je vais acheter un canapé"*, how do we know if the speaker is hungry or requires furniture? The algorithm should be able to decide and propose a translation for *"canapé"* in Spanish to choose between *"sandwich"* or *"sofá"* (sandwich or couch). Or in *"j'ai vu la dame avec les jumelles",* how do we know if the lady was seen along with twin sisters or if the speaker used binoculars to look at the lady? In the case of English, once again choosing between *"I saw the woman with the twins"* or *"I saw the woman through the binoculars"* is an issue that has not been completely solved with current systems. We consider that the automatic extraction of examples may help to clarify the meaning of words – and therefore improve interpretation – if the extracted fragments come from the definitions of the exemplified word. That is, a possible disambiguation method would rely on the

direct observation of the candidates for examples of the given lexical item (canapé, lady, furniture, and so on).

In the framework of the Genex example generator, we wanted to enable the possibility to enter a term in a language L1 and obtain candidates to exemplify in another language L2. In the occurrence, L1 = English, French and L2 = {Spanish}. As might be expected, this function should imply the least number of changes on the system resources (parser, example ranking algorithm, etc). The simple solution that was used was a word-by-word translation for the entry terms. Figure 2 shows the simplified architecture of the system. Genex's core (algorithm + Spanish corpus + TreeTagger) remains intact, and the translation module is only an interface in the system input. At the moment, only one-word terms are used to generate examples, therefore MT syntax and semantics problems are irrelevant to our approach. Bilingual dictionaries from another language to Spanish let the system generate the examples using the methodology described below. Currently, a multiterm module at the input is being developed.

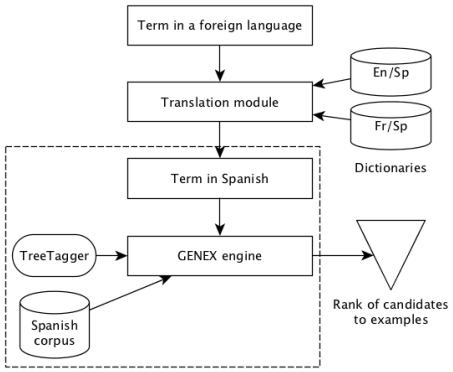

Fig. 2. Simplified Architecture of Genex+.

2.2 Dictionary Preprocessing

Lists of equivalences terms for English-Spanish and French-Spanish were required to carry out the corresponding tests with Genex+. These lists should proceed from an official source, thus the *Inter-Active Terminology for Europe* (IATE)[5] database was selected. This database contains 8.4 million terms and only includes the 24 official European Union languages, therefore Catalan could not be included.

IATE offers a distribution of their database with approximately 1.4 million entries and around 8 million terms in total. This database is downloaded as a TBX compressed file that can be consulted using a Java application which generates a subcorpus with the desired parameters introduced by the user. In this way, two working subcorpora were obtained, one for English-Spanish and one for French-Spanish, with equivalences

[5] http://iate.europa.eu/

shared by L1 and L2; this is, not all terms in Spanish have the same equivalences in other languages.

Once the working subcorpora were obtained, they were formatted into equivalence lists enabled for Genex+. IATExtract generates lists in TBX format[6] that are XML-based and allow information retrieval with software applications developed for this purpose or using a XML parser.

Because we required generating lists and not querying the corpus, a parser was developed in Python by implementing the *xml.sax*[7] library, to generate files with the format required by Genex+; only the tags that were necessary were kept and the rest were ignored. The work flow of the script in pseudo code is described below:

```
Read the files to be processed
Generate a new file
Find term code tags
Find language code tags
For each term tag
    If L1 and L2 share a term tag
        Group L1 and L2 terms by term code
        Write the data on the new file
        Apply the necessary format
Close used files
```

A first attempt was performed with the TERMCAT terminological dictionary, but given that the definitions were in Catalan, it could not be used by Genex+. For this reason, the terminological dictionary was obtained from the Spanish Association Against Cancer[8] and it was processed through the same procedure that had been applied to the IATE database. The dictionary's source code was downloaded and because it had been formatted as text in HTML, its processing was simpler: the HTML tags formatting the entries and definitions were read and the information was structured to match the format required for Genex+ afterwards. Once these elements were incorporated, the corresponding tests were carried out. The appearance of a dictionary edited with this methodology can be seen in figure 3.

2.3 Extraction of Examples Through Equivalent Terms

Once the new dictionaries and the lists of equivalences were obtained, we proceeded to generate examples in the target language (Spanish) from equivalences in the source language (French). The appearance of Genex+'s main menu can be seen in Figure 4.

Below are the results of five oncology terms in French, whose equivalents have been used to create a list of example candidates. There are only 2 cases of each term.Complete results for each one of them are shown in the Appendix[9]:

[6] http://iso.org/iso/catalogue_detail.htm?csnumber=4579

[7] https://docs.python.org/3.4/library/xml.sax.html

[8] https://www.aecc.es

[9] The experiment was carried out with 5 terms, resulting in a total of 25 example candidates. 44 people evaluated the 25 candidates, which yielded a total sample of 1100 variables.

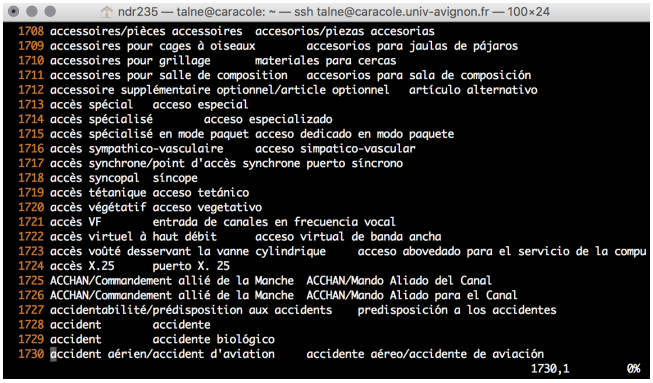

Fig. 3. Dictionary of Equivalent Terms.

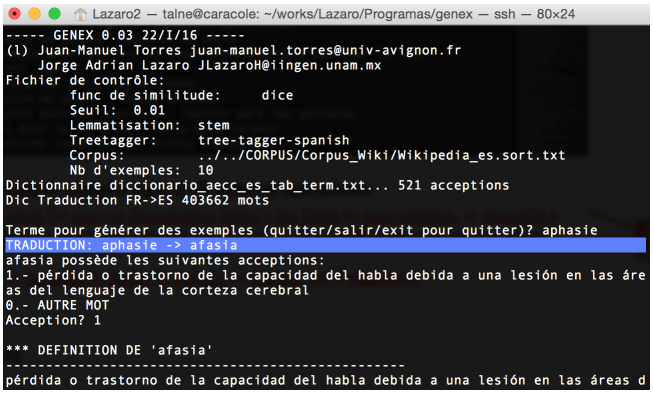

Fig. 4. Genex's Main Menu.

aphasie > afasia

1. Debido a esto, quedó traumatizada y durante dos años padeció afasia (pérdida del habla). 0.0889
2. Se distingue de una afasia motora en que no es un trastorno del lenguaje, sino del habla; es decir, el paciente manifiesta dificultades asociadas con la articulación de fonemas. 0.0815

carcinome > carcinoma

1. Estos son tumores caracterizados por células epiteliales grandes con citoplasma claro y abundante y con frecuencia se ven asociados con endometriosis o con un carcinoma endometrioide del ovario y con gran semejanza al carcinoma de células claras del endometrio. 0.0774
2. El carcinoma anaplásico de "células en avena" es el tipo histológico que con mayor frecuencia se asocia a la producción ectópica de hormonas y a síndromes paraneoplásicos. 0.0657

chimiothérapie > quimioterapia

1. Y es que los tratamientos a base de quimioterapia, de la que existen alrededor de 50 productos diferentes, si bien combaten las células cancerígenas, también destruyen las sanas. 0.0468
2. "En los casos en los que los pacientes sufren de enfermedades como el cáncer, tumores, melanoma, diabetes, sida, artritis reumatoide, alzheimer, esclerosis múltiple o lupus eritematoso severo pueden ser tratados con radiación o quimioterapia para destruir las células anormales. 0.0421

leucémie > leucemia

1. La leucemia linfoblástica precursora aguda de células B es un tipo de leucemia linfoide aguda que afecta en particular los precursores de los linfocitos B que están localizados en la médula ósea. 0.0882
2. Sin embargo, en algunos tipos de leucemias también pueden afectarse cualquiera de los precursores de las diferentes líneas celulares de la médula ósea, como los precursores mieloides, monocíticos, eritroides o megacariocíticos. 0.0641

tumeur > tumor

1. A veces el IGF -2 es producido en exceso por tumores de células islotes, causando hipoglucemia. 0.0815
2. Los tumores derivados de células del intestino posterior (carcinoide rectal) raramente producen un exceso de 5 -HIAA. 0.0774

2.4 Evaluation

The evaluation was carried out with the help of human agents. In total 44 people tested the relevance of the results provided by Genex+, of which 31 were women and 13 were men. Of all respondents only 6 did not hold a bachelor's degree, so they were considered a set of educated speakers. 95% of them were Spanish native speakers. Their average age was 31.24 years old.

For each term, 5 examples were extracted and each candidate was evaluated by the 44 evaluators. A total of 1,100 variables or answers were obtained to weight the relevance of the examples associated to the term. *The Pearson's correlation coefficient*, a measure of the linear correlation between two quantitative variables[9], was used to calculate the effectiveness of the system. This measure provided the degree of relation among the two variables: relevance of the examples produced by the system according to the criteria it uses and preference of speakers regarding the information these examples provide about the definition.

Pearson's correlation coefficient is represented by $\rho_{X,Y}$ and is calculated through Formula 2:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{2}$$

where XY is the covariance of (X,Y), σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Cohen's Kappa correlation measures the agreement between two evaluators, or two evaluation methods. It is represented by κ and calculated using Formula 3, where P_o is the observed agreement among raters and P_e is the hypothetical probability of chance agreement:

$$\kappa \equiv \frac{P_o - P_e}{1 - p_e} = 1 - \frac{1 - P_o}{1 - p_e}. \tag{3}$$

The data obtained is to be interpreted as the existent correlation between the system choices and what speakers prefer. These results range between 1 and -1. A value close to 1 means that the system has succeeded in the selection of a candidate for example, the more distant it gets, that is, when it gets closer to -1 it means the system failed at ordering as humans do. A value of 0 implies certain similarity between the human and system preference/aversion towards certain results.

The following table shows the results obtained by Genex+. The example's number corresponds to the order shown on the previous section, 1 being what the system considers an ideal example and 5 the least likely example according to their *lexicometrical density*. As it can be seen in Table 1, the Pearson column shows the extent to which the human evaluators agreed with this arrangement.

3 Discussion and Conclusions

As it can be seen, the system had a good performance overall, as both human evaluators and the system results mostly agree. There was an agreement above 0.3 almost on every instance, which indicates a trend towards total agreement. Furthermore, there are cases with agreement quite close to 1. For instance, according to this evaluation, *"chimiothérapie"* and *"leukemia"* obtained values as high as 0.7.

As in all automatic systems, we find complex cases, such as the example candidates for the term *"tumeur"* were evaluated almost inversely, in other words, closer to -1. Regarding this particular instance, it is worth mentioning that there is no aversion towards the candidates proposed by Genex+, but rather to their *lexicometrical density* ranking. Apparently human agents had a preference for the length of the example over its density. This is to be expected if we take into account that most of the times the explanatory function of the example or its pragmatic extension function are hampered by the brevity of a sentence. As can be seen, the most accepted example candidates (3 and 5) for this term contain words associated to the term, as well as specifications of the functions or affectations of *"tumor"*. This information can often be curtailed if we consider only the definition to which an example is associated and not its common use. Although the tool associates this fragment to the specialized definition in an acceptable way, and brings it to a less formal register, the fact is that there are also information specifiers or magnifiers that do not occur at word level (a word that determines another in a specific context, which is how Genex+ works) but at discursive level, where a whole sentence specifies a single word of another sentence; all of which complicates the analysis. This, as can be inferred, complicates the automatic processing of examples, because we would have

Table 1. Results obtained by Genex+.

Example	Human Response	Pearson	Kappa
aphasie >afasia			
1	9.38%	0.4	0.0
2	37.50%		
3	31.25%		
4	15.63%		
5	3.13%		
carcinome >carcinoma			
1	43.75%	0.3	0.25
2	12.50%		
3	15.63%		
4	0.00%		
5	25.00%		
chimiothérapie >quimioterapia			
1	34.38%	0.7	0.25
2	28.13%		
3	12.50%		
4	6.25%		
5	15.63%		
leucemie >leucemia			
1	28.13%	0.7	0.25
2	6.25%		
3	50.00%		
4	6.25%		
5	6.25%		
tumeur >tumor			
1	0.00%	-0.9	-0.25
2	3.13%		
3	40.63%		
4	6.25%		
5	46.88%		
Average		**0.24**	**0.1**

to think that sometimes the meaning of a term and its exemplification is not only in the phrase where it can be inserted, rather this exemplifying information is shared along larger structures such as subordinate clauses. Even with this, there are few cases that yield such results, so we can conclude that overall the Genex+ system works correctly and humans consider it an acceptable tool for selecting candidates as examples.

Finally, the results suggest that the best examples in a bilingual dictionary would be associated with the definition of the target language and not to that of the mother tongue[16], because the linguistic system is different. However, examples in both languages would complement each other to attest the way in which reality is apprehended

by the language that is required to know, as it has already been shown by Humble[12]. To achieve this, we could reverse the dictionaries and corpora to obtain a bidirectional tool, which is completely possible if it is taken into account that Genex+ is a language-independent system. This means that it would help the learner to discern those contexts where an equivalent can be used, and when this is not possible because the contextual usage restrictions have been automatically applied previously by the program.

Appendix: Complete list of results for each term

aphasie > afasia

1. Debido a esto, quedó traumatizada y durante dos años padeció afasia (pérdida del habla). 0.0889
2. Se distingue de una afasia motora en que no es un trastorno del lenguaje, sino del habla; es decir, el paciente manifiesta dificultades asociadas con la articulación de fonemas. 0.0815
3. Las disritmias son un trastorno de emisión, al igual que las dislasias y disartrias (trastornos de articulación por anomalías en los órganos articulatorios que pueden solucionarse con logopedia), disfasias y afasias (deterioro al adquirir el habla y pérdida de la misma) y disfonías y afonías (anomalías en la producción por afecciones laríngeas). 0.0653
4. El descubrimiento por este último de que las lesiones localizadas en una región del lóbulo frontal izquierdo (llamado después "área de de Broca") podrían entrañar una afasia (una incapacidad de hablar) infligieron un duro golpe a la doctrina holística. 0.0502
5. La afasia de Wernicke no sólo afecta a la comprensión del habla. 0.0491

carcinome > carcinoma

1. Estos son tumores caracterizados por células epiteliales grandes con citoplasma claro y abundante y con frecuencia se ven asociados con endometriosis o con un carcinoma endometrioide del ovario y con gran semejanza al carcinoma de células claras del endometrio. 0.0774
2. El carcinoma anaplásico de "células en avena" es el tipo histológico que con mayor frecuencia se asocia a la producción ectópica de hormonas y a síndromes paraneoplásicos. 0.0657
3. Los factores de riesgo de cáncer cervical están relacionados con características tanto del virus como del huésped, e incluyen: Varios tipos de VPH, particularmente el tipo 16, han sido hallados asociados con carcinoma orofaríngeo de células escamosas, una forma de cáncer de cabeza y cuello (en inglés). 0.0513
4. Se ha sugerido que un SNP extraño (rs11614913) que está superpuesto a la hsa-mir-196a2 está asociado al carcinoma pulmonar de las células pequeñas. 0.0461
5. Los tumores malignos, por su parte, son del tipo adenocarcinoma quístico formando un 40% de todos los carcinomas malignos de ovario, por lo general se ven en mujeres avanzadas de edad, frecuentemente asociados a casos familiares y un 66% de los casos de tumores malignos de ovario son bilaterales. 0.0410

chimiothérapie > quimioterapia

1. Y es que los tratamientos a base de quimioterapia, de la que existen alrededor de 50 productos diferentes, si bien combaten las células cancerígenas, también destruyen las sanas. 0.0468
2. "En los casos en los que los pacientes sufren de enfermedades como el cáncer, tumores, melanoma, diabetes, sida, artritis reumatoide, alzheimer, esclerosis múltiple o lupus eritematoso severo pueden ser tratados con radiación o quimioterapia para destruir las células anormales. 0.0421
3. Con los conocimientos acerca del efecto en conjunto de los tres genes se podráinvestigar de manera más precisa cómo adquieren las células cancerígenas las características con las que logran sobrevivir en el cuerpo, y probablemente promover la creación de medicamentos para cada tipo de cáncer, lo que disminuiría el uso de la quimioterapia y la radiación, que destruyen también el tejido normal. 0.0355
4. Y también los medicamentos y remedios necesarios se aplican bajo diferentes esquemas: los productos que destruyen los vasos sanguíneos se aplican por largos periodos, mientras que la quimioterapia se aplica por ciclos. 0.0188
5. Discos magnéticos ultrafinos de un micrón de diámetro y unos 60 nanómetros de espesor pueden destruir células cancerígenas, sin los efectos secundarios de las quimioterapias, indicó un estudio publicado el domingo por la revista científica Nature Material. 0.0148

leucémie > leucemia

1. La leucemia linfoblástica precursora aguda de células B es un tipo de leucemia linfoide aguda que afecta en particular los precursores de los linfocitos B que están localizados en la médula ósea. 0.0882
2. Sin embargo, en algunos tipos de leucemias también pueden afectarse cualquiera de los precursores de las diferentes líneas celulares de la médula ósea, como los precursores mieloides, monocíticos, eritroides o megacariocíticos. 0.0641
3. Madrigal considera que la leucemia -cáncer de sangre- no afecta a células secundarias sino a las células madres hematopoyéticas que son las que funcionan mal en los pacientes enfermos. 0.0641
4. En la leucemia mieloide crónica la translocación genética de los cromosomas 9 con 22 da lugar a una tirosincinasa que ejerce una importante acción en los mecanismos de adhesión, apoptosis y proliferación de las células mieloides afectadas. 0.0392
5. Los precursores de linfocitos B afectados tienen una serie de receptores en su membrana, lo cual permite identificar el tipo de leucemia como una leucemia linfoblástica precursora aguda de células B. Otras enfermedades: diagnóstico de leucemia aguda linfoblástica hace 5 años recibió quimioterapia se desconoce número de ciclos y medicamentos utilizados. 0.0390

tumeur > tumor

1. A veces el IGF -2 es producido en exceso por tumores de células islotes, causando hipoglucemia. 0.0815

2. Los tumores derivados de células del intestino posterior (carcinoide rectal) raramente producen un exceso de 5 -HIAA. 0.0774
3. Se caracteriza por el crecimiento de tumores benignos que puede desarrollarse a partir del tejido de un solo linfonodo o a partir de múltiples sitios simultáneamente El crecimiento de los linfonodos radica en la hiperproliferación de ciertas células B que con frecuencia son productoras de múltiples citoquinas. 0.0608
4. Se inyecta calcio en el páncreas, lo que causa que las células beta liberen insulina; si hay un tumor que causa un exceso de células beta, la insulina se producirá en demasía y el azúcar caerá demasiado y abruptamente. 0.0591
5. Warburg hipotetizó que el cáncer, el crecimiento maligno y el crecimiento de los tumores son causados por el hecho de que las células tumorales generan energía (producen ATP) principalmente por medio de una degradación no oxidativa de la glucosa (un proceso llamado glicólisis anaeróbica; al contrario de lo que ocurre con las células saludables, las cuales generan energía principalmente a partir de la degradación oxidativa del piruvato. 0.0580

References

1. Alvar Ezquerra, M.: Diccionario y gramática. LEA: Lingüística española actual 4(2), 151–212 (1982)
2. Atkins, B.T., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxfor University Press, Oxford (2008)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Moder Information Retrieval. ACM Press Addison-Wesley, New York (1999)
4. Cabré, T.: La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (1999)
5. Cabré, T.: El principio de la poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en terminología(i). Ibérica pp. 9–36 (2008)
6. Cowie, A.P.: The Language of Examples in English Learners' Dictionaries, pp. 55–65. No. 14, University of Exeter, Exeter (1989)
7. Didakowski, J., Lemnitzer, L., Geyken, A.: Automatic example sentence extraction for a contemporary german dictionary. In: Eurales 2012. pp. 343–349 (2012)
8. Drysdale, P.D.: The role of examples in a learner's dictionary. In: Cowie, A.P. (ed.) The Dictionary and the Language Learner. Papers from the EURALEX Seminar at the University of Leeds (1987)
9. Egghe, L., Leydesdorff, L.: The relation between Pearson's correlation coefficient r and Salton's cosine measure. ArXiv e-prints (Nov 2009)
10. Fox, G.: The case for examples. In: Sinclair, J. (ed.) Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. Collins ELT, London (1987)
11. Fuentes Morán, T., García Palacios, J.: Los ejemplos en el diccionario de especialidad. In: Fuentes Morán, T., García Palacios, J. (eds.) Texto, terminología y traducción. Almar, Salamanca (2002)
12. Humble, P.: The use of authentic, made-up, and controlled examples in foreign language dictionaries. In: Fontenelle, T. (ed.) EURALEX '98. University of Liège, Liège (1998)
13. Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., Rychlý, P.: Gdex: Automatically finding good dictionary examples in a corpus. In: DeCesaris, J. (ed.) 13th EURALEX International Congress. Series del Institut Universitari de Lingüística Aplicada (2008)

14. Kosem, I., Husak, M., McCarthy, D.: Gdex for slovene. In: Proceedings of eLex 2011. pp. 151–159 (2011)
15. Kuguel, I.: La activación del significado especializado. In: Lorente, M., Estopà, R., Freixa, J., Martí, J., Tebé, C. (eds.) Estudis de lingüística aplicada en honor de M. Teresa Cabré Castellví. Series del Institut Universitari de Lingüística Aplicada, Barcelona (2007)
16. Laufer, B.: Corpus-based versus lexicographer examples in comprehension and production of new words. In: Proceedings of the Fifth Euralex International Congress:. pp. 71–76 (1992)
17. Lázaro, J.: El ejemplo en terminología. Caracterización y extracción automática. Ph.D. thesis, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona (2015)
18. Minaeva, L.: Dictionary examples: Friends or foes? In: Tommola, H. (ed.) 5th EURALEX '92 Proceedings. University of Tampere, Tampere (1992)
19. Nesi, H.: The role of illustrative examples in productive dictionary use. Dictionaries: Journal of the Dictionary Society of North America 1(17), 198–206 (1996)
20. Paquot, M.: Exemplification in learner writing: a crosslinguistic perspective. In: Meunier, F., Granger, S. (eds.) Phraseology in Foreign Language Learning and Teaching. John Benjamins Publishing (2008)
21. Sinclair, J.: Collins COBUILD English Language Dictionary. Collins, Birmingham (1987)
22. Spärck-Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1(28), 11–21 (1972)
23. Volodina, E., Johansson, R., Johansson Kokkinakis, S.: Semi-automatic selection of best corpus examples for swedish: Initial algorithm evaluation. In: Proceedings of the SLTC 2012 workshop on NLP for CALL. pp. 59–70 (2012)
24. Ward Church, K., Hanks, P.: ord association, mutual information and lexicography. Computational Linguistics 1(16), 22–29 (1990)
25. Zgusta, L.: Manual of Lexicography. Walter de Gruyter, Paris (1971)