ISSN 2007-9737

Design of an Automatic Tagging Algorithm for the Development of a Non-Literal Language Corpus in Spanish

Ericka Ovando-Becerril, Hiram Calvo

Instituto Politécnico Nacional,
Centro de Investigación en Computación, Mexico City,

Mexico

{eovandob2021, hcalvo}@cic.ipn.mx

Abstract. The development of corpus in general represents an arduous task due to the analysis and labeling processes, as well as the elements to consider in its development. For its part, the study of literal language and non-literal language in its various literary figures represents an important area of study for Natural Language Processing. The study of this phenomenon in Spanish has been limited by the lack of corpora available for its study as well as the lexical and semantic complexity of the language. In accordance with the above, this project proposes a labeling algorithm for non-literal and literal language, likewise reviews the points to consider regarding design and experimentation and presents results from the CESS-ESP corpus.

Keywords. Metaphor, semantics, natural language processing.

1 Introduction

The development of the corpus and the technological development have allowed the evolution of natural language processing, either in a general way, or focused on linguistic phenomena [2, 20, 13, 19, 18]. Currently, the study of non-literal language in English contrasts significantly with the study of non-literal language in Spanish, on this problem one of the main limitations is the available corpora and in other hand the lexical and semantic complexity. Particularly in Spanish, there is no specific corpus for the study of non-literal language, nor is there one for the study of metaphorical language.

The relevance to develop new corpora is clear when it is studied the relevance of non-literal language and its role in everyday language.

The present work focuses its efforts on presenting an algorithm that allows to automatically label a set of CESS-ESP [8] sentences to generate a corpus of literal language and non-literal language in Spanish. The applications of the resulting corpus have a direct impact on the study of non-literal language since the NLP in Spanish and that is relevant in many tasks like translation, education [9, 15, 4] and discourse analysis to mention some of their applications.

The perspective of this algorithm has its origin in the syntactic analysis and it semantic tagging, as well as in the structure of the sentences presented. Likewise, some corpora already developed mainly in English are taken as a reference, without leaving aside the particularities of Spanish for this first proposal. At this article is presented the syntactic and semantic elements to consider, as well as the algorithm developed for corpus labeling and some examples of the first part of experimentation. For that work we consider two little definitions for [5]:

Literal language: It is the one that has a locution, a twist, etc., according to the sum of its meanings. For example, use the language established according to the dictionary.

Non-literal language: Language in which abundant rhetorical figures are used such as metaphor or metonymy, is frequent.

It is important to mention that although non-literal language was originally considered to be present

doi: 10.13053/CyS-27-4-4779

in a particular way in literary language, currently the cognitivist theory has made it possible to study its presence and importance in everyday language [5].

1.1 Corpus in Spanish

Some important elements to consider in the development of the corpus is the manual tagging witch is usually compared with any other tagging method considered in the corpus development. Likewise some corpora consider lexical tagging or the development of a lexicon [16, 6]; in other hand others relevant points is the decision of conformation of the corpus from paragraphs, sentences, or lists of words and the extension and the theme of the corpus [1, 14].

For example in English, the three most important corpora of literal and non-literal language are:

- MOH-X: it contains 646 sentences and doesn't contain a predetermined training and test set. This dataset is derived from the MOH set and the verbs are used as metaphors.
- The VUA sequence set: it contains 5323 sentences and has a predetermined training and test set. It represents text extracted from 117 British National Corpus fragments from four genres: academic, news, conversation and fiction; and corresponds to texts written between 1985-1994.
- TroFi database [11]: it contains 3737 sentences and doesn't contain a given training and test set. It corresponds to excerpts from the Wall Street Journal Corpus from the years 1987-1989.

One of the main characteristics of these three corpora is that they derive from established corpora that are representative of the English text. On the other hand, its conformation, as in the case of the TroFi database, corresponds to semantic analysis and tagging algorithms using wordnet [3].

The decision to work with corpora in Spanish implies, in addition to the grammatical and lexical complexity of the language, a complication in the corpora available for free download. Some of the identified corpus are Corpus del Español NOW [7], Molino Labs [17], Timestamped JSI web [21], Spanish News Text [12], el proyecto Aracne [10] most of these are restricted for download.

Finally, the search of the corpus and the considerations of its conformation, are a key work to understand the complexity of decision-making around this work. It is important to mention that the development of this work has been built largely as a result of experimentation and taking work in other languages as a reference because some points have not been studied, or at least have not been found in the investigation of the language at the state of the art.

2 Semantic Perspective of Literal and Non-Literal Language in Spanish

Formally, non-literal language modifies the use of a word according to its established meaning, for example, considering the figure of the metaphor, it can be understood as the extrapolation of a word or a set of words from one semantic to another.

On the other hand, a corpus or set of texts is formed, in its most essential form, from a set of words, that is, from a vocabulary.

Starting from the above, let w be a word, S the set of words that make up a sentence and V the set of vocabulary, then:

$$w \in S. \tag{1}$$

$$S \subset V. \tag{2}$$

It is important to consider the sets of main grammatical categories that can be considered more depending on the type of analysis: D set of determiners, U set of nouns, J set of adjectives, E set of verbs, P set of prepositions, A set of adverbs, C set of conjunctions, I set of interjections. That is to say:

$$\forall w \in S, w \in D \vee U \vee J \vee P \vee E \vee A \vee C \vee I. \tag{3}$$

Understanding that $w \in V$, where:

$$D, U, J, E, P, A, C, I \subset V. \tag{4}$$

In turn, a grammatical sentence contained in a set S corresponds within a text to a sequence of words that can be represented in a vector and in turn can be represented in the hyper-plane. Considering that the language is composed of rules that govern its behavior, we can understand that these sequences will correspond to a sequence of words w of the various grammatical categories.

Conventionally:

$$\forall w \in S, w \in H, \tag{5}$$

where H corresponds to the set of words of a semantic field that contains words of different grammatical types referring to the same topic or related topics. The above can be represented in the hyper-plane as a proximity between this set of words or in graphs as a short path that will allow them to be connected.

According to the above, in a concise way, let F be a non-literal sentence generated from S, where $\forall w \in S, w \in H$ and H and L are sets of semantic fields independent then:

$$\exists w \in F, w \notin H \wedge w \in L. \tag{6}$$

To better exemplify this proposal, the following two sentences can be considered where the use of the verb ***absorb*** stands out:

- *Non-literal: In no case did expenditures rise enough after the war to **absorb** the **revenue** that would have been available if the war taxes had been retained .*
- *Literal: This time, **the ground absorbed the shock** waves enough to transfer her images to the metal in bas-relief.*

According to the RAE, absorbing can be defined as ”attracting and retaining [something from the outside] inside” or ”to consume something entirely”; In this example, absorbing in the literal case implies attracting and retaining energy; where energy can be found in a semantic field similar to some other physical elements such as water or pollution and the idea of absorbing economic resources, as proposed in the sentence labeled as non-literal, proposes extrapolating completely physical elements to a non-tangible element such as economic resources, transferring from natural elements (for example) to economic elements.

In conclusion, these anomalies within the linguistic and semantic field are the guidelines of a non-literal idea, particularly the preamble of a metaphorical structure. This perspective allows us to understand the linguistic phenomenon in a delimited way from the computational field as representations of sets, graphs or vectors.

3 Automatic Tagging Algorithm Design

The main concepts that have been evaluated and considered for the approach of this work have been addressed, mostly giving rise to a series of questions and experimentation. It is important to point out that this writing corresponds to the first results.

3.1 Corpus

The corpus that was selected to test the algorithm was CESS-ESP available on the Natural Language Toolkit platform, which contains 6030 sentences and was developed by the Universitat de Barcelona.

Another point in common with corpora of this type in other languages is the topic related to news. It is important to mention that one of the peculiarities of the corpus is the length of the sentences that complicate its processing and labeling as it contains several subordinate clauses.

3.2 Syntactic Analysis

Starting from the general model where the syntactic analysis of the sentence allows to understand the interrelation of the words of the sentences to find the literal or non-literal meaning of the text, the method of the part of speech (POS) was used and tagging of morphological characteristics of Stanza to label each word, with the universal POS tags that is shown at Table 1.

Table 1. These tags are important to mark the core part-of-speech categories. That labels help to analyze lexical and grammatical properties of words, use the universal features

Universal POS tags		
Open class words	***Closed class words***	***Other***
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

3.3 Semantic Analysis

Once the words were tagged with POS, it was necessary to consider the semantic similarity, it was decided using WordNet, however one of the central questions was which words or which interrelationships are the most important to evaluate the similarity of the words? Why evaluate the semantic similarity of the words and not evaluate the sentence in general?

Likewise, it was necessary to consider in the first place, from the corpus in English, the direct relationship of the verb and the nouns, specifically considering the use of the verbs in a literal or non-literal way according to the closest nouns.

For this work, however, a second question arose around the length and complexity of the sentences, should all the nouns and verbs of the sentence be considered? After testing the semantic similarity between all the verbs, all the nouns, or considering other words like auxiliaries and adjectives or all the verbs and nouns, it was possible to generate some basic rules about which words.

According to the above, it was proposed to select the words according to the following rules to generate a list with the words of interest but trying to reduce the list according to our interest:

– The first word is either a noun or a verb, for example for a sentence that is started with *”La electricidad producida [...]”* the fisrt word to be consider in the list is *electricidad*.

  Another example is sentence that is started with *” Aumentó el producto interno bruto[...]”* the first word to be consider for the list is the verb *creció*.

– In relation to the previous point, all the verbs in the sentence must be considered.

– Nouns found after a verb, an adjective or an auxiliary are considered important.

  For example for phrase *”La electricidad producida pasará a la red eléctrica pública de México”* the noun *red* is important because that is related with the main verb of the sentence, and is after a verb. At the same sentence the noun *Mexico* is after the adjetive *Pública*.

  Another example is the phrase *Líbano puede tener en Oriente Medio una gasolinera* the noun *Oriente Medio* is after to the auxiliary *puede* but in this sentence the noun *gasolinera* is not consider at the list.

The result of the above process is a list of lemmas words in Spanish that have been selected because they are representative of the context within the sentence.

Each word in the set is translated in order to be processed with the similarity function of WordNet. Subsequently, the semantic similarity analysis of the bigrams of the words on the list is performed:

$$Select\_Words = [s_1, s_2, s_3, s_4, ...s_n]. \quad (7)$$

A vector is generated that corresponds to the similarity results between the words, but zero results for semantic similarity are not considered:

$$Similarity = [similarity(s_1, s_2), \\ similarity(s_2, s_3), ...similarity(s_{n-1}, s_n)]. \quad (8)$$

Finally, the mean of the vector is obtained, and to make the tagging process is considered a mean greater than 0.5 to label the sentence as literal language and less or equal than 0.5 to label as non-literal language.

4 Results

This section presents some examples of tagging sentences first as literal language and later as non-literal language. Likewise, some particular cases are considered where the algorithm fails and finally the relationship of the generated labels and the expected labels is presented to see the performance of the algorithm.

In the first table with the examples of sentences labeled as literal language, there are three clear examples of sentences in which the words extracted from the list allow the semantic similarity analysis, for example the relation and the use of the verbs is clearly what is expected according to the formal definition of the words, particularly the use of the verbs present in each of the sentences: “aumentar”, “conseguir”, “salvaguardar”, “ir” and “reunir”.

Let us see example of sentences tagged as literals.

Sentence 1:
Se trata de los cánceres de piel, huesos o pulmón, cataratas, anemias aplásticas y leucemias, cuya probabilidad aumenta con la dosis absorbida por los tejidos.
(*These are skin, bone or lung cancers, cataracts, aplastic anemias and leukemias, the probability of which increases with the dose absorbed by the tissues.*)
Word list (Spanish): [’tratar’, ’cáncer’, ’leucemia’, ’aumentar’, ’dosis’, ’tejido’]
Word list: [’be’, ’cancer’, ’leukemia’, ’increase’, ’dosis’, ’tissue’]
Media = 0.61
Label = Literal

Sentence 2:
Gobierno ha conseguido salvaguardar los intereses vitales de España dentro del proceso de construcción europeo, en el que *0* ha jugado un papel de animador del mismo”, sostuvo *0* en una conferencia de prensa.
(*The Government has managed to safeguard the vital interests of Spain within the European construction process, in which *0* it has played a role as its animator,” *0* said in a press conference.*)
Word list (Spanish): [’gobierno’, ’conseguir’, ’salvaguardar’, ’interés’, ’proceso’, ’jugar’, ’papel’, ’sostener’, ’conferencia’]
Word list: [’government’, ’manage’, ’safeguard’, ’interest’, ’process’, ’play’, ’role’, ’sostain/say’, ’conference’]
Media = 0.55
Label = Literal

Sentence 3:
Diversas actividades, que van desde un congreso mundial de prostitutas hasta una prueba ciclista, dificultarán la labor de la policía de Berlín durante este fin de semana en que se reúnen en la capital 14 líderes para una cumbre de reformistas.
(*Various activities, ranging from a world congress of prostitutes to a cycling test, will make the work of the Berlin police difficult during this weekend when 14 leaders meet in the capital for a summit of reformists.*)
Word list (Spanish): [’actividad’, ’ir’, ’congreso’, ’prostituta’, ’dificultar’, ’labor’, ’meet’, ’capital’]
Word list: [’activity’, ’range/go’, ’congress’, ’prostitute’, ’difficult’, ’work’, ’reunir’, ’capital’]
Media = 0.51
Label = Literal

Now let us see example of sentences tagged as non-literal.

Sentence 4:
La selección de fútbol de Ecuador, que dirige el colombiano Hernán Gómez, viajarán mañana a Buenos Aires con optimismo moderado para enfrentar al combinado de Argentina por la quinta jornada de las eliminatorias al Mundial del 2002.
(*The Ecuadorian soccer team, led by Colombian Hernán Gómez, will travel to Buenos Aires tomorrow with moderate optimism to face the Argentina team for the fifth round of the 2002 World Cup qualifiers*).
Word list (Spanish): [’selección’, ’dirigir’, ’colombiano’, ’viajar’, ’air’, ’enfrentar’, ’combinado’, ’jornada’]
Word list: [’team’, ’travel’, ’Colombian’, ’travel’, ’air’, ’face’, ’national team’, ’round’]
Media= 0.24
Label= Non-literal

Sentence 5:
Alrededor de 2.500 policías y otros numerosos agentes de seguridad nacionales y extranjeros deberán velar por la seguridad de los líderes y

al tiempo facilitar el tránsito de los ciudadanos para que *0* puedan gozar de sus propias celebraciones, como las numerosas fiestas infantiles organizadas por su Día Mundial .
(*Around 2,500 police officers and numerous other national and foreign security agents will have to ensure the safety of the leaders and at the same time facilitate the transit of citizens so that *0* can enjoy their own celebrations, such as the numerous children's parties organized by their International Day.*)
Word list (Spanish): ['policía', 'agente', 'velar', 'seguridad', 'facilitar', 'tránsito', 'gozar', 'celebración', 'fiesta', 'día']
Word list: ['police officer', 'agent', 'ensure', 'security', 'facilitate', 'transit', 'enjoy', 'celebration', 'part', 'day']
Media= 0.48
Label= Non-literal

The examples of sentences tagged as non-literal show the complexity of sentence extension and the need to establish rules to try to identify the most important interrelationships of words to evaluate the in semantic similarity. For that examples the verbs present in these sentences: "dirigir", gozar" and "apoderar", play a fundamental role in the semantic similarity analysis.

Most of the errors so far reported correspond to non-literal sentences labeled as literal language. For example, in the three sentences shown below, the phases: "ha jugado, un papel" and "se ha lanzado", that correspond to semantic innovations.

The second sentence is an important case because the word "ingenio" can be confusing even for a Spanish-speaking reader.

Let us see example of non-literal sentences tagged as literals.

Sentence 6:
Gobierno ha conseguido salvaguardar los intereses vitales de España dentro del proceso de construcción europeo, en el que *0* ha jugado un papel de animador del mismo, sostuvo *0* en una conferencia de prensa .
(*The Government has managed to safeguard the vital interests of Spain within the European construction process, in which *0* has played a role as its animator, *0* said in a press conference.*)
Word list (Spanish): ['gobierno', 'conseguir', 'salvaguardar', 'interés', 'proceso', 'jugar', 'papel', 'sostener', 'conferencia']
Word list: ['government', 'manage', 'safeguard', 'interest', 'process', 'play', 'role', 'say', 'conference']
Media= 0.55
Label= Literal

Sentence 7:
En la actualidad, un ingenio de similar poder destructivo podría ser transportado en un simple maletín de viaje.
(*Nowadays, a device of similar destructive power could be transported in a simple travel case.*)
Word list (Spanish): ['actualidad', 'poder', 'transportar', 'maletín']
Wod list: ['Nowadays', 'power', 'transport', 'travel case']
Media= 0.63
Label= Literal

Sentence 8:
Esta ha lanzado al mercado algunos programas, como el SimLife y el SimAnt, diseñados con las técnicas más revolucionarias de esta nueva ciencia.
(*It has launched some programs on the market, such as SimLife and SimAnt, designed with the most revolutionary techniques of this new science*).
Word list (Spanish): ['lanzar', 'mercado', 'técnica', 'ciencia']
Word list: ['launch', 'market', 'technique', 'science']
Media= 0.58
Label= Literal

Finally, as an experiment to show a small sample of the operation of the algorithm, the results of tagging 10 sentences of a randomly selection a a paragraph of CESS-ESP are presented and the tags obtained (TO) are compared against the expected tags (TE):

- AUNA es una nueva empresa "holding " creada por SIN , ET y UFINSA para el control común de estas cinco empresas de cable españolas.
  (*AUNA is a new holding company created by SIN, ET and UFINSA for the common control of these five Spanish cable companies.*)
  (TE= literal, TO= non-literal)

- Tras el examen preliminar de la documentación presentada, los servicios de competencia de la Comisión Europea estiman que la operación podría entrar dentro del campo de aplicación de la reglamentación comunitaria sobre concentraciones.
  (*After preliminary examination of the documentation presented, the competition services of the European Commission estimate that the operation could fall within the field of application of the Community regulations on concentrations.*)
  (TE= non-literal, TO= literal)

- La Comisión Europea invita a las terceras partes interesadas en esta operación a trasmitirle sus observaciones en el plazo de diez días sobre este proyecto de concentración.
  (*The European Commission invites third parties interested in this operation to send it their observations within ten days on this proposed concentration.*)
  (TE= literal, TO= literal)

- Brasil buscará a partir de mañana, viernes, el pase a su primera final de la Copa Davis ante los vigentes campeones, los australianos, y sobre la incómoda hierba del ANZ Stadium de Brisbane.
  (*The European Commission invites third parties interested in this operation to send it their observations within ten days on this proposed concentration.*)
  (TE= non-literal, TO= non-literal)

- Australia, actual campeona del torneo, intentará aprovechar su teórica ventaja en la hierba, en la que los luchadores jugadores brasileños, salvo el “todoterreno” Gustavo Kuerten, no se adaptan bien.
  (*Australia, current champion of the tournament, will try to take advantage of its theoretical advantage on grass, on which the struggling Brazilian players, except for the “all-rounder” Gustavo Kuerten, do not adapt well.*)
  (TE= non-literal, TO= non-literal)

- El capitán australiano John Newcombe no podrá contar con uno de sus mejores hombres, Mark Philippoussis, pero Lleyton Hewitt y un pletórico Patrick Rafter, que viene de ganar en Den Bosch y ser finalista en Wimbledon, aseguran un gran rendimiento para el equipo australiano.
  (*Australian captain John Newcombe will not be able to count on one of his best men, Mark Philippoussis, but Lleyton Hewitt and a bursting Patrick Rafter, who has just won in Den Bosch and been a finalist at Wimbledon, ensure a great performance for the Australian team.*)
  (TE= non-literal, TO= non-literal)

- Por su parte, el capitán brasileño, Ricardo Acioly, ha vuelto a confiar en Kuerten y Fernando Meligeni para enfrentarse al equipo “aussie”.
  (*For his part, the Brazilian captain, Ricardo Acioly, has once again trusted Kuerten and Fernando Meligeni to face the ”Aussie” team.*)
  (TE= non-literal, TO= non-literal)

- El doble que se enfrentará a la pareja Sandon Stolle-Mark Woodforde será el formado por “Guga” y Jaime Oncins.
  (*The double that will face the Sandon Stolle-Mark Woodforde couple will be the one formed by “Guga” and Jaime Oncins.*)
  (TE= literal, TO= literal)

- El sorteo ha deparado que los números uno de cada equipo, Rafter y Kuerten, abran el fuego en la primera jornada, mientras que Hewitt se medirá a Meligeni.
  (*The draw has led to the numbers one of each team, Rafter and Kuerten, opening the fire on the first day, while Hewitt will face Meligeni.*)
  (TE= non-literal, TO= non-literal)

- La otra semifinal enfrentará a España con Estados Unidos en Santander sobre tierra batida del 21 al 23 de julio , pero este fin de semana también se disputarán tres eliminatorias para la permanencia en el Grupo Mundial.
  (*The other semifinal will pit Spain against the United States in Santander on clay from July 21 to 23, but this weekend there will also be three qualifiers for permanence in the World Group.*)
  (TE= non-literal, TO= non-literal)

5 Conclusions

The present work has allowed us to fulfill the proposed objective: to automatically label the literal and non-literal language in sentences in Spanish. However, this work has shown the need to develop a new corpus of work in Spanish for the non-literal study and to consider the syntactic and semantic elements in the study of natural language.

Likewise, it also demonstrates the disadvantage of working with WordNet in Spanish due to the need to carry out translations or the ambiguity of some Spanish words; a relevant point is the limitations and lack of resources to work with PLN in Spanish. The strengths and weaknesses of the algorithm presented as well as its future projection have been reviewed.

For its part, this work has allowed experimenting between the semantic and syntactic interrelation to identify which types of words may be relevant and their different interrelationships.

On this topic, it is important to mention that sentence embeddings have a disadvantage when considering semantic similarity when working with Spanish sentences that contain subordinate clauses. The linguistic demands of Spanish suggest evaluating its particularities and adapting them to the development of the corpus.

As future work, it is considered relevant to complement this labeling and analysis with techniques such as machine learning, as well as the manual validation of each one of the labels, it is also considered important. test the performance of the algorithm with other types of corpora, perhaps with shorter sentences.

As a conclusion of this work, this algorithm is an important step in the conformation of the corpus in Spanish for the study of non-literal language and some results have been presented as part of the experimentation on this task.

Likewise, future work and the elements identified for it have been clear throughout the text, without a doubt, the need to develop new corpus and tools for NLP work in Spanish is very important, particularly in the study of non-literal language.

References

1. **Alsop, S., Nesi, H. (2009).** Issues in the development of the british academic written english (BAWE) corpus. Corpora, Vol. 4, No. 1, pp. 71–83. DOI: 10.3366/E1749503209000227.
2. **Birke, J., Sarkar, A. (2006).** A clustering approach for nearly unsupervised recognition of nonliteral language. 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 329–336.
3. **Birke, J., Sarkar, A. (2007).** Active learning for the identification of nonliteral language. Proceedings of the Workshop on Computational Approaches to Figurative Language, pp. 21–28.
4. **Burstein, J., Sabatini, J., Shore, J., Moulder, B., Lentini, J. (2013).** A user study: Technology to increase teachers' linguistic awareness to improve instructional language support for english language learners. Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, pp. 1–10.
5. **Calle-Rosingana, G. (2013).** Sentido literal y sentido figurado: Tratamiento cognitivo del Zeugma como estrategia estillística en la sombra del viento. Revista Española de Lingüística Aplicada, Vol. 26, pp. 91–106.
6. **Church, K., Hanks, P. (1990).** Word association norms, mutual information, and lexicography. Computational linguistics, Vol. 16, No. 1, pp. 22–29.
7. **Corpus del español (2023).** Corpus del español: Overview. https://www.corpusdelespanol.org.
8. **European Language Resources Association (2007).** CESS-ESP Spanish corpus. Syntactically and Semantically Annotated Corpora (Spanish, Catalan, Basque).
9. **Fenogenova, A., Kuzmenko, E. (2016).** Automatic generation of lexical exercises.

CEUR Workshop Proceedings, Vol. 1886, pp. 20–27.

10. **Fundéu BBVA (2023).** Proyecto Aracne. https://www.fundeu.es/aracne.
11. **Gao, G., Choi, E., Choi, Y., Zettlemoyer, L. (2018).** Neural metaphor detection in context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 607–613. DOI: 10.18653/v1/D18-1060.
12. **Graff, D., Gallegos, G. (1995).** Spanish news text. Linguistic Data Consortium. DOI: 10.35111/4dex-xm86.
13. **Kovecses, Z. (2010).** Metaphor: A practical introduction. Oxford University Press.
14. **Leech, G. (1992).** Corpora and theories of linguistic performance. Directions in Corpus Linguistics, pp. 105–126. DOI: 10.1515/9783110867275.105.
15. **Levy, M. (1997).** Computer-assisted language learning: Context and conceptualization. Oxford Linguisitcs, Oxford University Press.
16. **Meurers, D. (2012).** Natural language processing and language learning. Encyclopedia of Applied Linguistics, pp. 4193–4205. DOI: 10.1002/9781405198431.wbeal0858.pub2.
17. **Molino de ideas (2023).** Corpus Molinero. http://www.molinolabs.com/corpus.html.
18. **Ottolina, G., Palmonari, M., Alam, M., Vimercati, M. (2021).** On the impact of temporal representations on metaphor detection. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp. 623–632. DOI: 10.48550/arXiv.2111.03320.
19. **Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., Kircher, T. T. (2004).** Neural correlates of metaphor processing. Cognitive Brain Research, Vol. 20, No. 3, pp. 395–402. DOI: 10.1016/j.cogbrainres.2004.03.017.
20. **Shutova, E. (2010).** Models of metaphor in NLP. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 688–697.
21. **Sketch Engine (2023).** Learn how language works.

Article received on 06/06/2023; accepted on 20/09/2023. Corresponding author is Hiram Calvo.