An insulin pump? Identifying figurative links in the construction of the drug trafficking lexicon

Antonio Reyes† ‡ and **Rafael Saldívar**†
†Autonomous University of Baja California
School of languages
‡ Autonomous University of Queretaro
School of Languages and Literature
antonio.reyesp@uaq.mx, rafaelsaldivar@uabc.edu.mx

Abstract

One of the remarkable characteristics of the drug trafficking lexicon is its elusive nature. In order to communicate information related to drugs or drug trafficking, the community uses several terms that are mostly unknown to regular people, or even to the authorities. For instance, the terms jolly green, joystick, or jive are used to refer to marijuana. The selection of such terms is not necessarily a random or senseless process, but a communicative strategy in which figurative language plays a relevant role. In this study, we describe an ongoing research to identify drug-related terms by applying machine learning techniques. To this end, a data set regarding drug trafficking in Spanish was built. This data set was used to train a word embedding model to identify terms used by the community to creatively refer to drugs and related matters. The initial findings show an interesting repository of terms created to consciously veil drug-related contents by using figurative language devices, such as metaphor or metonymy. These findings can provide preliminary evidence to be applied by law agencies in order to address actions against crime, drug transactions on the internet, illicit activities, or human trafficking.

1 Introduction

Drug trafficking is a sensitive issue, apart from being a social taboo to some people. Unfortunately, this is a growing phenomenon that is impacting our lives on different layers. Our language is a sample of such impact. Nowadays, it is common to hear or read about drugs everywhere, but the words to name them are not necessarily the ones we are used to hear. Terms such as cocaine, marijuana, crack, or heroine have been replaced by new items, which at first glance seem to be totally unconnected to the context of drugs. Joy, candy, horse, or insulin are new labels used by the community to refer to drug-related contents. Some of them turn quite frequent in mass media and, consequently, in our daily communication; therefore, one can find them registered in specialized dictionaries or lexicons. Some others, on the other hand, are completely obscure to regular people, or even to the authorities.

In this context, the drug trafficking lexicon does not refer exclusively to the jargon to name drugs, but also to the terms used to refer to matters related to them. For instance, production of illegal substances (*colitas* (joint)), criminal gangs (*tacuaches*), or even, political speech (*sembrar* (use/fabricate false evidence)). In this respect, the drug trafficking lexicon is not only used by drug traffickers or by drug addicts. It has reached all social strata and is used by different actors. Furthermore, the drug trafficking lexicon underlines how this phenomenon permeates society through language: The fact of constantly being exposed to such lexicon makes this phenomenon something natural to everyone. Therefore, the consequences of violence, corruption, or institutional collapse derived from the use and sale of illegal drugs tend to be normalized.

Given this context, below we describe an ongoing research to identify drug-related terms by applying machine learning techniques. Our focus is on making explicit what people consciously (creatively) aim to veil regarding the drug trafficking lexicon. Specifically, we are interested in applying NLP techniques to identify figurative devices, such as metaphor or metonymy, as they are understood in Cognitive Grammar (see Langacker (1990)). To this end, we built a data set about drug trafficking in Spanish. This data set contains documents from different sources, such as press, blogs, song lyrics, or political speech. All of them were retrieved from Mexican sites; thus, the data set could be regarded as representative of the Mexican dialect and setting. The data set was used to train a simple word embedding model to identify links between the known terms and the ones created by means of figurative language. For instance, items such as cocaine, heroine, or drug (known terms)

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 118 - 124

share similar representations with the ones found in woman, or insulin (figurative terms).

With this study, we aim to provide preliminary evidence that can be applied by law agencies, for instance, to address actions against crime, drug transactions on the internet, illicit activities, human trafficking, among others.

The rest of the article is organized as follows: In Section 2 we describe and exemplify the notion of drug trafficking lexicon. Likewise, we provide a brief review about the scientific papers about the topic. In Section 3 we introduce the data set and detail the experiments that we carried out. In Section 4 we report the results and discuss the possible implications. Finally, in Section 5 we present the final remarks and some pointers to address the future work.

2 The drug trafficking lexicon

At first glance, the drug trafficking lexicon could be regarded as slang, or even as a non-standard vocabulary. Either way, it is common to think that it is only used by some isolated social groups. However, this type of language has gone from marginalization to daily speech in several countries. Mexico is a fair example of it. For instance, in the Mexican context, it is quite natural to hear in the news about *levantones* to refer to someone that has been kidnapped, and likely killed, by a criminal gang. In fact, the Dictionary of Mexican Spanish (El Colegio de México) has registered the term *levantón*, from the verb *levantar* (to lift up) as the action of kidnapping someone violently.

As noted from the example, the drug trafficking lexicon describes more than drug names. It used to depict a reality in which violence, corruption, and institutional collapse predominate; everything derived from the phenomenon of drug trafficking.

With respect to its features, it is necessary to specify that the drug trafficking lexicon is not a language properly; i.e. as far as it has been reported, it has no linguistic particularities to be considered an independent system. It is featured, on the contrary, by a set of lexical items (some of them neologisms) and phrases, whose meaning is often completely obscure to most people.

In this regard, the drug trafficking lexicon has compiled an interesting linguistic inventory, which has been fed from different sources, such as mass and social media, literature, political speech, and popular folklore. In this respect, in the field of Linguistics, some researchers have addressed their approaches from lexicographical perspectives. In particular, some of them have focused on the Latin American context. For instance, Acosta and Mora (2008) conducted a study about the criminal slang and drugs in Colombian prisons. They showed how such criminal jargon is characterized by a frequent use of linguistic devices, such as metaphors and metonymies (this seems to be repeated in the language of drug trafficking in Mexico, since the technical lexicon associated with crime is impregnated with metaphorical expressions (see Mattiello (2008)). More recently, in a research about the phenomenon of drug trafficking in the North American context, Saldívar (2022) described that one of the semantic fields in which this type of lexicon changes constantly is that of drug names. He stressed that this fact is evident in the creation of new terms, as well as in the reassignment of new meanings to the existing ones, both in English and Spanish. Likewise, Pressacco (2022) described violence and drug trafficking in Mexico from the so called narco language. The author distinguishes two classes to categorize this language: literal and figurative. Finally, she provides an interesting list of terms, phrases and constructions to exemplify the drug dealers argot.

In different locations and specialized areas, Sanmartín (1998) analyzed the jargon of the criminals in Spain. She described some linguistic mechanisms to characterize this lexicon, in particular, synonymy and polysemy. On the other hand, Torregrosa and Sánchez-Reyes (2015), in their study about English law enforcement, analyzed the use of conceptual metaphors related to drugs for educational purposes in the training of lawyers. Finally, in a computational approach, Reyes and Saldívar (2022) worked with narco language from a NLP perspective. They suggested a representation of narco-related concepts by identifying triggers of criminal content in corpus.

3 Unveiling figurative language

In this section, we firstly describe the data set used to build the word embedding model; then, we detail the processes to identify the figurative terms.

3.1 Data set

In order to train a vector model to represent the linguistic characteristics of this phenomenon, we gathered a specialized data set about drug traffick-

ing in Spanish. It is worth noting that, given the particularities of the topic (see Section 1), it is unlikely to find a large and public data set to be used. That is why we built a data set with documents of different genres to cover, as much as possible, a broad scenario about the topic. As we have previously pointed out, the documents come from Mexican sources only. This fact could be understood as a local application rather than a generic one. However, according to Bender and Friedman (2018) when explaining the notion of data statements, this reduction could provide the necessary context to allow the community to better understand how the experimental results could be generalized.

Table 1: General statistics per category.

Category	Tokens	Types
Blogs	229,338	29,585
Political	399,006	20,891
Essays	543,718	43,601
Literature	370,794	38,726
Narcocorridos	79,664	12,554
Press	728,165	83,514

The data set is divided in six categories according to the genres that we took into consideration to gather the documents: Blogs, political speech, essays, literature, *narcocorridos* (song lyrics about drug dealers), and press. Due to the linguistic differences across genres, the data set is imbalanced. For instance, with respect to the amount of documents, we collected a few set of texts for the categories *essays* and *literature*, compared to the amount of texts for the category *blogs* and *press*. This impacts on the size of each category. Thus, an essay about drug trafficking is more extensive (and elaborated) than a post in a blog; likewise, the specificity of information devoted to this phenomenon by the politicians in their speeches is completely different to the one reported by the journalists in the news. In addition, the amount of drug-related content is not necessarily the same across the six categories. For instance, compared to the *narcocorridos*, *literature* contains lesser specific information. This is due to the documents in the latter category are stories about drug trafficking within a narrative plot, while the former contains lyrics specifically written to drug dealers.

In order to balance the data set, we randomly select 50,000 words per category. In Table 1, we provide some statistics for each category based on the distinction type/token.

The data set is available upon request for academic purposes.

3.2 Word embeddings representation

In the past years, one of the most effective learning techniques employed in Machine Learning is word embeddings. They could be defined as representations of words in a vector space by grouping similar items (see Mikolov et al. (2013b)). This technique has been used to model linguistic information with excellent outcomes. For instance, Bakarov (2018) has explained that word embeddings are able to efficiently predict syntactic and semantic properties in natural language.

Almeida and Xexéo (2019) divide this technique into two main models: Prediction-based (local data models) and count-based (global data models). Although the use of word embeddings is growing in Machine Learning and other fields, some authors have reported a few drawbacks regarding their implementation in fine-grained tasks. One of the most important drawbacks is the unclear differentiation between semantic relatedness and semantic similarity (Bakarov, 2018).

Given the efficiency to represent linguistic properties, there are various word embeddings implementations. For instance, Word2Vec, FastText, or GloVe. In this study, we have adopted the Word2Vec algorithm, as described by Mikolov et al. (2013a,b).

The Word2Vec algorithm emphasizes the meaning and semantic relations between words by computing their co-occurrence in different documents. In this respect, Dessì et al. (2021) highlight that this algorithm is focused on modeling the context of words by exploiting ML and statistics in such a way the word vectors that share some regularities, regardless of the document they come from, are located nearly in the vector space. Therefore, the resulting representations allow the recognition of relatedness between words. This is why we have selected the algorithm to carry out the vector representation.

This algorithm can be trained using Continuous Bag-Of-Words (CBOW) or Skip-grams. We trained different models using the Skip-gram representation and modifying the vector dimension, window distance, and word frequency. It is worth mentioning that we also trained some models using a

CBOW representation in a preliminary setup; however, the outcomes were not as informative as with the skip-grams. Finally, in order to tune the vectors and come up with an integral model, we trained a final average model by finding the centroid of all the skip-grams representations. To this end, the Spearman's correlation coefficient was used to calculate the models' similarity (Hellrich and Hahn, 2016). The centroid model was used to run the experiments reported below.

3.3 Figurative terms identification

According to Saldívar (2022), apart from its cryptic-ity, one of the characteristics of the drug trafficking lexicon is its speed of change. This fact makes it elusive. Therefore, in order to identify the figurative terms, we firstly decided to use some known terms to build a dictionary. This resource groups items reported in the specialized literature as prototypical of the domain. Thus, they are used as seeds to identify a set of unrelated terms. For instance, a known term registered in the dictionary is *dinero* (money), this term is a seed to locate what others items appear close to it in the vector space. Some of the items are known terms (*morralla* (cash)), but there are others apparently unrelated (*cabezón* (big head)). The latter terms are the ones we are interested in, since they are likely figurative terms to refer to the known term. *Cabezón* is a metonymy to refer to the 100 dollars bills because Franklin's head in these bills is bigger. So, an utterance such as *Antes contaba morralla, ahora puros cabezones* (I was used to cash counting, now big heads counting only) makes sense both semantically and pragmatically.

The dictionary contains 439 terms. According tho the previous explanation, the first step was to look for the 439 terms in our data. Of those terms, only 183 appeared in our documents. The second step consisted in reducing the range of search. Thus, for each known term, we retrieved its 10 most similar words. This produces a total of 1,830 possible figurative terms to be analyzed. In Figure 1, we show the 10 most similar words for the term *coca* (abbreviation of cocaine).

In this figure, we can observe some known terms linked to the drug trafficking context: *Yerba* (marijuana), *cristal*, *chochos*, *ice* (cocaine), *heroína* (heroine), and *opio* (opium), which most people relate automatically to the drug trafficking lexicon. However, there are other items that, at first

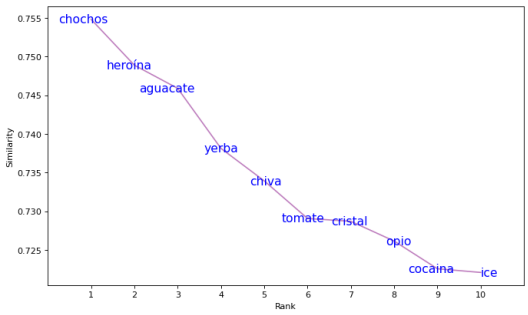

Figure 1: 10 most similar terms for the term *coca*.

glance, are totally unconnected to drugs: *tomate* (tomato), *aguacate* (avocado), or *chiva* (female goat). Initially, these items should be discarded due to they do not belong to the drug trafficking context. Nonetheless, given our interest in identifying figurative terms, they become the spotlight. If the vectors of these items are similar to the vectors of the known terms, then this could hypothetically be a sign about some semantic similarity.

In order to confirm this hypothesis, we focused on retrieving the vectors for each unknown term in such a way we could map the figurative usages. For instance, considering the information depicted in Figure 1, we first removed the known terms (*yerba*, *cristal*, *chochos*, *heroína*, *opio*, and *cocaína*); then, we retrieved the vectors for the unrelated terms (*tomate*, *aguacate*, and *chiva*). Finally, given the seed term (*coca* in this example), we mapped the known term and the unknown terms considering their distributional patterns in the vector space. In Figure 2, we show the 10 most similar terms for the presumably unrelated terms *aguacate* and *chiva*.

4 Results

The result of the previous processes is a set of 505 drug-related terms; i.e. an average of 3 unrelated items per known term.

Subsequently, we analyzed the vectors of the 505 candidates in order to recognize elements to connect them to the drug trafficking context. This is clearer if we observe Figure 2: From the 10 most similar words for *aguacate*, the items *churros* (marijuana), *mulas* (drug trafficker), *tachas* (cocaine), and *fuman* (inflectional form of the verb to smoke) are totally drug-related. The same fact for *chiva*. The words *inyecto*, *meto* (inflectional forms of the verbs to inject and to do drugs, respectively), and *chochos* are commonly used by the community to refer to drugs. This fact corroborates that some

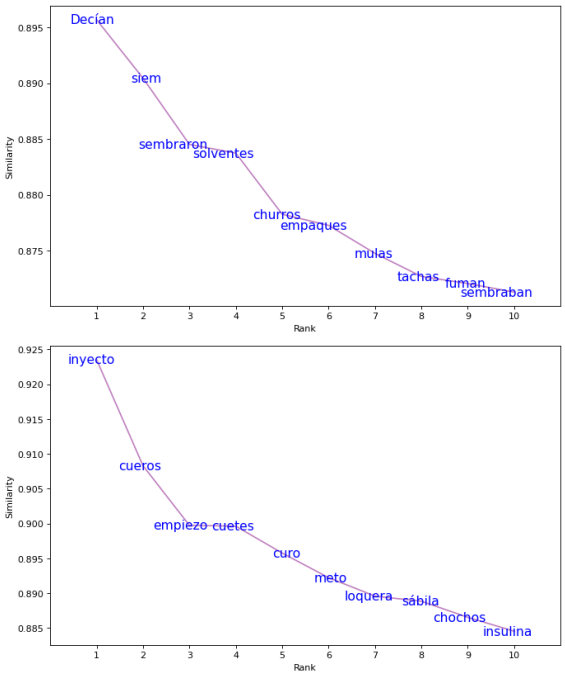

Figure 2: 10 most similar terms for *aguacate* and *chiva*.

of the unconnected items are closely linked to the context of drugs. However, there were others items whose vectors proved the contrary. For instance, *tomate*, from Fig. 1, whose 10 most similar words referred to food only.

Given this result, it could be stated that some unconnected items are, in fact, figurative terms to implicitly express drug-related content. Nonetheless, such assumption should be assessed by the experts; i. e. we could identify some unrelated terms regarding drug trafficking; however, we are not capable of saying that they mean anything to the community. In addition, there is not labeled data to compare our findings. Therefore, in order to provide arguments to validate our findings, we contacted an expert on the topic in Mexico. This expert has published several academic papers and some books about narco and, specifically, about narco language.

Prior to contacting the expert, we grouped the unrelated terms in clusters with the purpose of identifying an underlying semantic structure. To this end, we ran a similarity analysis for the 505 terms considering their co-occurrences in the whole data set. Figure 3 shows a sample of the clusters presented to the expert.

4.1 Evaluation

Once we contacted the expert, we asked him to revise the 505 terms to check whether or not they are terms used in a drug-related context. If so, to confirm, as far as he knows, whether or not they are used to refer or name any term cryptically.

The feedback provided by the expert is summarized as follows: With respect to the first task, he validated all the 505 terms as part of the domain. However, regarding the second task, he marked only 151 terms as terms used to refer to drug-related content in a cryptic manner; i. e. around 70% are already terms in usage in that context, although we did not know (it is worth stressing that we are not part of the community), and only 30% are items, cryptic enough, to be considered figurative terms. In addition, the expert provided the equivalents for the unknown terms. For instance, terms such as *cuete* (gun) and *insulina* (insulin), or *yongo* (yongo) were translated to syringe and place to do drugs, respectively[1].

4.2 Discussion

The feedback given by the expert confirmed that this approach is identifying drug-related terms efficiently. Although some of the 505 terms are already known in the domain, their usage is not very frequent to be registered in some lexical resource. For instance, they are not part of the terms we used to build our dictionary (see 3.3). In this respect, this is evidence about the dynamism of any language. The drug trafficking lexicon, in particular, must be very dynamic due to it expresses outlaw issues mainly.

With respect to the 151 cryptic terms, they confirm such dynamism. They are obscure enough to be able to determine what they mean in the drug trafficking context. However, beyond the fact that they can be considered as figurative terms, it is necessary to identify what kind of figurative device underlies them. In this regard, we are in the process of manually analyzing the linguistic contexts of the 151 terms in order to recognize patterns to explain their usage in this domain. Nonetheless this is work in progress, we have noticed that some of the terms, in fact, rely on figurative language to create an implicit link between the unrelated terms and the known term. For instance, considering the information depicted in Figure 3, terms

[1]To better understand the information given in Figure 3, we provide the translations (not the equivalents) for the known terms in each cluster: *aspirina* (aspirin), *cobija* (blanket), *raya* (line), *cajuela* (trunk), *cocinar* (to cook), *polvo*, *polvito* (dust), *hierba* (grass), *dulce* (candy), *hielo* (ice), *nieve* (snow), *cristal* (glass), *hielera* (icebox), *narcomensaje* (narco-message), *sabanas* (sheets), *goma* (gum), *enteipar* (to apply masking tape on someone), *arete* (earring), *bajón* (downer).

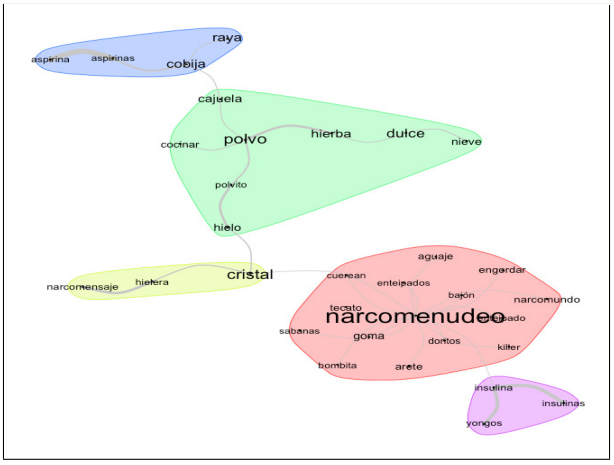

Figure 3: Sample of clusters given the similarity analysis.

such as *nieve* (referring to the drug named crystal) or *insulina* (referring to a syringe and/or the action of being injected) in the green and purple clusters, respectively, are understandable if we assume a metaphoric and metonymic frame. Thus, the term *nieve* (snow) is metaphorically mapped to *polvo* (powder) and then to *cristal* (crystal). First, a feature like the color is the link to secure the comparison. Subsequently, a component (*polvo*) of the whole (crystal) is used to connect to the same drug by profiling its rock-like appearance. Something similar happens to the second term. The *insulina* (insulin) is a legal drug to be injected into the diabetic patients. Many drugs are supposed to be injected to enhance the effect. Therefore, by using this term, the speaker is metonymically connecting a component (syringe) of the whole action (to be injected with one of such drugs).

It is also necessary to highlight that not all the terms can be explained by means of an underlying figurative device. There are terms that appear from other linguistic mechanisms. However, as have been reported by some experts on the topic, figurative language (especially metaphor, metonymy, and analogy) is quite frequent to create terms within the domain, either for cryptic purposes, attenuation, or as a simple exercise of creativity (see (Saldívar, 2022; Torregrosa and Sánchez-Reyes, 2015; Mattiello, 2008)).

5 Conclusions and future work

In this study we have approached an unusual phenomenon in Natural Language Processing: the drug trafficking lexicon. Our focus was on automatically identifying possible figurative terms to refer to drug-related contents in Spanish. To this end, we used a data set about drug trafficking in Mexico and built a word embedding model to identify the terms. The results showed that the model could identify a set of supposed unrelated terms to the domain. Those terms were validated by a human expert; however, only 30% of them are cryptic terms. This means, although they are known by the community, people out of the drug trafficking context do not know them. Therefore, they can be used to veil criminal content. Finally, we have outlined a possible explanation about their successful usage within the domain. In this respect, although this is still work in progress, we have suggested that this can be explained in terms of figurative devices, such as metaphor and metonymy, which according to the Cognitive Linguistics foundations (Langacker, 1990), are part of our conceptual structure.

As future work, it is planned to collect data from other variants to extend the scope of this approach, as well as to deepen the analysis of the linguistic mechanisms to better understand how this lexicon works to successfully communicate veiled information within a complex linguistic system. Thus, the insights could shed light on how this social phenomenon has linguistically permeated our society in broader terms. To conclude, we consider that works like this one could provide evidence to be applied to address actions against different illicit activities.

Limitations

Some of the limitations of this study rely on the data. As mentioned in the manuscript, the data set built to carry out the experiments was gathered considering only one specific dialect. Although the insights could be representative of the phenomenon, they cannot be generalized to the entire system. This is manly due to the social particularities of drug trafficking. For instance, the drug names or gangs depend on social elements extracted from the culture. However, the underlying figurative mechanism is consistent from dialect to dialect, and from language to language, as reported in the literature (see Lakoff (1987); Lakoff and Johnson (1980); Langacker (1987); Goldberg (1997), and others). Another issue regarding the data is the lack of labeled data to compare the results, as well as to use them to prove how our approach performs. In this regard, it is worth highlighting that topics like this one can represent a major challenge when collect-

ing data in some countries. Given the corruption and lawless of some governments, it could be very risky to find proper data and collect a representative corpus.

The human validation is also a limitation. It is unusual to have only one vision to validate the outcomes; however, it is very difficult to find specialists about the topic. This impacts on the number of available experts to assess our findings.

Finally, the manual analysis of the results must be concluded in order to provide a complete description of the figurative devices present in the data. In addition, although we have focused on the figurative terms, we notice that several of the known terms were generated by means of figurative language. Therefore, they should be explained to present a more comprehensive description of the phenomenon.

References

D. Acosta and C. Mora. 2008. Subcultura carcelaria. Diccionario de la jerga canera. *Escuela Penitenciaria Nacional*.

Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *CoRR*, abs/1901.09069.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *CoRR*, abs/1801.09536.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Danilo Dessì, Diego Reforgiato Recupero, and Harald Sack. 2021. An assessment of deep learning models and word embeddings for toxicity detection within online textual comments. *Electronics*, 10(7).

El Colegio de México. Diccionario del español de México. `http://dem.colmex.mx`. Online on February, 2022.

A. Goldberg. 1997. Construction grammar. In E.K. Brown and J.E. Miller, editors, *Concise Encyclopedia of Syntactic Theories*. Elsevier Science Limited.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

G. Lakoff. 1987. *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago.

G. Lakoff and M. Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

R. Langacker. 1987. *Foundations of Cognitive Grammar*. Stanford University Press.

R. Langacker. 1990. *Concept, Image and Symbol. The Cognitive Basis of Grammar*. Mounton de Gruyter.

E Mattiello. 2008. An introduction to english slang: a description of its morpholgy, semantics and sociology. *Polimetrico*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Coralie Pressacco. 2022. *La violencia del narcotráfico en México. Análisis lexicológico*. Universidad de Colima.

A. Reyes and R. Saldívar. 2022. Figurative language in atypical contexts: Searching for creativity in narco language. *Appl. Sci.*, 12, pages = 3: 1642, note = DOI: 10.3390/app12031642.

R. Saldívar. 2022. Metáforas y metonimias conceptuales en nombres de drogas en inglés y en español. *Forma y Función*, 35(1).

J Sanmartín. 1998. *Lenguaje y cultura marginal. El argot de la delincuencia*. Universitat de Valencia.

G Torregrosa and S Sánchez-Reyes. 2015. Raising metaphor awareness in english for law enforcement. *Procedia Soc. Behav. Sci.*, 212:304–308.