THE REPRESENTATION OF MIGRANTS IN SPANISH JUDICIAL DECISIONS: USING CORPUS DATA TO REFUTE HATE SPEECH

HOW TO CITE: Marín, M.J., Almela Sánchez-Lafuente, A. (2022). The representation of migrants in Spanish judicial decisions: using corpus data to refute hate speech. *Corpora*, 17 (2): 167–196.

Abstract

The phenomenon of immigration and its depiction in media texts have been examined profusely within the field of corpus-based discourse analysis (Gabrielatos and Baker, 2008; Baker *et al.*, 2013; Blinder and Allen, 2016). This research seeks to present it as reflected on a corpus of 600 judicial decisions issued by Spanish courts in the years 2016 and 2017.

This analysis was motivated by the rise of extreme right-wing parties in Europe in the recent years, which dehumanise immigrants and portray them as a threat to the welfare state. On a first approach, the results appear to dissociate immigration and crime since a considerable percentage of the keywords obtained (c. 20%) revolves around three major topoi, namely, *family, territory/access,* and *legal punishment,* not showing evidence of any major offences or crimes amongst the top-ranking lexicon.

The study of the collocate networks of the KWs within the category *legal punishment* confirms our initial perception, in fact, out of 21 collocates, only the word *delito* (*crime*) itself collocates with terms referring to typified crimes such as *violencia* (violence).

In parallel, the data were triangulated using the text-classification software *UMTextStats* (García-Díaz *et al*., 2018). The results of this second analysis confirm our initial observations.

Keywords*: migrants, corpus-based discourse analysis (CBDA), legal English, hate speech, keyword (KW) analysis*

1. Introduction

In recent years, Europe has witnessed the rise of extreme right parties, which have gained considerable parliamentary representation, moving from insignificant figures to becoming major political forces. In France, Le Pen's North Front obtained 34% of the votes in 2017 whereas the German *AfD* got 12.9% support from the voters in the same year and entered the German parliament for the first time. The results for the Italian extreme right were similar and Salvini's *Liga* gained 124 representatives (19% of the whole parliament). In Spain the situation is similar, since VOX, the Spanish far-right party, became the third most voted one in the 2019 election, gathering 3.64 million votes (15.09%)[1] and gaining 54 representatives at the parliament for the first time in Spanish history.

Immigration is one of the major arguments upon which these parties construct their speech, dehumanising the figure of migrants and presenting them as a threat to the welfare state. Donald Trump's speech and policies in this respect are used as a reference by the European far right. As Crawford states when analysing Trump's speech, the

[1] Statistical information retrieved from: https://www.rtve.es/noticias/20191111/vox-extrema-derecha-europa/1989668.shtml

rhetoric of dehumanisation revolves around images of migrants depicted as 'congenital criminals, lepers, thieves, unclean, garbage, animals (…), which denies them the dignity, consideration, compassion and empathy that we typically give other people' (2019: 2). Such arguments are pivotal to hate speech, which seeks to put migrants at the core of the political debate and to lay a smoke screen over other fundamental issues such as public services (education and health), economy or security.

Santiago Abascal, the leader of the Spanish far-right VOX, in the electoral debate held in November 2019, put special emphasis on migration policies and, amongst other strategies, related the figure of migrants to crime rates. By doing so, he attempted to picture migrants as a potential threat to 'legitimate' Spanish citizens. Migrants were accused of major crimes and the official figures were distorted to the extent of presenting false data, as put forward by Wodak (2105: 23), who points at the mechanisms to construct fear by means of proposing migrants as scapegoats "that are blamed for threatening or actually damaging our societies, in Europe and beyond".

Nevertheless, the association between the figure of migrants and crime could be refuted by simply consulting the official statistics provided by the Spanish Ministry of Justice, where crime rates are divided into different categories and linked to the nationality of those who commit such crimes. However, what do judicial decisions say about the relationship between crime and immigration? Could the statistical data from the Ministry of Justice be supported by those extracted automatically from a corpus of legal texts? This research aims at exploring a corpus of 600 judicial decisions related to immigration and issued by Spanish courts between the years 2016 and 2017 to find linguistic evidence capable of dissociating the concept of immigration from that of crime and contesting the arguments proffered by the Spanish far-right.

To that end, a legal corpus was processed automatically using software tools such as *Wordsmith 8.0* (Scott, 2020a) which allowed us to identify the corpus keywords, that is, those words which are statistically more relevant in the specialised text collection by comparison with a 100 million-word reference corpus of general Spanish (Reese *et al*., 2010). This software facilitated the automatic extraction of the most relevant terms in our corpus using *log-likelihood* (Dunning, 1993; Rayson & Garside, 2000). The list of KWs (keywords), once filtered as described below, acted as a point of departure for the closer scrutiny of the collocate networks of the most significant terms found in it. Wordsmith offers users the possibility of obtaining the list of collocates of a given term, nonetheless, Brezina *et al.*'s *Lancsbox* (2015) favours the task greatly by visualizing the collocate networks and allowing the user to expand a term's context to further collocational levels other than merely the term's most frequent immediate collocates, as *Wordsmith* does.

Lancsbox (Brezina, McEnery and Wattam, 2015), which includes several applications such as *Graphcoll*, was therefore used for the examination and visual implementation of the collocate networks of those terms related to the thematic categories which KWs were classified into. Such scrutiny led to the identification of major thematic areas which depart considerably from the idea that immigration is closely linked to crime, as supported by the far right. It was rather observed, in line with the findings by other atuhors like Baker *et al.* (2008), Pérez-Paredes *et al.* (2017) or Alcaraz Mármol and Soto-Almela (2016, 2018) that the major concerns for migrants entering Spain relate to the migration processes itself as well as to their family environments.

In parallel, the data were triangulated using the text-classification software *UMTextStats* (García-Díaz *et al*., 2018). Its structure is similar to the well-known text analysis tool Language Inquiry and Word Count –LIWC (Pennebaker and Francis, 1999)–, yet it includes a linguistic basis of European Spanish and several categories that are not word-based. *UMTextStats* approaches language study from a different perspective

as vocabulary items are mapped onto morphosemantic categories through the implementation of psycholinguistic criteria. In spite of the diverging nature of the tools used to process the corpus, the results obtained using *UMTextStats* point in a similar direction to the observations made above, since the representativeness of categories such as crima or violence is practically inexistent.

In this research paper, foundational information including a review of corpus linguistics approaches to the phenomenon of immigration is presented in Section 2, followed by Section 3 where some details about our data collection, corpus tools and data analysis are offered. Next, Section 4 provides the discussion of our results, and Section 5 presents our conclusions by restating the main points of our study.

2. Literature Review

The description of legal language has traditionally been accomplished by reputed specialists (Mellinkoff, 1963; Alcaraz, 1994; Tiersma, 1999; Borja, 2000) whose conclusions, deriving from deep knowledge and extensive expertise in the field, are usually based on intuition or on relatively reduced samples of texts. Their work often presents a top-down characterisation of the major traits of *legalese*, following a deductive approach whereby the rule usually precedes the actual description of the examples provided.

Legal genres have been described across systems and their numbers vary depending on the perspective of analysis. Judicial decisions or sentences, which form the corpus employed herein, appear in generic classifications as part of the oral mode (Danet, 1980), within the category "recording and law making" (Maley, 1994), or as public unenacted law (Orts, 2009), amongst others. Their relevance within the variety is fundamental both in civil and common law systems. Judicial decisions or sentences, which constitute the so-called jurisprudence, stand as a major source of law in the Spanish legal system (Calvo Vidal, 1992). Their form and content make them particularly prone to linguistic analysis since they not only display their own specific language features but may also touch upon other legal genres like statutes, wills, or contracts, hence their representativeness.

As regards the study of legal English, in recent years, there has been a growing tendency towards corpus-based and corpus-driven[2] descriptions of language varieties, which provide a bottom-up characterisation of legal English (Author, 2012; Biel and Engberg, 2013; Pontrandolfo and Goźdź-Roszkowski, 2014; Breeze, 2015), yet, there is still a clear need for a greater number of these studies. As a matter of fact, corpus-driven studies are capable of unveiling specific features of language which might otherwise remain unnoticed. This is the case of Author's (2019) work on the expression of appraisal in judicial decisions, a legal genre that is generally assumed to leave little room for subjectivity. Nonetheless, as proven by textual evidence, it was found that a relevant proportion of vocabulary items expressing appraisal was present in two legal corpora obtained from Spanish and British sources.

Immigration in public discourse (mostly in the media) has been examined with the aid of corpus linguistics techniques. The work by Baker *et al.* (2009; 2013) is seminal in this respect. In Baker *et al.* (2013) we are presented with the portrayal of Muslims in the British press from a longitudinal perspective. This is achieved through the

[2] In corpus-based linguistic studies, a query is formulated in advance so as to find evidence in a corpus, whereas corpus-driven analyses base their conclusions solely on linguistic findings obtained from corpora and adopt an inductive approach to language description.

identification of the major topoi which the collocates of the keywords obtained revolved around. One of the major concepts revealed by their analysis was that Muslims were frequently constructed in terms of homogeneity and connected to conflict (2013: 275), being presented as prone to taking offence and dangerously connected to radicalism. In Baker *et al*. (2008), the authors introduce a novel approach to critical discourse analysis based on the triangulation of data both through a quantitative and a qualitative examination of two sets of texts which, in the end, offered different though harmonising perspectives on the same issue, immigration.

As Egbert and Baker put it (2020), methodological triangulation in linguistic research including corpus data is experiencing an upward trend. Over the last few years, a wide array of specialised text analysis software has been used in empirical linguistic research, which brings considerable benefits to the humanities in general, and to corpus linguistics in particular. Accordingly, confirmation bias can be mitigated by combining several research methods and empirical materials in the study of the same linguistic phenomenon. Thus, researchers' metalinguistic awareness, that is to say, the ability to think about how language is used to convey meaning and what that meaning is, must be informed by reliable research methods in order to minimize potential weaknesses in research design and to try to avoid confirmation bias.

In connection with the work by *Baker et al*., Taylor (2014) introduces a comparative analysis of two corpora of Italian and British newspaper articles. She observes a clear tendency to connect immigration and crime in the Italian corpus, a typical idea sustained by the European far-rights, as already stated. On the contrary, the British text collection does not convey that idea so explicitly, probably due to the fact that the British corpus did not comprise any texts from tabloids, more prone to sensationalism.

In line with previous research, Alcaraz-Mármol and Soto-Almela (2016) explore the semantic prosody of the words *inmigración* (immigration) and *inmigrante* (immigrant) in a corpus collected from two Spanish newspapers with different political ideology. Consistent with the results in Taylor (2014), the authors find that most of the words co-occurring with the lexical items under study have a negative meaning, thus confirming the feeling of rejection towards the phenomenon of immigration in the Spanish press. Similarly, Alcaraz-Mármol and Soto-Almela (2018) find that a related term with negative semantic prosody in the Spanish press is *refugiado* (refugee). Specifically, the authors study that lexical item from a diachronic perspective over a 7-year period, concluding that the negativity associated with ***refugiado*** has grown in the last two years under study, in parallel to its increasing frequency in the media.

On a different note, only Pérez-Paredes *et al*. (2017) and Sánchez *et al.* (2019) have explored legal texts (UK and Spanish legislation and official information) **to probe into** the view which is offered in the official documents issued by state institutions on such a complex phenomenon as immigration. They come to the conclusion that, on the one hand, UK legislative and informative texts appear to depict migrants (the term *immigrant* tends to be avoided) as citizens whose integration bears almost no relationship with the image projected of them on the information published by official institutions, which presents them as subject to control processes (Pérez-Paredes *et al*., 2017). On the other hand, Sánchez *et al.* (2019: 88) acknowledge the attempt by the Spanish administration to favour the integration of immigrants through the usage of terms like *ciudadano* (*citizen*), *personas* (*people*) or *población* (*population*) in association with the idea of *immigrant*.

Nevertheless, to the best of our knowledge, judicial decisions have not been scrutinised in search of linguistic evidence on the construction of immigration from a legal perspective. This research was therefore conceived to try and bridge this gap.

3. Methodology

3.1 Corpus Description

As shown in Table 1, a Spanish corpus of judicial decisions of 600 texts was employed in this study. It contains 2.4 million tokens and 20,236 types. The source used to obtain the texts, produced between 2016 and 2017, was *CENDOJ*[3], the Spanish legal documentation centre.

A reference Spanish corpus was also processed to automatically identify the keywords in the legal corpus. It is a roughly 100 million-word text collection of Wikipedia articles, as illustrated in Table 1, distributed in 94 texts. This is just a section of a larger corpus, the *Wikicorpus* (Reese *et al*., 2010), available online and downloadable in plain text format[4]. The rationale behind the choice was the need to resort to a large text collection so as to compare our legal corpus with one from the general field. The Wikicorpus seemed particularly fit for this purpose since it was considerably larger than the legal one under examination and it contained articles on various issues such as history, science, medicine or literature, amongst many others, covering a plethora of language areas other than the legal field.

Table 1. Corpora description

The selection of the texts in the legal corpus was carried out using the search engines offered by the *CENDOJ* website, which allows for advanced searches. The search terms included *inmigración* (immigration), *inmigrante* (immigrant), *extranjero* (literally *foreigner*) and *extranjería* (this term is used in Spanish to refer to the laws and regulations on immigration, for instance, the *Ley de Extranjería* 4/2000 —Immigration Act 4/2000—). In the Spanish system there are no specific courts exclusively dedicated to trying cases of asylum requests or irregular migration as we find, for instance, in the United Kingdom, therefore, the texts were retrieved randomly from any court at any level within the judicial pyramid by introducing the search terms above.

The search engine provided by the CENDOJ website is a rather sophisticated one which facilitates the search by offering the possibility to consider only one area of law (criminal, social, civil, etc.), one specific type of court (from the lowest local courts to the supreme court) or a given type of sentence (auto v. sentencia in the Spanish system) amongst many other options. None of them was activated to avoid any kind of bias that might have skewed the results, this way, the texts obtained might have been produced by any judge at any court within any jurisdiction. The only activated option was the time span, which was set between 1st January 2016 and 31st December 2017. Such settings led to the identification of 667 sentences, out of which 600 were randomly selected. As for the search terms, they were all introduced simultaneously, that is, all the texts retrieved by the search engine might include one or more than one of these terms. The terms *extranjero* (*foreigner*) and *inmigrante* (*migrant*) were the most frequent ones ranking 180th and 581st respectively in the frequency list once the corpus was processed.

[3] http://www.poderjudicial.es/search/indexAN.jsp

[4] https://www.cs.upc.edu/~nlp/wikicorpus/

3.2 Corpus Tools and Procedure of Analysis

3.2.1 *Wordsmith Tools 8.0*

Regarding the first software used, *Wordsmith tools,* keywords (KWs) has been the tool used in order to unveil the most significant themes in the corpus. For this research, the frequency threshold was set at 5, that is, a word had to occur at least five times in the corpus to be considered for the comparison, and the measure implemented for their identification was *log-likelihood* (Rayson and Garside, 2000; Dunning, 1993). Table 2 illustrates the top 20 lexical KWs (function words were filtered) obtained from the legal corpus, which were used as the basis for the analysis implemented below. The table also displays their raw frequency in the specialised corpus and their keyness value, used to rank them.

Table 2. List of lexical keywords

3.2.2 *Lancsbox*

Closely linked to the automatic identification of keywords is the relevance of other words which tend to co-occur with them, that is, their collocates. Collocational patterns reveal the context in which a word occurs and provide plenty of information about the meanings and connotations associated with a word in context. Nevertheless, for the identification of collocational patterns in a text collection, especially if it is a large corpus, it is necessary to employ automatic tools that facilitate the task. Let us first define and consider some theoretical questions.

Broadly speaking, in Firth's words, a collocate is the company a word keeps (1957: 6). The concept collocation has been revisited since then (Cruse, 1986; Gries, 2013; Sinclair, 1991; Stubbs, 2001) and more specific and accurate definitions have been provided, Sinclair (1991; 2005) deems the statistical data associated with two co-occurring words as fundamental for their identification, as collocates can be mined automatically by applying measures of association like mutual information (Church and Hanks, 1990) or *log-likelihood* (Rayson and Garside, 2000; Dunning, 1993), amongst others.

However, as Baker (2016) acknowledges, the study of collocates has been limited to the analysis of word pairs until recently, often due to the limitations of tools like *AntConc* (Anthony, 2014) or *Wordsmith* (Scott, 2020a), only capable of extracting pairs of collocates, disregarding the potentiality of collocational or lexical networks (Williams, 2001) in the study of the interaction amongst terms and their vicinity in a corpus. Williams' (2001) idea that collocational or lexical networks may enhance quantitatively and, above all, qualitatively our understanding of specialised vocabulary meant a step forward in the study of term usage and meaning and authors like Baker (2005; 2016), McEnery (2006) or Author (2016) acknowledge this fact. Yet, in spite of the above, the process undergone in the production of lexical networks could be time consuming, requiring the manual arrangement of the networks (often populated by thousands of elements), since automatic corpus tools only allow for the study of one collocational level.

The analysis and interpretation of lexical networks has been envisaged from different angles. On the one hand, there is a large body of research within the field of Natural Language Processing (NLP). These studies often establish a link between the concept of lexical networks and that of neural networks where lexical hierarchies are established by automatic systems often aimed at word sense disambiguation (WSD). The work by Barceló-Coblijn et al. (2017), Stuart and Botella (2009) or Guilquin (2008) illustrate this trend. As a matter of fact, there exists a plethora of tools capable of processing electronic texts designed with different purposes, although not many of them can obtain the lexical networks of a term automatically. This is the case of the software package *Lancsbox* (Brezina *et al*., 2015), which was specifically designed to that end. One of the advantages of using *Lancsbox* is that it not only manages to obtain a word's network quickly, but it also visually represents its network through a graph that displays the node's collocates, connecting them with vectors whose size varies according to the strength of the collocational bond calculated by the tool (the shorter the vector, the stronger the link between words) and indicating collocate directionality. The data associated with each of the constituents of a term's lexical network can also be read in detail and saved in .csv format[5].

Corpus-Based Discourse Analysis (CBDA) has benefitted greatly from the use of tools like *Lancsbox*, as well as data-driven learning (DDL). As for the former, we find a plethora of studies like Baker's (2018), who delves into the connection between language and sexuality and highlights the advantages of using tools like Lancsbox, also questioning the need to go beyond the mere data and to explore texts to provide an insightful, unbiased interpretation of these data. Brezina (2018) also emphasises the usefulness of the tool, capable of unveiling fundamental aspects of the semantic structure of a text and its multiple layers of meaning. Some of the most relevant keywords found in our legal corpus were analysed using this tool, as illustrated by Fig. 1 below, which shows the network of *multa* (*fine/penalty*) and will be examined in greater detail in the forthcoming sections.

Nevertheless, in spite of the multiple applications of lexical/collocational networks, the scarcity of research on legal texts from lexical network and CBDA perspectives is manifest. This study therefore seeks to fill that gap.

3.2.3 *UMTextStats*

Lastly, the study of morphological and semantically meaningful categories in corpora has proved highly valuable as well. Judging from our experience, some Natural Language Processing (henceforth, NLP) tools can be used together with more standard Critical Discourse Analysis methods like keyword or collocate analyses, as suggested by Bisceglia, Calabrese, and Leone (2014), and Jumaquio-Ardales, Oco, and Madula (2017). Indeed, this can be considered as an important methodological innovation of the present study. Thus, after examining KWs and some of their collocates, the data were triangulated by virtue of this further analysis. The first and best-known text analysis tool of this nature is *Language Inquiry and Word Count –LIWC* (Pennebaker and Francis, 1999). This software was developed to provide an efficient method for studying English language psycholinguistic concerns. Specifically, the categories used were related to standard linguistic processes, psychological processes, relativity, and personal matters; a detailed description of the individual categories can be found in Pennebaker and Graybeal (2001). It has also been adapted and translated into more than ten languages, including Spanish (Ramírez-Esparza, Pennebaker, García, and Suriá, 2007). It provides an effective tool for

[5] The extension .csv stands for 'comma separated values', which can be easily imported into an excel spreadsheet.

studying the emotional, cognitive, and structural components contained in language on a word-by-word basis, working out the percentage of words which fall into those categories. Over the last few years, it has been widely used in fields like sentiment analysis (Salas-Zárate *et al*., 2014), forensic linguistics (i.e. Mihalcea and Strapparava, 2009), and psycholinguistics (i.e. Hancock *et al*., 2011).

In the light of those results, the text-classification software *UMTextStats* (García-Díaz *et al*., 2018; García-Díaz, Cánovas-García, and Valencia-García, 2020) has been built on a similar technology. However, as compared to LIWC, it brings two major advantages: a linguistic basis of European Spanish, and several categories that are not word-based. In this novel tool, developed at Universidad de Murcia, the input is a set of natural language texts and the result is a vector consisting of different features; thus, it is especially appropriate for automatic classification experiments, as the resulting values can be used for the training of different machine learning classifiers. This NLP tool currently contains 112 Spanish dictionaries that comprise more than 50,000 word stems and regular expressions. Each word stem can be mapped to 125 different features, amongst which grammatical information such as total of pronouns, articles, negations, and auxiliary verbs is offered, as well as emotions, named entities, and cognitive processes, to name but a few examples of the psycholinguistic categories.

It is worth noting that in the dictionaries used by the software, lexical items have been formalised by means of regular expressions, that is to say, search strings that can be used to specify sequences of characters to be extracted from a text or corpus (Jurafsky and Martin, 2018). Thus, for instance, *doméstico/a/os/as* (*domestic*) has been formalised as *doméstic[oa]s*?, which is interpreted by the software as the string of characters *domestic-* followed either by *-o* or *-a*, and after that sequence, an optional *-s*. Some other examples comprise broader possibilities, such as the regular expression *abraz\w**, which matches the string *abraz-* followed by any repetitions (*) of any alphanumeric character *(\w*), allowing for the retrieval from the corpus of the whole verbal conjugation of *abrazar* (*to hug*), the noun *abrazo(s*) (*hug/s*), or, in general, any word built on the stem *abraz-*.

As *UMTextStats* is currently at the beta testing stage, we have been able to make the most of it by including an ad-hoc category, namely *crime*, comprising lexical items such as *homicidio*(s) (*homicide*), *asesinato*(s) (*murder*), *lesiones* (*injuries*), and *amenazas* (*threats*). All of the items included in this category are typified in the Spanish Criminal Code and they are also reflected on official crime statistics, hence their adequacy.

Taking all these features into consideration, this software was selected as an alternative method of analysis of the data which might contribute to the triangulation of the results. The psycholinguistic criteria applied to its design allow for the creation of morphosemantic categories which the KWs analysed below were mapped onto, including those related to crime. The results of the analysis support our interpretation of the findings as they point at the almost irrelevant presence of lexical items comprised in the theme *crime,* as evidenced below, specifically when compared against other categories such as *family* or *home*.

3.3 Analysis and discussion

3.3.1 *Keywords*

As a first approach, two frequency wordlists were obtained using *Wordsmith* (Scott, 2020a) after processing both the legal and the general corpus. We began the analysis by manually classifying the top 1,000 KWs into different thematic categories. Prior to this,

a selection of those KWs was made based on their lexical content, that is, function words (only a few of these appeared amongst the top-ranking 1,000 KWs) or those content words which referred to general legal terms such as *sentencia* (sentence), *recurso* (appeal) *jurisprudencia* (jurisprudence) or *tribunal* (court), were discarded since they did not provide any information on the actual content of the texts.

Three major themes were identified amongst the top 1,000 KWs, namely, *family*, *territory/access* and *punishment*. The total number of the vocabulary items comprised in the three categories represent 20% of the top 1,000 KWs selected for the analysis, while none of them points at any of the crimes typified in the Spanish criminal code and recorded in the official statistics provided by the Ministry of Justice for the years 2016 and 2017.

The procedure applied to select the sample sentences below goes as follows. For a concordance line (or an extension of it) to illustrate the analysis provided, a thorough examination of the texts they were extracted from was accomplished. This way, we made sure that we were interpreting the information contained in them correctly and avoided any ambiguity or bias on our part.

Table 3. Family keywords

Table 3 displays a sample of the KWs which fall under the category *family,* whose average frequency and keyness[6] are 493.36 and 1,997.68 respectively. These data highlight their statistical relevance as they are twice as frequent as the average for the whole corpus, while their keyness value is slightly lower than the average for the top 1,000 KWs, standing 13.25% below.

The category *family* ranks third amongst the ones examined as regards frequency and keyness, since *territory/access* and *legal punishment* stand in the first and second positions, however, it provides valuable information about one of the major concerns or issues which relate to migrants who enter a territory and are brought to court due to some irregularity, as found in *Baker at al.* (2008). Let us insist on the fact that we are dealing with texts which inform of legal procedures (of varied nature) which were initiated against a migrant or a group of them and their result. The statistical salience of a topic like *family* deserves specific attention, particularly when it comes to supporting our stance against the association between immigration and crime as put forward by the Spanish far right. As seen in Wodak (2015: 28), parties like the Spanish VOX, often blame migrants for all our woes and present them as a threat to our nations. Yet, having examined a considerable amount of the texts in the corpus and as illustrated by the excerpts below, the portrayal of those who migrate to a new territory as seen in Spanish sentences differs greatly from the image put forward by far right parties, as will be discussed further on.

Interestingly, the discourse of the far right is reflected on the Spanish press. Alcaraz-Mármol and Soto-Almela (2016) find that association in corpora of Spanish newspapers with different political ideology, finding negative semantic prosody for the words *inmigración* (immigration) and *inmigrante* (immigrant), since most of the words co-occurring with those lexical items have a negative meaning. The term *refugiado* (refugee) has also been depicted negatively in the Spanish media, judging by the lexical company it keeps in the corpus collected by Alcaraz-Mármol and Soto-Almela (2018).

The figures provided above highlight the salience of this set of KWs with respect to the whole corpus as their average frequency is twice as high as the same value for the whole corpus and its keyness, although lower than the average, is still remarkable for a

[6] Obtained after implementing the *log-likelihood* algorithm on both frequency lists, the legal and the general one, using the software cited above.

single set of terms, as already stated.

The context of usage of some of these KWs reinforces this perception, as it links immigration to family issues and, by extension to irregular situations where the defendants seek asylum for humanitarian reasons or submit family reasons which may or may not grant them legal access to the country. The concordance lines extracted from our corpus, which relate to the terms *familiar* (familiar), *reagrupante* (person with whom migrants are reunited) or *esposo* (*husband*), instantiate this fact. The adjective *familiar* often occurs throughout the entire text collection and illustrates quite clearly the difficulties that migrant children go through when they live with their parents in a precarious situation. It may be worth noting that it is a polysemic word which can also refer to something that is easily recognisable, yet, over 70% of the top 100 concordances obtained from the corpus relate to the noun *family*. The samples where this word was used as a partial synonym of *recognisable*, as in any other similar instances, were not taken into consideration for the analysis.

The first concordance lines tell us about a child not being granted a stable home (*el padre no garantizaba un domicilio estable para la unidad familiar*) because she lived in an overcrowded flat with her father (*un piso compartido por muchas personas*), while the second sample refers to the fact that children's interests must become a priority (*el interés superior del niño*) when making decisions about their families' permission to enter or stay in a country.

FAMILIAR

- *(...) de la niña, la menor quedó desprotegida. El padre no garantizaba un domicilio estable para la unidad* ***familiar*** *(...) sino un piso compartido por muchas personas.*
- *(...) debemos tener debidamente en cuenta el interés superior del niño, la vida* ***familiar*** *(el arraigo* ***familiar*** *en nuestro ordenamiento), y el estado de salud del extranjero (...)*

The term *reagrupante* (the person with whom migrant families are reunited) points in the same direction as *familiar* by relating the status of migrants to the possibility of their reuniting their family in the country where one of them is living and working. In the first sample, we find a judicial decision which allows for the request to reunite a migrant family although partially (*reagrupación parcial*), not all the applicants are granted a permit to stay (*no del entero núcleo familiar)*. In the second sample, the residence permit is granted to temporarily reunite the whole family (*conceder al actor autorización de residencia temporal)* thanks to the wife's request (*a instancia de la esposa* ***reagrupante*****).**

REAGRUPANTE

- *(...) nos pronunciamos a favor de la admisión en nuestro Derecho de la reagrupación parcial, que supone la de algún o algunos miembros de la familia del* ***reagrupante*** *y no del entero núcleo familiar (...)*
- *(...) Con fecha 30 de mayo de 2016 la Subdelegación del Gobierno en Barcelona resolvió conceder al actor autorización de residencia temporal inicial por reagrupación familiar a instancia de la esposa* ***reagrupante****.*

Table 4. Territory/Access

Table 4 comprises a sample of the set of KWs within the topic *territory/access,*

which is, by far, the one with the highest frequency average, 1,219.48 (six times higher than the same value for the whole corpus applying the >5 threshold: 236.97) and the highest keyness value out of the three categories, namely, 3,776.50 (66% higher than the average for the top 1,000 KWs: 2,262.48).

These data reveal the enormous relevance of this topic in the legal corpus, something that we had already observed in the analysis of the category *family,* which pointed in this direction. Migrants are brought to court for different reasons, however, as already stated, it is not crime that stands out as the major area which these texts revolve around. As will be discussed in greater detail in section 3.3.2., figure 3 informs of the only instances in which crime nouns are identified as statistically relevant in the corpus. When describing the collocate network of the term *delito* (*crime*) itself we find other terms like *malversación* (*embezzlement*), *lesiones* (*injuries*) or *robo* (*theft*). Even so, their average frequency as collocates is almost 9 times lower than the collocates of other terms like *multa* or *antecedentes*, which, in fact, do not explicitly refer to crime.

On the contrary, we find clearer evidence of the figure of migrants as related to their attempt to gain access to a territory other than their home countries or to request a residence permit. However, they are often involved in irregular immigration processes appearing as victims of human trafficking or simply trying to stay or bring their families to a European country in an irregular way. The concordance lines below illustrate this.

One of the KWs in this category is the term *permanencia* (residence time), which occurs 2,020 times in the corpus and can be found in different contexts instantiating various situations. In the first excerpt, we find evidence of human trafficking as the defendants admit having organized a profit-oriented network (*los acusados han manifestado participar, con ánimo de lucro),* helping Asiatic citizens enter the country irregularly (*la entrada en territorio español de asiáticos carentes de la documentación oficial*). On the other hand, the second extract reflects the situation of a migrant who has requested a residence permit and has proved that she is the partner of a legal resident *(constando además que convive con residente legal).* The migrant also proves that she does not have a criminal record *(careciendo de antecedents penales).* Once more these samples signal the relevance of both themes, family and access to the territory, as fundamental to understand the image of the phenomenon of immigration in our country as seen through the lens of the judiciary.

PERMANENCIA

- *(...) cada uno de los acusados ha manifestado participar, con ánimo de lucro, en la entrada en territorio español de asiáticos carentes de la documentación oficial (...) para su* ***permanencia*** *en España o bien para que pudieran desplazarse por el resto*
- *(...) constando además que convive con residente legal y que ha tenido una larga* ***permanencia*** *en territorio Español, careciendo de antecedentes penales, sin que se haya sancionado al recurrente anteriormente por la misma infracción (...)*

The noun *retorno* (return) is particularly interesting as its contexts of usage reveal the desperate need that migrants have to flee from their countries and the dangers implied in returning to their native homes. The examination of a considerable contexts of usage associated to it (500 out of 2,227), which were sorted according to its left-hand collocates to facilitate the task, confirms our interpretation.

In the first two excerpts, we are told that migrants could risk their lives if they go back to their home countries (*grave peligro que para su vida supone el retorno a Nigeria*), where their relatives are receiving death threats (*su familia sigue recibiendo*

amenazas) and where they will be condemned to certain death if they return (*sería condenarle a una muerte segura*). Conversely, the third sample reflects a different situation. A migrant requests to be readmitted and return to Spain but her appeal is dismissed (*desestimó el recurso*) after having been banned from national territory for two years (*con prohibición de **retorno** de dos años*).

RETORNO

- *(...) Ministerio del Interior otras alegaciones manuscritas insistiendo en el grave peligro que para su vida supone el **retorno** a Nigeria (...)*
- *(...) según expone fue obligado por las amenazas de un grupo criminal de su país. Su familia sigue recibiendo amenazas y un posible **retorno** a su país sería condenarle a una muerte segura. Introduce drogas porque pide (...)*
- *(...) enero pasado que desestimó el recurso y confirmó la legalidad de la Resolución de la Delegación de Gobierno de 5 de junio de 2014 que expulsó a la recurrente y ahora apelante del territorio nacional con prohibición de **retorno** de dos años.*

Table 5. Legal punishment

The category *legal punishment* ranks second amongst the three topics analysed in this section. The average frequency for the whole category is 810.30, that is, four times as high as the corpus frequency average. Its average keyness is the second highest one (3,351.31), being 48% above the same value for the top 1,000 KWs: 2,262.48, hence its statistical relevance.

When we approached this KW category, we expected to find a considerable amount of KWs that could relate to the semantic field of crime. However, out of 21 KWs, only the words *delito* (*crime/offence*) and *trata* (*human trafficking*) themselves explicitly referred to it. As regards the reference corpus, none of the KWs extracted refers to legal punishment. The term *antecedentes* (criminal record) was not considered as explicitly related to the semantic field of crime because, after examining a considerable amount of concordance lines associated to it, we found that it was often stated that the defendants had no criminal record, that is, that they had not committed any crime previously. This seems to be in line with Pérez-Paredes et al. (2017), who state that the view put forward by UK legislation and official information bears almost no relationship with criminal activities.

The rest of terms within this group connected to procedural terminology which indicated, in its majority, the punishment that migrants received (*condena*, *multa*, *pena*) for breaking the law (*infracción*, *delito*) and accessing the territory or willing to remain in it irregularly. These findings agree with the those presented by Alcaraz-Mármol and Soto-Almela (2018) and Khosravinik's (2008) about the binomy in the portrayal of migrants as victimized or victimizers. In their analyses they find that lexical items like *victimization*, *vulnerability*, *suffering*, and *desperation* tend to co-occur with *refugee* in the corpora under scrutiny, often depicting them as unprotected victims.

Nonetheless, this does not mean that the texts, which are judicial decisions, do not contain any crime-related vocabulary. As a matter of fact, they do, yet such lexicon does display any statistical significance owing to its lesser relevance with regard to the whole text collection, hence our interpretation of the results at this point.

By examining the extracts below, it is inferred that the term *multa* (*fine/penalty*)

is often used in connection with procedural content which explains how the law is applied. The first two excerpts actually link illegal stay (*en los casos de permanencia illegal; la estancia irregular*) to the punishment (*multa/sanción*) received in each case. Once more and as pinpointed above, the third sample reflects a situation where the defendants admit (*responden en concepto de autores*) having committed a crime against the rights of migrants (*contra los derechos de los ciudadanos extranjeros)* for which they are sentenced to prison and have to pay a fine (*una multa*).

MULTA

- *(...) en los casos de permanencia ilegal, la Administración, según los casos, puede imponer o bien la sanción de **multa** o bien la sanción de expulsión.*
- *LOEX sanciona con **multa** la estancia irregular sin prever una decisión de retorno, por lo que no colmaría las exigencias del artículo (...)*
- *(...) los acusados, (...) responden en concepto de autores, de un delito contra los derechos de los ciudadanos extranjeros (...) para los que se interesó la pena nueve meses de prisión y **multa** de seis meses a razón de seis euros por día, (...)*

The examination of the concordances related to each of the terms frequently implied discarding many of the samples, particularly when the term was polysemic. This is the case of *trata*, which could act as a verb meaning *deal with* or *be about* and also as a noun referring to *human trafficking*. Only the latter interpretation was considered for this analysis.

The term *trata* is actually defined in the law and it is understood, as shown in the first excerpt, as a profit-oriented network (*la fuente principal de ingresos y el motivo económico impulsor del delito*) which benefits from the sexual exploitation of the women who are brought illegally to a country (*la explotación de las víctimas en la prostitución*). The texts also clarify the difference between illegal immigration and human trafficking since the latter not only implies an irregular migration process (*la trata de seres humanos puede tener carácter trasnacional o no*) but also the sexual exploitation of women, which could also happen within Spanish territory without necessarily depending on those women coming from other non-European countries (*las víctimas pueden ser ciudadanos europeos, o incluso españoles*), as stated in the second extract. The third sample depicts the process by which women are captured by these criminal organisations (*captación de las víctimas*) which take advantage of the precarious living conditions in their home countries (*aprovechando su precaria situación económica en Nigeria*) to introduce them in Spanish territory for prostitution (*para introducirlas en nuestro país de forma irregular con el propósito de que ejercieran la prostitución).*

TRATA

- *En el supuesto de la **trata** de personas, la fuente principal de ingresos para los delincuentes y el motivo económico impulsor del delito es el producto obtenido con la explotación de las víctimas en la prostitución (...)*

- *La otra gran diferencia básica entre la inmigración ilegal y la **trata** radica en que la primera siempre tiene un carácter transnacional, (...) mientras que la trata de seres humanos puede tener carácter trasnacional o no, ya que las víctimas pueden ser*

ciudadanos europeos, o incluso españoles.

- *En el caso enjuiciado se aprecia fácilmente la concurrencia de estos elementos típicos típicos de la* ***trata*** *(...) captación de las víctimas, aprovechando su precaria situación económica en Nigeria (...) para introducirlas en nuestro país de forma irregular con el propósito de que ejercieran la prostitución.*

3.3.2 *Collocate Networks*

In order to support the results obtained in the examination of the KWs above, an analysis of the collocate networks of some of these items was carried out. The collocates of each of the members of the category *legal punishment* were also processed with *Graphcoll* adjusting the settings to apply the MI3 *measure* (the cubed version of Church and Hanks' (1990) mutual information measure), a word association measure which, according to Brezina *et al.* (2015: 160) tends to push more frequent combinations to the top of the rank, leaving the most unusual patterns aside or either relegating them to the bottom of the collocate inventories, in other words, 'the measure gives more weight to observed frequencies and thus gives high scores to collocations which occur relatively frequently in the corpus'. By applying this measure, as illustrated in the figures below, except for the term *crime* itself, none of the items within the category was linked to criminal offences but rather to other general legal terms, or terms related to territory/access.

Although crime-related nouns do occur in the corpus, the statistical salience of this set of terms with respect to the whole text collection is irrelevant in comparison with other terms from the categories *family* or *territory,* as already stated. Likewise, the relevance of crime nouns when contrasted with the reference corpus is reduced to a minimum or almost inexistent, hence their absence from the keyword list.

Three out of the top five most frequent terms within the category *legal punishment, multa* (F=4263)*, delito* (F=1308) and *antecedentes* (F=1137)*,* were singled out for the analysis of their collocational networks. Let us start by examining the network of the term *multa* (*fine*, *pelnalty*), by looking at Fig. 1, as generated by *Graphcoll*. As well as providing an image of the networks, the tool *Graphcoll* produces tables with all the items which configure them, so it becomes easier to get to know all the collocates associated to each term and their data by extension. In the first place, the lexical network of *multa* is an overpopulated one (the tables facilitate greatly the reading of data), as the term itself occurs 4,263 times in the corpus, hence the large number of collocates it generates. Out of them, none refers explicitly to crime, finding words like *expulsion* (*deportation* F=1,978)[7], *territorio* (*territory* F=353), *español* (*Spanish* F=234), *nacional* (*national/domestic* F=158), or *estancia* (*stay* F=117) amongst its most relevant collocates, all of them connected with the topic *territory/access.*

Fig. 1 Lexical network of *multa*

[7] The F stands for the frequency of both terms as collocates, that is, the number of times they occur together within the corpus.

Something similar can be observed within the network of the term *antecedentes* . In the same way as *trata* or *familiar*, the term *antecedentes* might have more than one interpretation. On the one hand, this term refers to the defendants' criminal record and, on the other hand, to a specific section found in Spanish judicial decisions where a summary of the legal process prior to its hearing at court is presented which serves as context for the sentence itself. In this particular case, for the study of the lexical network of *antecedentes*, all the collocates displayed below, having examined the concordances associated to each of them, relate to the meaning "criminal record".

As seen in Fig. 2, the top collocates of the term *antecendentes* bear no relationship with any of the crime categories recorded in the official statistics such as *homicidio* (*murder*), *lesiones* (*injuries*) *robo* (*theft*) or *tráfico de drogas* (*drug dealing*). Conversely, we find terms like *penales* (*criminal* F=349), *policiales* (*police-related F=92*), *abogado* (*solicitor/lawyer F=126*) or *estado* (*state F=162*), most of them being considered as general legal terms used to describe legal procedures.

Fig. 2 Lexical network of *antecedentes* (criminal record)

Finally, only the lexical network of the term *delito* (*crime*) itself, as illustrated by Fig. 3, as was to be expected, contains lexical elements which explicitly refer to crime. The most significant are *falsedad documental* (*forgery* F=56), *malversación* (*embezzlement* F= 44), *lesiones* (injuries F=40), *robo* (theft F=34) or *tráfico de drogas* (*drug dealing* F=32). In fact, we find the greatest proportion or better said, the only proportion of crime nouns amongst the collocates of *delito* (*crime*). Nevertheless, their statistical relevance as collocates, that is, how many times they occur in the corpus as collocates, is much lesser than it is in the other two networks examined above, displaying 33.33 average collocate frequency (not considering function words in any case) as opposed to 286.45 for all the collocates of *multa (fine)* and 77.96 for *antecedentes* (*criminal record*).

Fig. 3 Lexical network of *delito* (crime)

3.3.3 *Data Triangulation Using UMTextStats*

A third method, *UMTextStats* text analysis, was adopted in our study in an attempt to triangulate the results, trying to offer a new perspective that might enhance our interpretation of the results, as suggested above. For the sake of clarity, it is worth noting that the second column of Table 6 shows some of the examples from each category as they occur in our corpus. Regarding the last column of the table, it indicates the relative frequency of these word categories, that is to say, the ratio of the words in each category to the total word count of the corpus expressed as a percentage.

Special attention deserves the *ad-hoc* category (Table 6) included in the tool for the present study, *crime*, as its results confirm our initial observation: only 0.03% of the words in the whole corpus fall into this category. Similarly, the category *violence* is even less significant, as only 0.01% of the words are mapped to it. This stands in stark contrast to the meaningful ratios included in Table 6, like those of *affiliation* (0.64%) and *family* (0.32%), intimately related to our first topic: *home* (0.74%), connected to our second theme (*territory/access*); and the category *risk* (0.43%), which includes elements

semantically linked to *punishment*, the third outstanding topos advanced in our previous analyses. Under close scrutiny, the lexical elements mapped to this category are mainly connected to terminology indicating the risk that migrants take when accessing the territory and/or staying there illegally. Interestingly, *job* also stands out as one of the most frequent categories (0.58%), which is coherent with the main motivation for migratory movements. Thus, these results seem to confirm the findings from the two previous analyses, supporting the idea that the keyness of the corpus lies in the three major themes *family, territory/access,* and *legal punishment*.

Table 6. Relevant UMTextStats categories

All in all, as observed above, the results appear to dissociate immigration and crime or violence, conversely associating this social phenomenon to more positive or neutral categories. These findings are in line with the evidence presented above, which highlighted the absence of any major offences amongst the top-ranking lexicon. Thus, it can be stated that the triangulation of our data with UMTextStats confirms the results yielded by the keywords analysis and the collocate networks: even if the perspective gained by this NLP tool provides a different type of analysis based on the salience of morphosemantic categories, its results indicate that categories like *crime* or *violence* do not have a strong representation in our corpus, as opposed to *family* or *home*.

4. Conclusion

This research has examined the linguistic evidence found in a 2.4 million-word corpus of Spanish sentences on immigration which, according to the analysis of the keywords extracted and their collocational networks, points at the dissociation between the phenomenon of immigration itself and crime. This study arose from the urge to refute hate speech, which directly connects both concepts by distorting the official data available and attempting to dehumanise migrants, who are presented as a threat to the welfare state in developed countries, as described by authors like Wodak (2005).

The analysis began with the identification of the major themes which the keywords (KWs) in the legal corpus can be categorised into, namely, *territory/access*, *legal punishment* and *family*. Specifically, the category *territory/access* ranked first and reached the highest keyness value out of the three categories. The mere quantitative comparison between this category, the whole type list and the top 1,000 KWs highlights its representativeness and significance in the corpus.

None of the keyword terms analysed referred explicitly to any of the crime categories typified in the Spanish Criminal Code and present in the official statistics provided by the Spanish Home Office such as *asesinato* (*murder*), *robo* (*robbery*, *theft*) or *lesiones* (*injuries*). Consequently, and from a purely quantitative perspective, the statistical salience of crime-related terms as opposed to other categories such as *family* or *territory* is practically inexistent. Only the words *delito* (*crime/offence*) and *trata* (*human trafficking*) themselves explicitly refer to crime (roughly 3% of the whole set). The examination of the collocates of some of the items in each category reinforces the idea

that, once more, the presence of crime-related nouns in association with the word *family* of *inmigración* is merely incidental.

The triangulation of these data using *UMUTextStats* confirmed the results discussed above, as we found no evidence in our legal corpus (in itself and as compared against a general corpus) that supports the link between crime and immigration. The use of this NLP tool, which envisages linguistic analysis from a different perspective, added psycholinguistic criteria onto the mapping of the morphosemantic categories identified in the corpus and allowed us to establish the percentage of terms which fell under the categories *crime* or *violence*, amongst others. The figures clearly determine that these themes do not have a strong representation in the corpus either, as opposed to those within the categories *family* or *home,* since *crime* and *violence* represent 0.03% and 0.01% respectively of the total amount of corpus tokens.

On the other hand, the qualitative analysis of the contexts of usage of those KWs and their collocates also reveals the fundamental role of families in the life of migrants, who often seek asylum for humanitarian reasons or submit family reasons to their permit applications. Gaining access to a territory other than their home countries or requesting a residence permit appears as another major issue that the texts in the corpus revolve around. However, migrant women are frequently presented as victims of human trafficking and are brought to Spain under threat to be sexually exploited.

In sum, all these data gain even greater significance if we consider the genre the corpus stems from, judicial decisions, which reflect the legal proceedings which migrants are subject to in one way or another. Migrants are brought to court and sentenced, often because of irregular migration processes yet, based on these findings, we found no sound statistical ground to affirm that there is a strong link between immigration and crime as found in the texts and can thus refute the arguments of the Spanish far-right with objective statistical arguments.

REFERENCES

Anthony, L. 2014. *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. www.laurenceanthony.net

Alcaraz, E. 1994. *El inglés jurídico*. Madrid: Ariel Derecho.

Alcaraz-Mármol, G. and J. Soto-Almela. 2016. 'The semantic prosody of the words inmigración and inmigrante in the Spanish written media: A corpus-based study of two national newspapers', *Revista Signos* 49 (91), pp. 145-167.

Alcaraz-Mármol, G. and J. Soto-Almela. 2018. Refugees in the Spanish press: A corpus-assisted study of the semantic prosody of the term refugiado from a diachronic perspective. *Sintagma* 30, pp. 95-113.

Author. 2012

Author. 2014

Author. 2016

Author. 2019

Baker, P. 2005. *Public Discourses of Gay Men*. London: Routledge.

Baker, P. 2016. 'The shapes of collocation', *International Journal of Corpus Linguistics* 21 (2), pp. 139-164.

Baker, P. 2018. Language, sexuality and corpus linguistics. Concerns and future directions. *Journal of Language and Sexuality*, 7 (2): 263 - 279.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19 (3), pp. 273-306.

Baker, P., C. Gabrielatos, and T. McEnery. 2013. *Discourse Analysis and Media Attitudes. The Representation of Islam in the British Press*. Cambridge: CUP.

Barceló-Coblijn, Ll., Serna Salazar, D., Isaza, G., Castillo, L.F., Ossa, M.G. (2017) Netlang: A software for the linguistic analysis of corpora by means of complex networks. *PLoS ONE*, 12 (08). https://doi.org/10.1371/journal.pone.0181341

Bisceglia, B., Calabrese, R., and Leone, L. (2014). 'Combining Critical Discourse Analysis and NLP tools in investigations of religious prose', *LRE-REL2* 24.

Borja, A. 2000. *Legal English.* Valencia: Cámara de Comercio de Valencia.

Egbert, J. and P. Baker. 2020. *Using Corpus Methods to Triangulate Linguistic Analysis*. New York/London: Routledge.

Blinder, S. and W. L. Allen. 2016. 'Constructing immigrants: Portrayals of migrant groups in British national newspapers, 2010–2012', *International Migration Review* 50 (1), pp. 3-40.

Brezina, V. 2018. Statistical choices in corpus-based discourse analysis. In Taylor, C. & Marchi, A. *Corpus Approaches to Discourse: A Critical Review*. London: Rouledge.

Brezina, V., T. McEnery, and S. Wattam. 2015. 'A new perspective on collocation networks', *International Journal of Corpus Linguistics*, 20 (2), pp. 139-173.

Calvo Vidal, Félix M. 1992. *La Jurisprudencia ¿Fuente del Derecho?* Valladolid: Lex Nova.

Church, K. W. and P. Hanks. 1990. 'Word association norms, mutual information, and lexicography', *Computational Linguistics* 16 (1), pp. 22-29. dl.acm.org/citation.cfm?id=89095 (accessed 15 March 2020).

Crawford, B. (2019). *The Dehumanization of Immigrants and the Rise of the Extreme Right.* https://www.aicgs.org/publication/the-dehumanization-of-immigrants-and-the-rise-of-the-extreme-right/ (accessed 20 March 2020).

Cruse, D. A. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.

Danet, B. 1980. "Language in the Legal Process". *Law and Society Review*, 14 (3): 445-564.

Dunning, T. 1993. 'Accurate Methods for the Statistics of Surprise and Coincidence', *Computational Linguistics* 19 (1), pp. 61-74. dl.acm.org/citation.cfm?id=972454 (accessed 15 March 2020).

Firth, J. R. 1957. *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Gabrielatos, C., and P. Baker. 2008. 'Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005', *Journal of English Linguistics* 36 (1), pp. 5-38.

García-Díaz, J. A., M. Cánovas-García, and R. Valencia-García (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*, 112, 641-657. doi:10.1016/j.future.2020.06.019

García-Díaz, J. A., M. P. Salas-Zárate, M. L. Hernández-Alcaraz, R. Valencia-García, and J. M. Gómez-Berbís. 2018. 'Machine Learning Based Sentiment Analysis on Spanish Financial Tweets', *WorldCIST* 1, pp. 305-311.

Guilquin, G. 2008. Taking a New Look at Lexical Networks. *Lexis*, 1. https://doi.org/10.4000/lexis.757

Gries, S. T. 2013. '50-something years of work on collocations: What is or should be next', *International Journal of Corpus Linguistics* 18 (1), pp. 137-166. https://doi.org/10.1075/ijcl.18.1.09gri (accessed 16 March 2020).

Hancock, J. T., M. T. Woodworth, and S. Porter. 2011. 'Hungry like the wolf: A word-pattern analysis of the language of psychopaths', *Legal and Criminological Psychology* 18 (1), pp. 1-13.

Jumaquio-Ardales, A., N. Oco, and R. Madula. 2017. 'Click-analysis of a Lesbian Online Community in Facebook Using the CDA and NLP', *Humanities Diliman: A Philippine Journal of Humanities* 14(1).

Jurafsky, D. and J. H. Martin. 2018. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Third Edition). https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed 30 March 2020).

Khosravinik, M. (2008). British newspapers and the representation of refugees, asylum Seekers and immigrants between 1996 and 2006. Centre for Language in Social Life, 128, 3-40.

Retrieved from http://www.lancaster.ac.uk/fass/groups/clsl/docs/clsl128.pdf

Maley, Y. 1994. "The Language of the Law". In Gibbons, J. (ed). *Language and the Law*. London: Longman.

Mellinkoff, D. 1963. *The Language of the Law*. Boston: Little, Brown.

McEnery, T. 2006. *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Abington, UK: Routledge.

Mihalcea, R. and C. Strapparava. 2009. 'The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language' in K. Y. Su, J. Su, J. Wiebe, H. Li (eds.) *Proceedings of the Association for Computational Linguistics, ACL-IJCNLP 2009*, pp. 309-312. Singapore: ACL.

Orts Llopis, M.A. 2009. "Legal genres in English and Spanish: some attempts of analysis". *Iberica*, 18: 109-130.

Pennebaker, J. W., and A. Graybeal. 2001. 'Patterns of natural language use: Disclosure, personality, and social integration', *Current Directions in Psychological Science* 10, pp. 90-93.

Pennebaker, J. W. and M. Francis. 1999. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers.

Pérez-Paredes, P., P. Aguado, and P. Sánchez. 2017. 'Constructing immigrants in UK legislation and Administration informative texts: A corpus-driven study (2007–2011)', *Discourse and Society* 28 (1), pp. 81-103.

Ramírez-Esparza, N., J. W. Pennebaker, F. A. García, and R. Suriá. 2007. 'La psicología del uso de las palabras: Un programa de computadora que analiza textos en español', *Revista Mexicana de Psicología* 24(1), pp. 85-99.

Rayson, P. and R. Garside. 2000. 'Comparing corpora using frequency profiling' in A. Kilgarriff and T. Berber Sardinha (eds.) *Proceedings of the Workshop on Comparing Corpora*, held in conjunction with the *38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pp. 1-6. Hong Kong: ACL.

Reese, S., G. Boleda, M. Cuadros, G. Padró, and L. Rigau. 2010. 'Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus' in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias (eds.) *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, pp. 1418-1421. La Valleta, Malta: European Language Resources Association.

Sánchez, P., P. Aguado, and P. Pérez-Paredes. 2019. 'Featuring immigrants and citizens: A comparison between Spanish and English primary legislation and administration information texts (2007–2011)' in L. Viola and A. Musolff (eds.) *Migration and Media: Discourses about identities in crisis*, pp. 63-90. Amsterdam: John Benjamins.

Scott, M. 2020a. *WordSmith Tools version 8.* Stroud: Lexical Analysis Software.

Scott, M. 2020b. *WordSmith Tools Help*. Stroud: Lexical Analysis Software.

Sinclair, J. 1991. *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

Sinclair, J. 2005. 'Corpus and Text: Basic Principles' in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, pp. 1-16. Oxford: Oxbow Books. http://ahds.ac.uk/linguistic-corpora/ (accessed 16 March 2020).

Stuart, K., Botella, A. 2009. Corpus Linguistics, Network Analysis and Co-occurrence Matrices. *International Journal of English Studies*, special issue: 1-28.

Taylor, C. 2014. 'Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis', *International Journal of Corpus Linguistics* 19 (3), pp. 368-400.

Tiersma, P. 1999. *Legal Language*. Chicago: The University of Chicago Press.

Williams, G. 2001. 'Mediating between lexis and texts: collocational networks in specialised corpora', *ASP, la revue du GERAS* 31, pp. 63-76. asp.revues.org/1782 (accessed 15 March 2020).

Wodak, R. 2015. *The Politics of Fear. What Right-Wing Populist Discourses Mean.* London: Sage.

Table 1 Corpora description

CORPORA	TEXTS	TOKENS	TYPES
Spanish legal corpus *CENDOJ*	600	2,396,985	20,236
Spanish general corpus *WIKIPEDIA*	94	101,322,383	732,795

Table 2 List of lexical keywords

N	Keyword	Freq.	Keyness
1	SENTENCIA (sentence)	11,926	66,558.5
2	RECURSO (appeal)	9,596	53,973.37
3	ADMINISTRATIVO (administrative)	5,419	33,402.94
4	CONTENCIOSO (adversarial procedure)	4,690	30,212.81
5	MULTA (fine)	4,263	24,724.75
6	RECURRENTE (recurring)	3,383	23,831.76
7	JURISPRUDENCIA (jurisprudence)	3,110	21,878.72
8	TRIBUNAL (court)	6,240	21,334.13
9	APARTADO (section)	4,037	21,007.32
10	DIRECTIVA (directive)	4,271	20,904.46
11	PROCEDIMIENTO (procedure)	4,028	20,356.25
12	TERRITORIO (territory)	4,427	18,212.66
13	ART	3,328	16,037.17
14	SALA (court section)	5,244	16,029.28
15	IRREGULAR (irregular)	2,783	15,759.55
16	LEY (act)	6,114	15,576.40
17	DERECHO (law)	5,520	13,974.32
18	CIRCUNSTANCIAS (circumstances)	2,889	12,843.32
19	EXPEDIENTE (record)	2,588	12,708.45
20	APELANTE (apellant)	1,590	12,087.12

Table 3 Family keywords

	Keyword	**Freq.**	**Keyness**
Rank #			
47	ARRAIGO (family roots/ties)	1,199	7,903.78
74	FAMILIAR (familiar)	1,950	5,185.44
224	MATRIMONIO (marriage)	875	2,062.15
231	NULIDAD (invalidity)	367	2,019.75
236	REAGRUPANTE (person with whom migrants are reunited)	258	1,968.22
560	TUTELA (legal guardianship)	214	690.59
728	REAGRUPADO (person who has been reunited)	75	491.72
776	REAGRUPAR (action of reuniting)	78	450.00
793	MATRIMONIOS (marriages)	145	436.39
833	ESPOSO (husband)	197	398.05

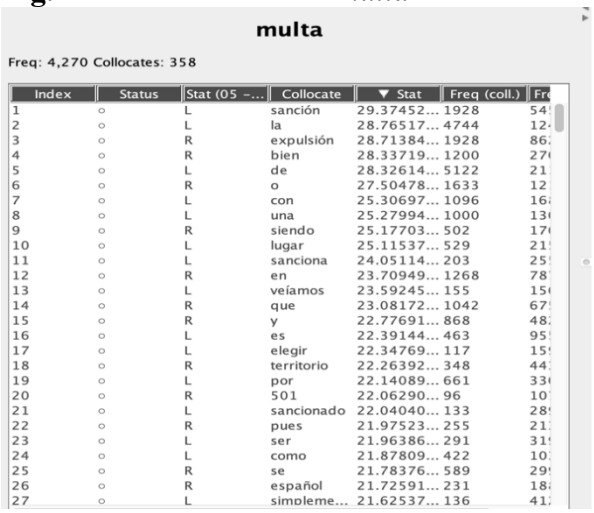
Table 4 Territory/Access

	Keyword	**Frequency**	**Keyness**
Rank#			
17	TERRITORIO (territory)	4,427	18,212.66
20	IRREGULAR (irregular)	2,783	15,759.55
28	PERMANENCIA (residence time)	2,020	11,812.33
34	RETORNO (return)	2,203	10,494.70
48	RESIDENCIA (residence)	2,214	7,790.02
55	NACIONAL (national/domestic)	4,078	6,654.68
72	ESTADO (state)	4,551	5,251.27
75	VISADO (visa)	984	5,160.15
81	EXTRANJERO (foreigner/migrant)	1,522	5,083.37
91	ESTANCIA (stay)	1,863	4,388.22

Table 5 Legal punishment

9	MULTA (fine)	4,263	24,724.75
79	ANTECEDENTES (criminal record)	1,137	5,085.36
100	PENAL (criminal)	1,298	4,017.51
119	DELITO (crime/offence)	1,308	3,496.88
1223	TRATA (human trafficking)	756	214.07
228	PENALES (criminal)	503	2,043.00
262	INFRACCIONES (infringement/violation)	403	1,698.45
320	CONDENA (sentence/conviction)	669	1,389.22
336	PENA (penalty/conviction)	840	1,344.56
360	SANCIÓN (punishment/penalty)	492	1,262.54

Fig. 1 Lexical network of *multa*



multa

Freq: 4,270 Collocates: 358

Index	Status	Stat (05 –...	Collocate	▼ Stat	Freq (coll.)	Fr
1	○	L	sanción	29.37452...	1928	54
2	○	L	la	28.76517...	4744	12
3	○	R	expulsión	28.71384...	1928	86
4	○	R	bien	28.33719...	1200	27
5	○	L	de	28.32614...	5122	21
6	○	R	o	27.50478...	1633	12
7	○	L	con	25.30697...	1096	16
8	○	L	una	25.27994...	1000	13
9	○	R	siendo	25.17703...	502	17
10	○	L	lugar	25.11537...	529	21
11	○	L	sanciona	24.05114...	203	25
12	○	R	en	23.70949...	1268	78
13	○	L	veíamos	23.59245...	155	15
14	○	R	que	23.08172...	1042	67
15	○	R	y	22.77691...	868	48
16	○	L	es	22.39144...	463	95
17	○	L	elegir	22.34769...	117	15
18	○	R	territorio	22.26392...	348	44
19	○	L	por	22.14089...	661	33
20	○	R	501	22.06290...	96	10
21	○	L	sancionado	22.04040...	133	28
22	○	R	pues	21.97523...	255	21
23	○	L	ser	21.96386...	291	31
24	○	L	como	21.87809...	422	10
25	○	R	se	21.78376...	589	29
26	○	R	español	21.72591...	231	18
27	○	L	simpleme...	21.62537...	136	41

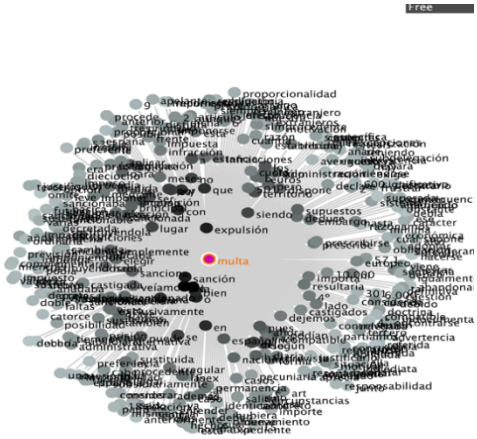


Fig. 2 Lexical network of *antecedentes* (criminal record)

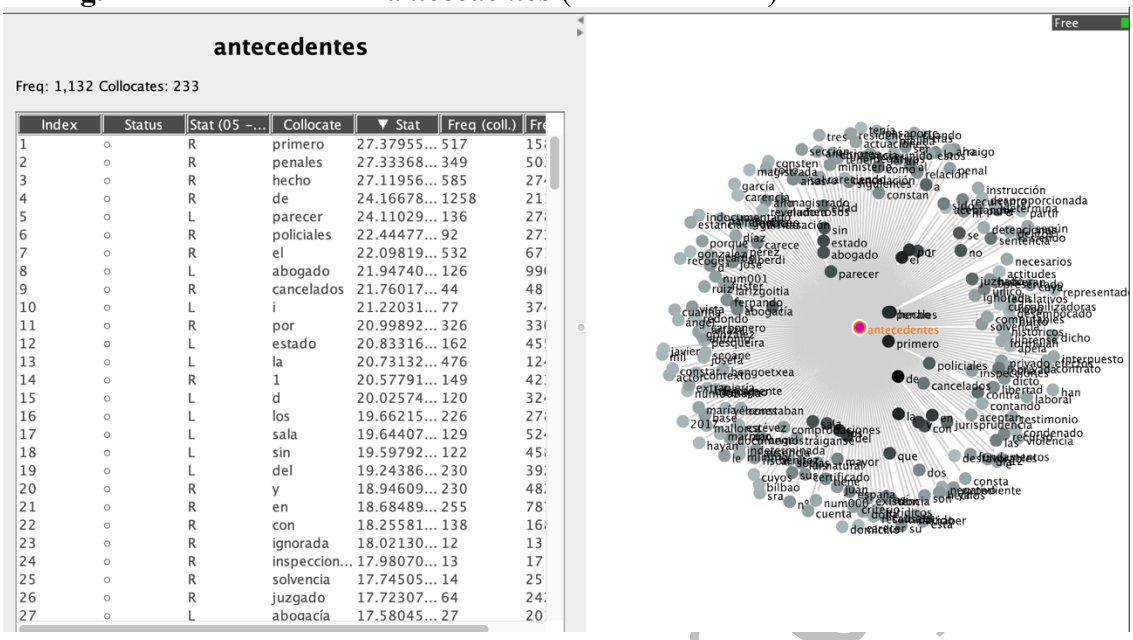


Fig. 3 Lexical network of *delito* (crime)

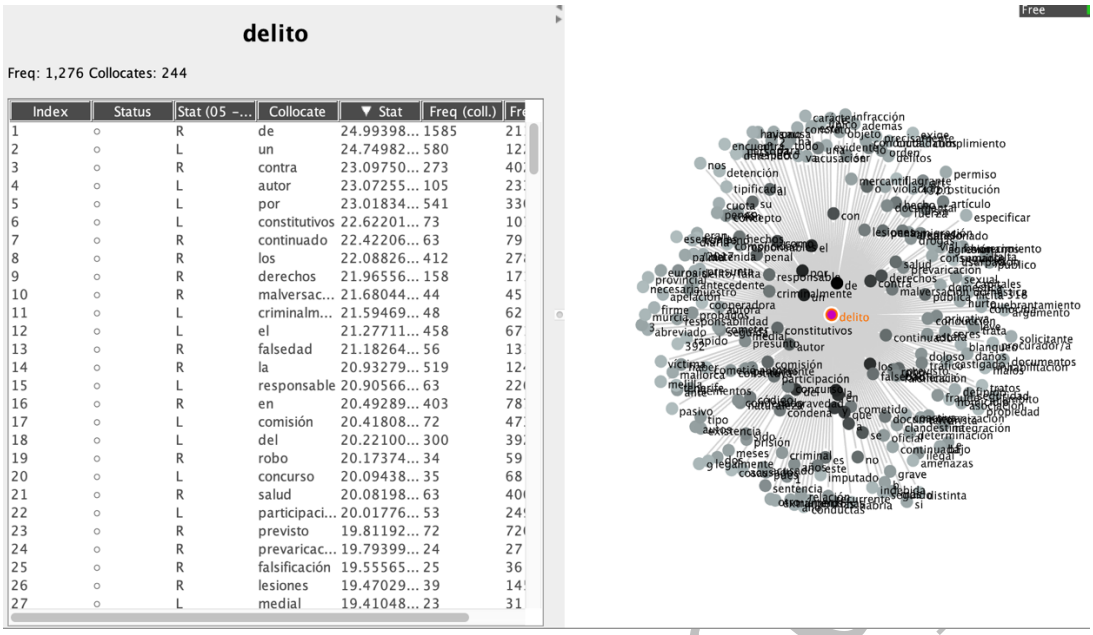


Table 6 Relevant UMTextStats categories

Category	Examples	Relative frequency
HOME	*doméstico/a/os/as* (*domestic*), *hogar(es)* (*home*, sg./pl.), *residencia* (*residence*)	0.74%
AFFILIATION	*abrazo/ar* (*hug*, sg./pl., as well as *to hold*, verbal conjugation), *amigo/a/os/as* (*friend*, sg./pl.)	0.64%
JOB	*abogado/a/os/as* (*lawyer*, sg./pl.), *enfermero/a/os/as* (*nurse*, sg./pl.), *pediatra(s)* (*pediatrician*, sg./pl.)	0.58%
RISK	*amenaza(s)* (*threat*, sg./pl.), *provocación(es)* (*provocation*, sg./pl.), *riesgo(s)* (*risk*, sg./pl.)	0.43%
FAMILY	*madre(s)* (*mother*, sg./pl.), *abuelo/a/os/as* (*grandmother/grandfather/grandparents*), *nuera(s)* (*daughter-in-law*, sg./pl.)	0.32%
CRIME	*homicidio(s)* (*homicide*, sg./pl.), *asesinato(s)* (*murder*, sg./pl.), *lesiones* (*assault and battery*)	0.03%
VIOLENCE	*puñetazo(s)* (*punch*, sg./pl.), *violento/a/os/as* (*violent*), *arma(s)* (*weapon*, sg./pl.)	0.01%