Interactive Analysis and Visualisation of Annotated Collocations in Spanish (AVAnCES)

Simon Gonzalez
The Australian National University / Canberra, ACT, Australia
u1037706@anu.edu.au

Abstract

Phraseology studies have been enhanced by Corpus Linguistics, which has become an interdisciplinary field where current technologies play an important role in its development. Computational tools have been implemented in the last decades with positive results on the identification of phrases in different languages. One specific technology that has impacted these studies is social media. As researchers, we have turned our attention to collecting data from these platforms, which comes with great advantages and its own challenges. One of the challenges is the way we design and build corpora relevant to the questions emerging in this type of language expression. This has been approached from different angles, but one that has given invaluable outputs is the building of linguistic corpora with the use of online web applications. In this paper, we take a multidimensional approach to the collection, design, and deployment of a phraseology corpus for Latin American Spanish from Twitter data, extracting features using NLP techniques, and presenting it in an interactive online web application. We expect to contribute to the methodologies used for Corpus Linguistics in the current technological age. Finally, we make this tool publicly available to be used by any researcher interested in the data itself and also on the technological tools developed here.

1 Introduction

Advances in current technologies have played a pivotal role in the development of academic fields, such as corpus-based phraseology. One of the most tangible results is the development of corpora based on digitised books (Michel et al., 2011), Google books (Zieba, 2018), and social media (Caselli et al.). Contributions from Corpus Linguistics have also been invaluable. Corpus Linguistics has been identified as one of the fastest growing linguistic methods in language studies (Abdumanapovna, 2018). This growth has gone hand in hand with advances in technologies, and it is clearly tangible in the tools that are now available for us as linguistic researchers to create exhaustive corpora to be accessed all around the world (a comprehensive list can be found in Tools for Corpus Linguistics). This has made Corpus Linguistics strongly dependent on the internet, where many websites have been deployed specifically for this purpose. All these factors render working with linguistic corpora a very interdisciplinary field, combining linguistics, data processing, data visualisation, and app development.

Another technological development that has influenced corpus-based phraseology has been the birth and development of social media platforms since the early 2000s. With these, we can create corpora that are based on natural language, from text to speech sources. Among these social media platforms, Twitter is one of the most influential ones and most widely used around the globe for the last two decades. A positive take on this is that Twitter offers free APIs that can be used to build tools for linguistic purposes. Researchers have made positive use of this and have maximised the potential to collect data and use it for language research (Dijkstra et al., 2021; Goel et al., 2016; Shoemark, 2020).

Within the field of Computational Linguistics, language studies have also found invaluable tools that have positively influenced the way we approach phraseology studies. Natural Language Processing techniques allow us to do a wide range of tasks on a large amount of data in relatively quick time. This has changed the focus from analysing small amounts of data, generally limited to the time human coders could process data, to processing massive amounts of data, where the limit is now on

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 21–30
November 20, 2022. ©2022 Association for Computational Linguistics

the computational capability.

Taking these technological contributions, namely social media, and open-source computational tools, we present in this paper the development of an online tool for the querying, analysis, and visualisation of collocations in Latin American Spanish based on a social media corpus. We discuss the emerging challenges when creating a corpus from social media and propose methodological processes appropriate for building digital language corpora to efficiently analyse collocations. The main motivation is to bring more depth to the presentation and analysis of linguistic patterns in a more interactive way. This type of implementation gives users powerful tools oriented towards finding patterns in available corpora. The final product aims to give researchers full control of the corpus by combining linguistic analysis, Natural Language Processing outputs, and visualisation techniques. With this holistic approach, we offer a deeper understanding of the complexity of collocations through exploration tools.

The goal of this research is then to present a new approach to analyse collocations. In this paper, we focus on Spanish, but this methodology can be used for any language that has outputs in social media platforms. We apply the analytical framework of Network Analysis to the study of collocations, and we also look at syntactic relationships and statistical measurements. In this sense, we aim to bridge the gap between the Continental tradition (Hausmann, 1991; Melcuk, 2007) and the British Contextualists tradition (Sinclair, 1991; Sinclair et al., 1970; Jones and Sinclair, 1974).

This paper is organised as follows. In Section 2, we present the technologies implemented in a more contextualised way, relevant to our study. We also present related work and our approach to the analysis and app development. We present the Methodology in Section 3 and the Analysis in Section 4. The Final product is presented in Section 5, with the Conclusions in Section 6.

2 Background and Rationale

The development of this new technology is created within three frameworks: Social Media, Computational Linguistics, and Internet Technologies. These are briefly discussed in the next sections below.

2.1 Social Media and Corpus Linguistics

One relevant premise in Corpus Linguistics is to collect reliable representative data, and this is achieved by selecting resources that allow language expression in a natural context (Abdumanapovna, 2018), and social media allows the study of language in contexts used for everyday communication (Rudiger and Dayter, 2020). This integration of social media on Corpus Linguistics is becoming more common practice, and it has been implemented, explored, and documented (Dunn, 2022; Rudiger and Dayter, 2020; Sun et al., 2021). Because of the complexity that social media language entails, it has not been widely explored, despite its prevalence in current communication processes (Sardinha, 2022). It has been therefore suggested to implement multidimensional (MD) analysis to approach the study of language in social media platforms, so we can capture its complexities. MD approaches were initially proposed by Biber (1988) and they are still widely implemented in current studies (Gardner et al., 2019; Jin, 2021; Sardinha, 2022). This method consists of analysing multiple linguistic characteristics of texts in a comprehensive way, examining a range of linguistic features across sources, which in turn helps identify correlations across features in whole corpora. The nature of this task requires the appropriate tools for achieving the correct results. That is why, the implementation of Natural Language Processing (NLP) tools helps in this methodological approach.

2.2 The Role of NLP in Corpus Linguistics

NLP allows Corpus Linguistics to have more statistical (Gerlach and Font-Clos, 2020; Lafferty et al., 2001; Manning and Schütze, 1999; Schmid, 1994) and machine learning (Karkaletsis et al., 2015) approaches to analyse language. This growing overlap between these two fields has experienced strong consolidation in the last decade. It is now common practice to implement NLP techniques in the design, modelling, and querying of linguistic corpora (Almujaiwel, 2018; Amri et al., 2017; Gentzkow et al., 2018), especially, in the analysis of linguistic forms within large datasets. This has positively contributed to more established corpus analysis approaches that focus on frequency counts, which helps us examine patterns of individual words and words in contact with other words. Other established methodologies that have been

reinforced with NLP techniques include analysis of collocations, n-grams, and word distributions. But NLP techniques can also provide other layers of analysis beyond word features. With NLP approaches, we can also analyse syntactic relationships and dependencies in sentences, examine semantic relationships, and automate identification of specific words in large corpora. A common application is the recognition of Named Entities, which consider textual distributions, word relationships, and syntactical positions. This is particularly useful when tagging geographic locations, proper names and institutions mentioned in the corpus. In summary, NLP tools are generally implemented for text chunking, word sense disambiguation, Named Entity Recognition, syntactic parsing, semantic role labelling, and semantic parsing (Amri et al., 2019). A clear advantage of NLP techniques is that they facilitate the quantification of features, which is the bases for statistical approaches to language data analysis. This does not substitute qualitative approaches to Corpus Linguistics, but rather complements the way we explore and analyse our linguistic data.

2.3 The Internet and Corpus Linguistics

The advancement of the internet and the computational power of current resources allow Corpus Linguistics to carry out tasks with intensive processing power and storage capacity. These help in both the processing and retrieval of large datasets (Abdumanapovna, 2018; Biber et al., 2006; Kennedy, 1998). In fact, Fisas et al. (2016) argue that this gives Corpus Linguistics more outcome feasibility and real-time access to corpora, regardless of physical location. The use of internet technologies has already been exploited for corpus purposes (Andersen, 2012; Collins, 2019; Hardie, 2012) and there are available corpora maximising this technology, e.g. *The Corpus of Contemporary American English (COCA)* (Davies, 2008), *The British National Corpus* (Clear, 1993), and the *Czech National Corpus* (Hnatkova et al., 2014).

2.4 Purpose of Current Corpus

The aim of our corpus is to capture the linguistic complexities of collocations in Spanish used on Twitter and explore the differences between the structures and patterns across users in thirteen Latin American countries. There has been a growing interest in linguistic studies using Twitter data for different purposes. The areas include phonological variation (Dijkstra et al., 2021; Eisenstein, 2013), stylistic and lexical variation on writing (Blodgett et al., 2016; Nguyen, 2017; Shoemark, 2020; Wurschinger, 2021; Pavalanathan and Eisenstein, 2015), dialectal studies (Eisenstein, 2017; Jorgensen et al., 2015), and language change (Goel et al., 2016). In this corpus, we prepare the data holistically, in such a way that it gives opportunities for users to focus their analysis on a wide range of linguistic features. This is explained in the following sections.

The focus of this study is on collocations, which can be defined as words occurring together in high frequencies with their semantic properties (Corpas-Pastor, 2017). In the computational sense, collocations are described as a distinct type of *multi-word expression (MWE)* which occurs in high frequency relative to the individual words that make the expression (Baldwin and Kim, 2010). In this sense, this is based on statistical quantification for all combinations (Jones and Sinclair, 1974; Stubbs, 2002). Apart from statistical approaches to identifying MWEs, other methods have been proposed in the literature. One of this is based on n-gram frequencies, also known as collocational networks. A limitation of this approach is that it can only identify continuous co-occurrences. The statistical approaches aim to overcome this limitation and are purposed to discover discontinuous co-occurrences. Hybrid models have therefore been developed to capture both continuous and discontinuous occurrences. These can combine measurements of linguistic features (e.g., semantic patterns), statistical calculations, and psychological approaches (Stefanowitsch, 2013). In this paper, we implement a multi-modal approach based on the hybrid models previously proposed, where we combine syntactic dependencies and n-gram patterns.

3 Methodology

Among other computational languages and software available, shiny R (Chang et al., 2019), within R (R Core Team, 2022), offers an invaluable infrastructure that, if well implemented, can facilitate the integration of the necessary methods mentioned above to produce high quality linguistic corpora. The app developed as part of this study and all its functionality were developed in R, which has been widely used for Corpus Linguistics development and related tasks (Abeille and Godard, 2000; S.Th., 2009). The main framework was within

Filter	Count	Percentage
URLs	$\sim 10,000$	1.3%
Re-tweets	$\sim 258,000$	35%
Quote tweets	$\sim 60,000$	8%
Non-Spanish tweets	$\sim 95,000$	13%
Less than 10 Words	$\sim 137,000$	19%

Table 1: Filters applied to the raw data, showing the type of filter, the total number of tweets filtered, and the percentage from the total extracted corpus.

shiny R. Shiny apps allow great interactivity and responsiveness. Interactivity allows users to explore visualisations in effective ways, and responsiveness allows users to navigate contents in real time, with the use of clicks and dropdown menus. Other libraries that we used for the creation of visuals were *ggplot2* (Wickham, 2016) and *echarts4r* (Coene, 2022). *echarts4r* is used to create a wide variety of interactive visuals, and *ggplot2* allows a great degree of flexibility when creating figures, which is relevant to explore complex linguistic data. But this allows complex ideas to be presented in a digestible way. Another advantage of this is that it allows users to see data points within the general context, as well as being able to narrow down into more specific analysis. This creates a seamless navigation of linguistic data in an efficient way.

3.1 Corpus

A preliminary research was done to identify relevant Twitter accounts to build the corpus from. For this, we aimed to choose Latin American users whose accounts had a relatively large number of posts. The reason was to gather as much data as allowed in the free API (3,250 tweets per account at a given moment). The filters below show that there is a lot of data that is lost to keep more comparable content. The second criterion was that the posts had to be in Spanish, and finally, the accounts had to be active at the moment of the data extraction. The motivation was to capture synchronous language use. This is especially relevant when analysing the use of phrases, which can be compared across sociolinguistically related groups of speakers in similar timeframes. Initially, there was a total of over 744,000 tweets. From this, we applied the filters presented in Table 1.

The final output was a total of 307,000 tweets. This is the main body of the corpus. For the demonstration of the app, we chose a subset of the whole corpus. Large corpora require substantial computational power to process the data in real time. For this reason, we selected approximately 17,000 sentences from the original corpus, distributed across all users from the thirteen countries. We left in only sentences with 15 to 17 words. The motivation was to select tweets with similar structures and character length. The final data contains 239 individual users, with an average of 73 sentences per user. The distributions per country and gender are shown in Table 2. Due to the limitations on the use of Twitter data for individual identification, account usernames are not presented, and the source data is not available for download. We only present analysis on the phrases, n-grams, and syntactic dependencies, which encompasses the aim of the tool. However, following Twitter regulations, we can only share the Tweet IDs as a request sent to the author of this paper.

Country	Females (120)	Males (119)
Argentina	210 (33%)	425 (67%)
Bolivia	513 (27%)	1397 (73%)
Chile	160 (28%)	410 (72%)
Colombia	711 (39%)	1130 (61%)
Costa Rica	745 (59%)	518 (41%)
Cuba	313 (31%)	703 (69%)
Ecuador	669 (49%)	680 (51%)
Mexico	762 (52%)	715 (48%)
Panama	848 (54%)	727 (46%)
Peru	437 (57%)	335 (43%)
Puerto Rico	606 (40%)	911 (60%)
Dominican Rep.	1177 (55%)	952 (45%)
Venezuela	633 (44%)	801 (56%)
TOTAL	**7784**	**9694**

Table 2: Total number of sentences per country and gender in the corpus.

The data extraction was done through an R script developed by the first author. We used the *rTweet* (Kearney, 2019) package, which allows users to gather Twitter posts by the free Twitter API. After collecting the data, the next step was the development of computational algorithms used to create linguistic annotations. This is described in the following sections.

3.2 Corpus Processing

The corpus was processed for two separate yet related tasks. The first one was to extract all the morphological and syntactic information. The main purpose was to give morphosyntactic infor-

Country	ADJ	ADP	ADV	AUX	DET	NOUN	PRON	VERB
Argentina	9%	20%	7%	5%	15%	24%	10%	14%
Bolivia	8%	21%	4%	4%	17%	26%	6%	14%
Chile	9%	20%	6%	4%	15%	25%	8%	13%
Colombia	8%	22%	5%	4%	16%	25%	7%	13%
Costa Rica	9%	22%	5%	4%	14%	24%	8%	14%
Cuba	9%	20%	5%	5%	16%	25%	7%	13%
Ecuador	8%	21%	5%	5%	15%	24%	8%	14%
Mexico	8%	22%	4%	4%	17%	25%	7%	13%
Panama	8%	21%	5%	4%	15%	25%	8%	14%
Peru	8%	19%	6%	4%	16%	23%	10%	14%
Puerto Rico	7%	23%	5%	4%	16%	25%	7%	13%
Dominican Rep.	8%	21%	4%	4%	17%	26%	7%	13%
Venezuela	8%	22%	5%	4%	15%	24%	8%	14%
TOTAL	**16129**	**42810**	**9735**	**8348**	**32122**	**49745**	**14649**	**26609**

Table 3: Total number and percentages of Parts of Speech per country in the corpus.

mation to collocations and the contexts in which they appear. The second task carried out statistical measurements on the collocations to be displayed through the corresponding visualisations.

3.2.1 Morphosyntactic Tagging

The morphosyntactic processing of this dataset was preprocessed outside the app and before launching it. For each sentence, we tagged each word and added their morphological and syntactic information. We implemented a wide range of NLP techniques for the data processing and analysis. The data was processed using the *UDPipe* (Straka and Strakova, 2017) package as the main tool for the NLP tasks. We used the *Spanish Ancora* model available in the package. The algorithm tokenises each sentence, identifies word lemmas, and then assigns a range of features based on the positions and functions of words in the sentence. Three main features extracted were the part of speech, morphological information (e.g., gender and number for nouns, tense and aspect for verbs), and their syntactic function in the given sentence (e.g., subject, object). The total distribution per country of Parts of Speech tagging is shown in Table 3. As observed, their distributions are similar across all countries.

3.2.2 Statistical Analysis of the Data

Unlike the morphosyntactic tagging, the statistical processing of this dataset is done interactively within the app. The user chooses the corresponding country, and then all the calculations are made. This is done following the processes from (Schweinberger, 2022) and using the *quanteda* (Benoit et al., 2018) and text mining – *tm* (Feinerer and Hornik, 2020) – packages. The first step is to concatenate all sentences in a single vector and then tokenise all words. From this point onwards, the process splits into two workflows. The first one is to calculate collocations across all words in the data, and the second one is to calculate all the collocations that can occur with a word selected in the app by the user. These processes are expanded below.

3.2.3 Overall Collocations Processing

In this process, the user first has the option to filter out stop words in Spanish using the *stopwords* (Benoit et al., 2021) package. The default option is to include stops words to capture collocations where stop words are included, for example, prepositions. We calculate the stats for the collocations running the function `textstat_collocations()` in the *quanteda* package, which calculates the lambda value as computed in Blaheta and Johnson (2001). Here, the user selects two parameters. The first one is the size of the collocations, e.g., number of words in the unit, from two to five. The second parameter is the minimum count. This refers to the number of times the collocation appears. The larger the size of the data, the more rigorous it can be to capture more frequent collocations. On the other hand, for smaller datasets, higher minimum counts could filter out relevant collocations. Here we maximise the power of interactivity, where users choose their

Collocation	Lambda	z
Golpe De Estado	4.87003	2.30594
Estado De Derecho	4.41775	2.04069
Democracia Y Libertad	3.1469	1.72487
Abuso De Poder	1.95931	0.87952
Libertad De Expresion	1.72231	0.73930
Poder Y Placer	1.24259	0.54335
Ministro De Gobierno	1.22566	0.7563
TOTAL	**7784**	**9694**

Table 4: Three-word collocations for tweets from Bolivia in the Overall Collocations.

parameters to better explore the corpus.

3.2.4 Word-based Collocations Processing

The first step in this process is to convert the sentences into a quanteda *Corpus* object. It contains the original sentences, document-level variables and metadata, corpus-level metadata, and features that are used for subsequent processing of the corpus. Like the **3.2.3 Overall Collocations Processing**, users can choose to filter out stop words. The non-optional filters are removing punctuation characters and numbers. This corpus is then converted to a *Document Term Matrix* object, which contains a sparse term-document matrix. This is a mathematical matrix that stores information on the frequency of terms that occur in the sentences, where rows correspond to the sentences in the collection and columns correspond to the terms. For statistical purposes, this is used to calculate co-occurrences counts from the word selected to all the other words in the data, as shown in Table 4. Table 5 shows the strength of specific words in relation to a reference word, which adds another layer of information for collocations.

4 Analysis and Visualisation

In this paper, we implement an analysis approach driven by visualisations of collocations. The visualisations are based on the mathematical measures done in the data processing stage, for both overall collocations and word-based selections. The driving approach is on ***Network Analysis (NA)***, which has been widely implemented in different fields, including causal distribution research (Kelly, 1983), archaeology (Golitko and Feinman, 1981; Orengo and Livarda, 2016), psychological studies (Jones et al., 2021; Mullarkey et al., 2019), and social network research (Clifton and Webster, 2017). The

Term	Strength	Term	Strength
abuso	18.42	consultiva	5.53
placer	12.85	quiso	5.53
estrategia	9.99	avenidas	5.53
segundo	9.99	casas	5.53
corrupcion	9.65	conductores	5.53
médicos	8.41	semáforo	5.53
horas	7.82	vuelven	5.53
opinión	7.82	sola	5.53
luis	5.80	públicos	5.53
ejerciendo	5.53	ruta	5.53

Table 5: Collocation strength for the term “poder” (”power”). Top 20 collocations shown. Note that the term “abuso” (”abuse) is the strongest term, and the strength stabilises at term “ejerciendo” (”exercising”).

main purpose of NA is to identify relationships within the components of a network. The assumption is that meaningful relationships between two or more elements will always reflect better and stronger connections than random or weaker relationships. The working components from which NA operates are based on relational data organised in a matrix form. This is where the relationship between the matrix output from the data processing and the methods in NA converge. We take the numeric output of the matrix and feed it into a network analysis visualisation function from the *visNetwork* (Almende, 2021) package. An example of a Network is shown in Figure 1.

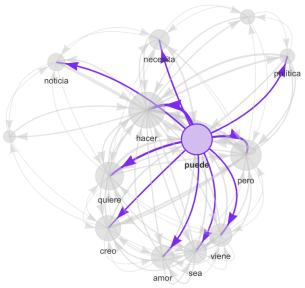

Figure 1: Network for the term “puede”.

4.1 Parts of Speech Networks

Network Analysis is also applied to the parts of speech tagging of the data. This can be used to observe relationships at the morphological level. It complements the analysis of collocations and provides another perspective to examine. Like in the collocations’ visualisation, we use the functionality from the *visNetwork* package, and users can change the parameters of analysis, including the number

of links between nodes, and the base frequency for all the tags, as shown in Figure 2.

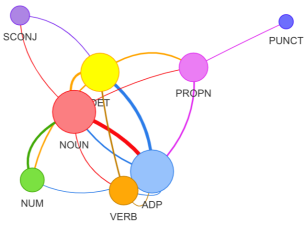

Figure 2: Network Analysis of Parts of Speech relationships in data selected.

4.2 Syntactic Dependencies

Another relevant implementation of the analysis targets syntactic dependencies. Here we use the output from the Morphosyntactic tagging step. The visualisation is done using the *textplot* (Wijffels et al., 2021) package. The main functionality of this package is to read the syntactic information from *UDPipe* outputs and then plot the dependencies in a text visualisation output. This can be done for all the sentences in the corpus. This is a powerful functionality that can be used to explore syntactic patterns of all collocations, and to understand all their contexts, as shown in Figure 3.

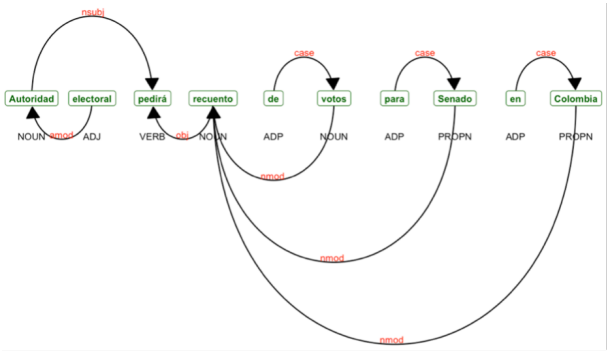

Figure 3: Syntactic Dependencies visualisation output, showing morphological and syntactic relationships between words.

4.3 Other Visualisations

Other visualisations are provided to examine a range of parameters that are important in understanding patterns and distributions of collocations in the corpus (See Figure 4). This gives users more tools to understand the patterns. These are presented in bar plots and radius pie charts from the *eachrts4r* package, which are used for examining of n-grams and parts of speech patterns.

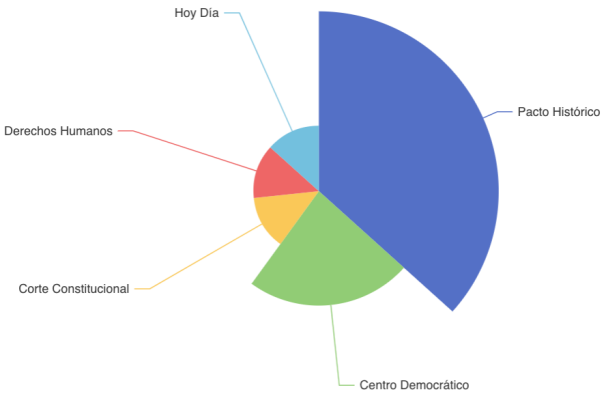

Figure 4: Radius Pie Chart of top five collocations within selected data.

5 Final Product

The final product is an app that gives users the opportunity to explore all the data, and the results from the different analyses. The code and application can be accessed through the GitHub repository: https://github.com/simongonzalez/AVANCES. The app is organised into five main sections. The first one is the visualisation of the distributions of speakers based on countries and occupations in the data. The second section shows the distributions of n-grams and parts of speech through network visualisations, pie charts, and bar plots. The third section presents results from the Network Analysis, looking at overall and word-based collocations. The fourth section shows the syntactic dependencies plots, and the sentences are selected by the user. The fifth and final section has a searching capability. In this tab, users can search for syntactic patterns in the data. The source tagging comes from the *UDPipe* output, showing the morphosyntactic patterns. The main usability is to allow users to identify in advance the potential sequences that can be relevant to explore in more depth. All these five sections then gather all the pre-processed data and also process the data based on user requests. This gives a full control on the data processing to have sophisticated exploration tools.

6 Conclusions and future work

In this paper, we have presented the development and deployment of a Spanish linguistic corpus built from Twitter posts. We combined NLP techniques, linguistic analysis, and app development approaches to create a holistic framework to analyse and explore collocations across Twitter users from thirteen Latin American countries. In future

versions of the app, we aim to include more language features, as well as more data from other Spanish-speaking countries. We also aim to carry out more linguistic analysis relevant for corpus research, such as language variation, stylistics, sentiment analysis, for example. Finally, this is an open-source tool with the potential to be expanded and customised based on user needs.

References

S.A. Abdumanapovna. 2018. The contemporary language studies with corpus linguistics. In *Proceedings of the 2nd International Conference on Digital Technology in Education (ICDTE 2018)*, pages 82–85, New York, NY, USA.

A. Abeille and D. Godard. 2000. French word order and lexical weight. *Syntax and Semantics*, pages 325–358.

B.V. Almende. 2021. visnetwork: Network visualization using 'vis.js' library. *R Package*. Version 2.1.0.

S. Almujaiwel. 2018. Integrating nlp with corpus linguistics and vice versa. In *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications (LOPAL '18)*, pages 1–6, New York, NY, USA.

S. Amri, L. Zenkouar, and R. Benkhouya. 2019. A comparative study on the efficiency of pos tagging techniques on amazigh corpus. In *Proceedings of the 2nd International Conference on Networking, Information Systems Security (NISS19)*, pages 1–5, New York, NY, USA.

S. Amri, L. Zenkouar, and M. Outahajala. 2017. Build a morphosyntaxically annotated amazigh corpus. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (BDCA'17)*, pages 1–7, New York, NY, USA.

G. Andersen. 2012. *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, volume 49. John Benjamins Publishing.

T. Baldwin and S.N. Kim. 2010. Multiword expressions. *Handbook of Natural Language Processing*, pages 267–292. 2nd edn.

K. Benoit, D. Muhr, and K. Watanabe. 2021. stopwords: Multilingual stopword lists. *R package*.

K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30).

D. Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

D. Biber, Conrad S., and R. Reppen. 2006. *Corpus Linguistics: investigating language structure and use*. Cambridge University Press.

D. Blaheta and M. Johnson. 2001. Unsupervised learning of multi-word verbs. In *ACLEACL Workshop on the Computational Extraction, Analysis and Exploitation of Col-locations*.

S.L. Blodgett, L. Green, and B.T. O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben mmerman, and Malvina Nissim. Dalc: the dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. 2019. shiny: Web application frame-work for r. *R Package Version 1.3.2*.

J.H. Clear. 1993. *The British national corpus. The digital word: text-based computing in the humanities*. MIT Press.

A. Clifton and G. D. Webster. 2017. An introduction to social network analysis for personality and social psychologists. *Social Psychological and Personality Science*, 8(4):442–453.

J. Coene. 2022. echarts4r: Create interactive graphs with 'echarts javascript'. *R Package Version 5*.

L.C. Collins. 2019. *Corpus Linguistics for Online Communication: A Guide for Research*. Routledge.

G. Corpas-Pastor. 2017. Collocational constructions in translated spanish: What corpora reveal. In *Euphoria 2017, LNAI*, pages 29–40.

M. Davies. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*.

J. Dijkstra, W. Heeringa, L. Jongbloed-Faber, and H. Van de Velde. 2021. Using twitter data for the study of language change in low-resource languages. a panel study of relative pronouns in frisian. *Frontiers in Artificial Intelligence*.

J. Dunn. 2022. *Natural Language Processing for Corpus Linguistics (Elements in Corpus Linguistics)*, volume 1. Cambridge University Press, Cambridge.

J. Eisenstein. 2013. Phonological factors in social media writing. In *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, pages 11–19.

J. Eisenstein. 2017. Identifying regional dialects in online social media. *The Handbook of Dialectology*.

I. Feinerer and K. Hornik. 2020. tm: Text mining package. *R Package Version 0.7-8*.

B. Fisas, F. Ronzano, and H. Saggion. 2016. A multilayered annotated corpus of scientific papers. In *LREC*.

S. Gardner, H. Nesi, and D. Biber. 2019. Discipline, level, genre: Integrating situational perspectives in a new md analysis of university student writing. *Applied Linguistics*, 40(4):646–674.

M. Gentzkow, J. Shapiro, and M. Taddy. 2018. Congressional record for the 43rd–114th congresses: Parsed speeches and phrase counts (tech. rep.). In *Palo Alto, CA: Stanford Libraries*.

M. Gerlach and F. Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126–132.

R. Goel, S. Soni, N. Goyal, J. Paparrizos, H.M. Wallach, F.D. Diaz, and J. Eisenstein. 2016. The social dynamics of language change in online networks. *ArXiv*.

M. Golitko and G. M. Feinman. 1981. Procurement and distribution of prehispanic mesoamerican obsidian 900 bc-ad 1520: A social network analysis. *Journal of Archaeological Method and Theory*, 22(1):206–247.

A. Hardie. 2012. Cqpweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

F.J. Hausmann. 1991. Collocations in the bilingual dictionary. In *An International Encylopedia of Lexicography*, pages 2775–2778.

M. Hnatkova, M. Kren, P. Prochazka, and H. Skoumalova. 2014. The syn-series corpora of written czech. In *Proceedings of LREC2014*, pages 160–164.

B. Jin. 2021. A multi-dimensional analysis of research article discussion sections in an engineering discipline: Corpus explorations and scientists' perceptions. *SAGE Open*, 28(1):114–133.

P.-J. Jones, R. Ma, and R.-J. McNally. 2021. Bridge centrality: A network approach to under-standing comorbidity. *Multivariate Behavioral Research*, 56(2):353–367.

S. Jones and J. Sinclair. 1974. English lexical collocations. a study in computational linguistics. *Cahiers de Lexicology*, 24:15–21.

A.K. Jorgensen, D. Hovy, and A. Sogaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 9–18.

G. Karkaletsis, G. Petasis, and V. Paliouras. 2015. *Using machine learning techniques for part-of-speech tagging in the Greek language*. World Scientific Publishing Company, Singapore.

M.W. Kearney. 2019. rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software*, 4(42). 0.7.0.

H. H. Kelly. 1983. Perceived causal structures. *Attribution theory and research: Conceptual, developmental and social dimensions*.

G. Kennedy. 1998. *An Introduction to Corpus Linguistics*. Longman, London.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the ICML-01*, pages 282–289.

C. Manning and H. Schütze. 1999. *oundations of Statistical Natural Language Processing*. The MIT Press.

I. Melcuk. 2007. Lexical functions. *Phraseology. An International Handbook of Contemporary Research*, 1:119–131.

J. Michel, Y. Shen, and et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

M.-C. Mullarkey, I. Marchetti, and C.-G. Beevers. 2019. Using network analysis to identify central symptoms of adolescent depression. *Journal of Clinical Child Adolescent Psychology*, 48(4):656–668.

D. Nguyen. 2017. *Text as social and cultural data : a computational perspective on variation in text*. University of Twente, The Netherlands. PhD Dissertation.

H.A. Orengo and A. Livarda. 2016. The seeds of commerce: A network analysis-based approach to the romano-british transport system. *Journal of Archaeological Science*, 66:21–35.

U. Pavalanathan and J. Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90:187–213.

R Core Team R Core Team. 2022. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

S. Rudiger and D. Dayter. 2020. *Corpus Approaches to Social Media*. John Benjamins.

T. B. Sardinha. 2022. Corpus linguistics and the study of social media: a case study using multi-dimensional analysis. *The Routledge Handbook of Corpus Linguistics*, pages 656–674.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

M. Schweinberger. 2022. Analyzing co-occurrences and collocations in r. *R Tutorial*. 2022.05.04.

P. Shoemark. 2020. *Discovering and analysing lexical variation in social media text*. The University of Edinburgh. PhD Dissertation.

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

J. Sinclair, S. Jones, and R. Daley. 1970. *English lexical studies: report to OSTI on project C/LP/08*. Department of English, University of Birmingham. Final report for period January 1967-September 1969.

A. Stefanowitsch. 2013. Collostructional analysis. *The Oxford Handbook of Construction Grammar*, pages 290–306.

Gries. S.Th. 2009. *Quantitative Corpus Linguistics with R*, volume 1. Routledge, London and New York.

M. Straka and J. Strakova. 2017. Pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

M. Stubbs. 2002. Two quantitative methods of studying phraseology in english. *International Journal of Corpus Linguist*, 7(12):215–244.

Y. Sun, G. Wang, and H. Feng. 2021. Linguistic studies on social media: A bibliometric analysis. *SAGE Open*.

H. Wickham. 2016. ggplot2: Elegant graphics for data analysis. *R Package*.

J. Wijffels, S. Epskamp, I. Feinerer, and K. Hornik. 2021. textplot: Visualise complex relations in texts. *R Package*. 0.2.0.

Q. Wurschinger. 2021. Social networks of lexical innovation. investigating the social dynamics of diffusion of neologisms on twitter. *Frontiers in Artificial Intelligence*.

A. Zieba. 2018. Google books ngram viewer in sociocultural research. *Res. Lang.*, 16:357–376.