Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the article:

García Pellicer, Remo; Fernández Montraveta, Ana, dir. Corpus Linguistics : Developing a Multianalysis Text Tool. 2021. 37 pag. (1482 Grau en Estudis Anglesos 801 Grau en Estudis Anglesos)

This version is available at https://ddd.uab.cat/record/248824

under the terms of the BY-NC-ND license

Universitat Autònoma de Barcelona

DEPARTAMENT DE FILOLOGIA ANGLESA I DE GERMANÍSTICA

Corpus Linguistics: Developing a Multianalysis Text Tool

Treball de Fi de Grau

Author: Remo Garcia Pellicer

Supervisor: Ana Fernández Montraveta

Grau d'Estudis Anglesos

June 2021

ACKNOWLEDGEMENTS

Thanks to my supervisor, Ana Fernández, for guiding me through the process of completing my TFG.

TABLE OF CONTENTS

Index of figures ii

Abstract 1

1. Introduction 2

2. A brief history of corpus linguistics 3

3. The use of corpora in different fields of research 6

3.1 Corpora in translation 6

3.2 Corpora in English language teaching 7

4. Methodology of corpus creation 8

4.1 Annotation layers 9

4.2 Providing multiple analysis tools within corpora 11

5. Building an original corpus structure 13

5.1 Collecting and storing text from web data 14

5.2 Adding annotation layers 14

5.3 Adding analysis tools 15

6. Conclusion 17

References 18

Appendix 19

INDEX OF FIGURES

Figure 1. Corpora growth in number of words since the 1960s.....5

Figure 2. Results of a bigram search of the word *military* in a piece of news.....12

Figure 3. Results of a concordance search of the word *military* in a piece of news.....13

Figure 4. Results of a sentence parsed through *Spacy* from a piece of news.....15

Abstract

This thesis presents an overview of Corpus Linguistics, highlighting how the advancements in the field have influenced English Language Teaching. It is a well-known fact that textbooks for foreign language students do not present real examples of language use but rather prefabricated texts and dialogs adapted to the class level. In previous research, the advantages of allowing students to access corpora in class have already been discussed. However, corpus resources are not always easily or freely accessible for students or teachers.

I aim to contribute with a new tool that can be used in English classes. It allows both teachers and students to create corpora based on news articles available online and it incorporates a variety of analysis tools, such as a part of speech tagger and a syntactic parser that allow users to visualize the syntactic relations in sentences. Furthermore, the front end of the software is in website format to make it as accessible as possible. As a result, the use of the tool proves to be interactive and engaging: a simplified design makes it highly user-friendly, and all the features it provides make it very versatile. In addition, the program can be used as a data bank for the development of materials for future lessons, allowing the teacher to find new examples and prepare exercises beforehand.

Keywords: Corpus linguistics, English Language Teaching, Corpus analysis, Web scraping, Part of speech tagger, Syntactic parser, Concordance.

1. Introduction

Corpus linguistics is the study of language used in a real context. The present study is not only centred around corpus linguistics itself and its history, but also around the different disciplines impacted by the advancements in corpus research, such as translation studies or English Language Teaching (ELT). The main goal is to delve into the importance of using corpora in any field of study related to linguistics, with a particular focus on the use of corpora in the classroom. After all, English learners do not usually have access to real instances of the language. Instead, classes are typically based on textbooks, which present prefabricated language, with examples stripped from all context. I would argue that using corpora to extract real examples or create genuine exercises would benefit students in order to be exposed to real English.

In addition, the study also describes how metalinguistic information may be added onto the corpus by incorporating layers of annotation apart from the raw text itself. The main analyses reviewed are part of speech tagging and parsing, along with built-in tools to search for collocations and visualize specific elements in context and their frequency of appearance.

Lastly, the program developed in this work (Annex 1 and Annex 2) is provided. The tool allows users to interact with the corpus, allowing them to enlarge the corpus with new texts extracted from news articles. The tool I have created is meant to be easily accessible and designed to be used in ELT classes, providing both students and teachers with sentences from real contexts that may then be analysed in a variety of ways.

The study is divided into four main sections. Section 2 presents an overview of the history of corpus linguistics, mainly during the twentieth century. Section 3 deals with

the use of corpora in various fields of research. In Section 4, the process of building a corpus is described, emphasizing on the extra layers of linguistic information that can be used to enrich corpora. Lastly, Section 5 contains the program's implementation, providing a detailed description of how the tool operates.

2. A brief history of corpus linguistics

A corpus is a selection of texts which share a text type and have a particular set of criteria in common. Corpora have been around for centuries, as texts are to be compiled and preserved in order to withstand the passage of time. Some of the earliest manual corpora date back to the Vedic era. Pratishakhya literature, also known as Parsada, were texts written in Sanskrit which were meant to educate readers on the correct pronunciation of words. Had it not been for Parsada, Vedic texts would have arguably remained a mystery to modern scholars. In addition, as Parsada dealt with the phonetics of Sanskrit, it has allowed researchers to analyse the Vedas in a very precise way, allowing them even to replicate the ritual recitations of the ancient Indian religious texts.

Corpora and their many applications have been studied by philologists ever since. However, the accessibility of texts and the nature of corpora remained unaltered for a long period of time. It was not until the twentieth century, more specifically the introduction of computers, that the field of corpus linguistics was able to make its huge leap forward. The technological advancements that computers brought in terms of information processing allowed for corpora to be compiled and accessed more straightforwardly, thus making it possible for corpora to be made larger than ever before. What is more, the enlargement of corpora did not result in a more restricted or slower

access to texts, but the other way around. Electronically compiling and accessing data drastically changed the course of corpus linguistics.

The first corpora stored and accessed electronically, were developed during the 1960s, the most recognized one from that time being the Brown Corpus (Brown) from the Brown University in Rhode Island, USA. The main goal of the Brown Corpus was to store a corpus consisting of around a million words, all of which had been taken from American publications from the year 1961. After the Brown Corpus, other corpora began to take form during the following decades. During the 1960s and 1970s, the size and structure of corpora tended to follow that of the Brown Corpus, i.e., corpora always consisted of around a million words. During those two decades, the most notorious ones were the Lancaster-Oslo-Bergen Corpus (LOB) in 1978, compiled by Johansson et al., which had also been tagged, i.e., annotated with part-of-speech identifications, and the English for Science and Technology corpus in 1985 (JDEST), compiled by Yang Hui-Zhong from the Shanghai Jiao Tong University.

It was not until the 1990s that researchers were able to make even larger corpora, mainly due to the limited storage capabilities of computers prior to that date. The Bank of English (COBUILD) corpus was then launched in 1991, a project initially led by Professor John Sinclair, after its compilers had been adding texts to it for a couple of years. Corpora containing a million words had been considered huge but became small corpora instead by the time the Bank of English corpus was released, as it contained more than 400 million words by that time. What is more, even such a number of words within a corpus grows pale if compared to the capabilities of data storage of contemporary computers. Corpora from the 21st-century such as the Oxford English Corpus (OEC), which has been compiled by the authors of the Oxford English Dictionary, contains within

it around 2.1 billion words, which originate from a variety of sources such as news, magazines or novels. The growth in number of words and the steep increase that corpora have been able to withstand once the necessary technology has been able to support such a feat can be seen in the following Figure 1, which shows the size of the aforementioned corpora:

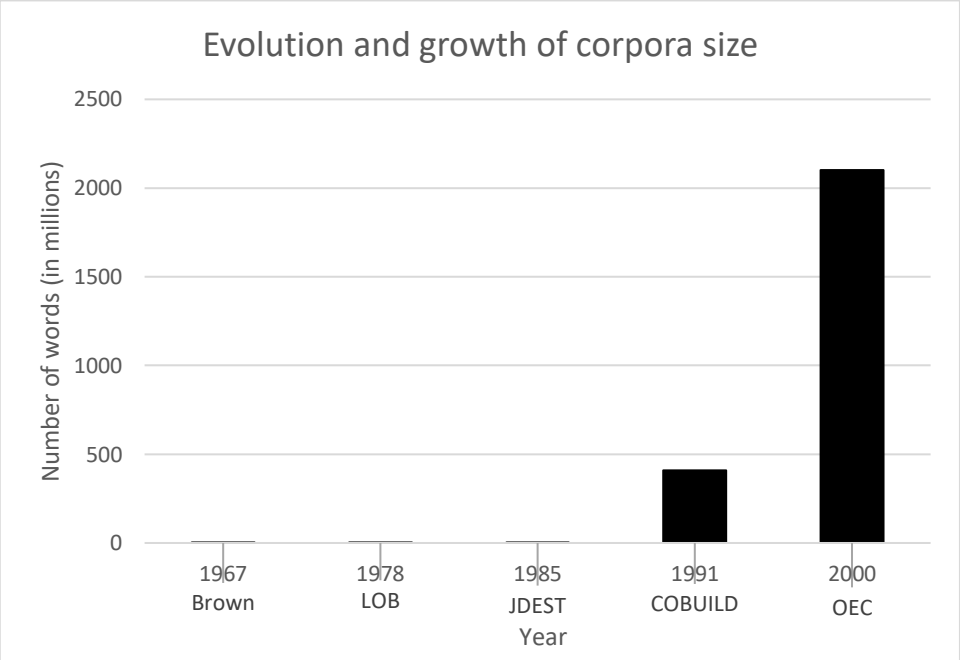


Figure 1. Corpora growth in number of words since the 1960s.

Sinclair observes that corpus size varies with time and that bigger corpora are more desirable than small corpora. In his words:

> So there is a kind of relativity in corpus sizing – the dimensions of a "small" corpus vary with the date it is compiled; the apparently massive corpora of a few years ago are now perceived as tiny, and in another decade or two, anything less than a few billion words will count as a small corpus, because there is every reason to make bigger and bigger corpora, and the job becomes easier as the size goes up. (Ghadessy, Henry and Roseberry 2001: ix).

3. The use of corpora in different fields of research

The wide availability of corpora and the modern way of accessing texts allows for any discipline in which language needs to be taken into account to make considerable progress. For example, the field of translation has been hugely influenced by the advancements of corpora.

Nevertheless, translation is not the only discipline which requires new and updated tools for corpus compilation and access. Sociolinguistics or ELT, among many others, is also in demand of innovative corpora resources and text analysis tools. For example, corpora may provide teachers with new angles to organize their materials or even their entire in-class lessons.

3.1 Corpora in translation

Baker describes the impact that corpus research has had and will continue to have on translation studies as follows:

> There is no doubt that the availability of corpora and of corpus-driven methodology will soon provide valuable insights in the applied branch of translation studies, and that the impact of corpus-based research will be felt there long before it begins to trickle into the theoretical and descriptive branches of the discipline. (1993: 242).

In applied translation, the most innovating field of study which has been affected by corpus linguistics would be automatic or computer-assisted translation. Advancements in corpus linguistics have allowed researchers to develop contemporary text treatment tools that make computer-assisted translation possible, e.g., concordancers. However, advancements in machine translation not only have an impact on the quality of automatic translation, but also in all translated texts in general, as the data that is obtained from

translated corpora is, in turn, bringing researchers closer to describing the natural structure of languages. As Sinclair states it:

> The new corpus resources are expected to have a profound effect on the translations of the future. Attempts at machine translation have consistently demonstrated to linguists that they do not know enough about the languages concerned to effect an acceptable translation. In principle, the corpora can provide the information. (Sinclair 1992: 395, cited in Baker 1993: 242).

3.2 Corpora in English language teaching

A possible and valuable way of incorporating corpus access in the classroom would be gathering data on the vocabulary ranges of learners and then comparing it against a small corpus of word lists and families, thus being able to evaluate the lexical richness within the writing production of the learners. The results may then guide the teachers, shedding some light on the particular vocabulary needs of their students. Additionally, if access to such corpora is also partially given to students, they would then be able to notice their own mistakes themselves and focus on reinforcing those areas of study. As is suggested by Nation in his study of vocabulary richness in learners' written production, "teachers can use the learners' compositions with the words marked up according to their frequency level as a way of commenting on learners' vocabulary use in their writing" (Nation 2001, cited in Ghadessy et al. 2001: 43)

Furthermore, the design of ELT lessons can be based, to a certain extent, on annotated corpus data. After all, English courses, be it English as a First Language (EFL) or English as a Second Language (ESL), are not meant to be focused around teaching all possible vocabulary items to students, as it is estimated that native speakers know an average of around 17000 words (Goulden, Nation, Read 1990). On the contrary, English courses tend to educate their students on fewer words. Sinclair and Renouf (1988)

analysed a variety of EFL courses in order to show that learners were taught between 1156 and 3963 different word forms in their books. Then, the focus of English learning courses is not to be set on the quantity of vocabulary to be taught but rather on the quality. Corpora have proven to be very useful when it comes to selecting which words are more likely to become useful for learners. As Flowerdew explains it:

> A decision must therefore have been made to include only a fraction of the words known by the average native speaker. However, the great power of the corpus-based word list is in that the course designer can be sure that the words selected are the most useful (i.e. the most frequently used). (Ghadessy et al. 2001: 75).

4. Methodology of corpus creation

As already seen in the previous section, the use of corpora in research is highly valuable. The debate is then focused on the analyses which may be facilitated along with corpora and the value of annotated corpora.

The following sections will be focused on the most used annotation layers, which provide lots of helpful information while keeping text easy to read. Once the relevance of annotation and corpus size has been established, the next valuable aspect that should be considered when building a corpus is the number of analysis tools provided with it. After all, as already mentioned, information within big corpora, such as the OEC, may be too much for researchers to manually handle. When it comes to annotation layers, I will mainly focus on both part of speech tagging and parsing, which might be the most valuable resources for ELT, as the tool provided in this study is meant to be used in the classroom.

4.1 Annotation layers

Some researchers see raw corpus data as easier to read, keeping language patterns more visually evident to readers. On the other hand, other scholars claim that annotated corpora can prove to be more useful, as annotation allows them to incorporate an automatic linguistic model within the corpus without having to notice the patterns manually. Additionally, annotation is arguably a necessary step for researchers to be able to test and contrast their linguistic theories against the compiled data, as manually analysing the information could prove to be unfeasible for researchers, depending on the size of the corpus. However, researchers who claim that an unannotated corpus is more valuable would describe the annotation processes as contamination of the originality of data with preconceived ideas of how it linguistically operates. The solution to the debate, however, is rather simple. As Anthony describes it:

> Again, the debate on the value of annotation can be easily resolved by refocusing the discussion on the tools used to analyze corpora. Modern corpus tools are easily able to show or hide different layers of annotation or markup of texts. If a researcher would like to analyze the raw texts, the various layers of annotation can be hidden. On the other hand, if a researcher needs to count verb tenses or any other linguistic feature and has tagged the corpus for them, the software can then utilize this additional information and provide the researcher with a result almost instantaneously. (2013: 148).

In other words, corpora should consist not only of the raw data within text files, but also of the various layers of annotation which different fields of research may be interested in. Configuring corpora in such a way makes it possible for researchers to turn the variety of annotated layers on or off at their will, allowing for a healthy way of contrasting the information that the naked eye may be able to notice against the automatic information that the corpus analysis tools may have found during the compilation process.

Arguably, the most valuable annotation layer would be part of speech tagging, which is a process in which individual words within the corpus are marked depending on the part of speech, i.e., the grammatical category, to which they correspond. The tags which are applied to each element are dependent not only on the definition of the words themselves, but also on the context in which they appear. Consequently, part of speech tagging depends on rule-based algorithms that can figure out in which context a given element is set, as some words may belong to more than one category depending on their environment. For example, in the case of the Brown Corpus, the rules for determining which part of speech corresponded to each element were manually typed into the program, by providing a list of words with different part of speech tags within a variety of different contexts. From then on, those initial rules were evaluated and corrected throughout the years, improving the corpus' tagger performance with every revision until the 1970s, as by then, the annotated layer of tags it provided mainly was accurate and made very few mistakes, if any.

The second most valuable layer of annotation in corpora would arguably be parsing, as it allows for an initial analysis of the syntactic structure of the texts within a corpus. In corpus linguistics, parsers are extremely useful for visualizing the syntax of sentences within texts, as they provide a layer of information beyond tags, being able to check how the words in a given sentence are related to each other. Once a given text has been parsed, it can then be either manually or automatically converted into a syntax tree, thus allowing the user to visually represent the structure of a given sentence within the corpus. Figure 2 below shows a sentence which has been inputted through a parser and then automatically converted into a tree. For researchers to be able to visually represent the structure of sentences within corpora is highly desirable, as it makes it easier for the

user to have an initial syntactic analysis which can then be used for a variety of purposes, be it a grammar class or research itself. However, as it happens with taggers, parsers need a previous set of rules from which they can learn how to analyse new sentences which are fed to them. In the case of parsers, those rules are embedded within a grammar.

One possible way for a corpus to be able to automatically decide which clause analysis to provide for a given element is for it to have been previously fed annotated and parsed data so that the system can then draw its conclusions on the syntax of sentences from previous information which it can already value as correct. In other words, the rules for syntactic analysis are to be deduced from previous examples. Nonetheless, statistical deep learning is not the only option for establishing grammars. Researchers may also manually type in the needed grammar for a given sentence every time they want said sentence to be syntactically analysed. However, as that method would not really require a layer of parser annotation, it would not be able to provide analyses automatically, but instead manually, and also a grammar would need to be constructed a priori for every sentence.

4.2 Providing multiple analysis tools within corpora

Apart from annotation layers, there exist other resources that would allow users to perform quick searches for particular words within the corpus. One of the possible purposes of such searches would be finding lexical bundles, for example. Lexical bundles, or n-grams, allow users to check bundles of words which typically appear together. The result of an n-gram search, or lexical bundle search, is any given sequence of n words in which one of those elements is the inputted word. Information returned by n-grams may then be analysed to reveal which are the most frequently appearing words around the

inputted element, thus allowing users to see which are the possible collocations of the target word. Figure 3 below corresponds to an example of an n-gram search within a piece of news. The number to the right of each pair is indicating the frequency of the occurrence of said pair, unveiling that the word *military* appears mostly next to *Israeli* in this text. However, not all high frequency pairs are collocations. This piece of news in particular is about Israel's military operations, thus the high frequency of the pair *Israeli military*. In other words, frequencies of n-grams do not provide enough information on their own for linguists to decide which elements are collocations.

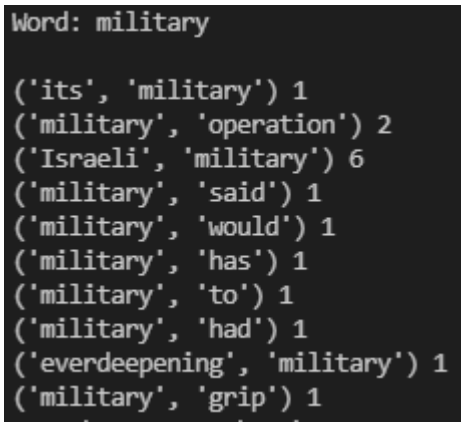


Figure 2. Results of a bigram search of the word *military* in a piece of news. (The Guardian - Israel, May 12)

As previously mentioned in section 2.1, another analysis tool that is highly valuable in corpora would be concordancing, especially when said corpora are used for translating purposes. The results of a concordancer search work in a similar way as n-grams, in the sense that they allow the user to understand better the context of a given word in a particular text or corpus. However, instead of just returning a limited bundle of words which appear next to the element being searched, concordancer searches return a fraction of the sentence in which the word appears, i.e., the immediate context of the word that is being searched or analysed. Such immediate contexts of certain words have proven

to be highly desirable for translation systems, as concordancers are most often used as a first step within computer-assisted or machine translation. Figure 4 below corresponds to an example of a concordancer search within a piece of news, more specifically, the same piece of news as in Figure 3. Similarly, it can be seen that the word *military* appears right next to Israeli most of the times, as was already revealed by the n-gram search. However, the concordancer provides a more significant amount of the context in the close vicinities of the word.

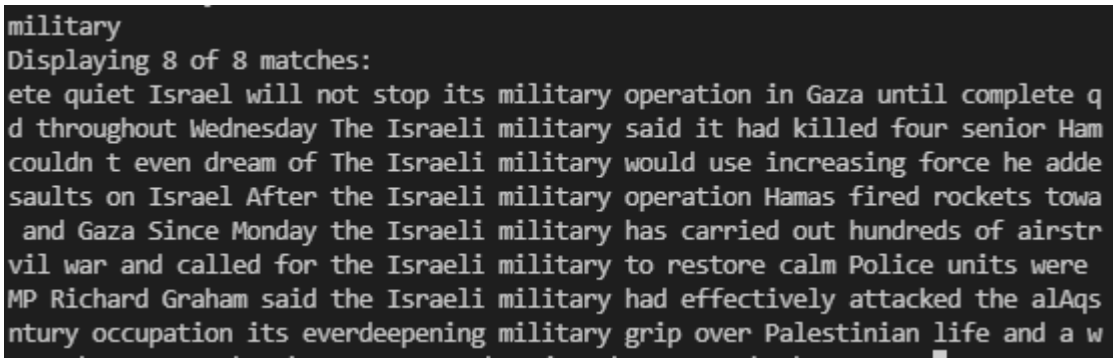


Figure 3. Results of a concordance search of the word *military* in a piece of news. (The Guardian - Israel, May 12)

5. Building an original corpus structure

Natural English is found within corpora, as corpora are collected in their actual contexts without any external interference. The program brings natural English close to the users, as they may interact with text from any given piece of news, being able to analyse it in detail. When it comes to the program's details, it has been deployed on a cloud server (see Annex 1) and is open to the public at Open Corpus. The advantages the program brings are that changes done to the project may be deployed onto the server in an easy and fast manner. Moreover, the server is compatible with the market's latest technologies. Therefore, the project may be developed in any programming language and still be compatible with the server. What the program does better than others, is include

a wide range of corpus analysis tools while maintaining the interface extremely simple and easy to use, keeping the interaction as user-friendly as possible.

5.1 Collecting and storing text from web data

After establishing the importance of both having layers of annotation and multiple analysis tools within a corpus, this section of the study will focus on describing the corpus building program that has been developed to exemplify the range of possible tools that corpora may use incorporate. The interface is shown in detail in Annex 1. Naturally, the first step in corpus building would be collecting the texts which are to be studied. In this case, texts are scraped from web data. To be more specific, this corpus deals with news which may be extracted from *The Guardian*'s website, a British daily newspaper founded in the nineteenth century. The program's input is an URL, i.e., a web address from *The Guardian*, from which it can extract both the title and the body of the piece of news given to it. However, it does not work with links to live news. It then stores all URLs which are inputted through the program into the database as raw text. After the text data has been compiled into the corpus, the next logical step is to add annotation layers and a variety of tools to analyse the data.

5.2 Adding annotation layers

The second utility which this corpus brings is adding annotation onto the text which has been downloaded. As described in section 3 within this paper, the most valuable annotation layers are arguably parsing and tagging. The program allows the user to process the raw text data through both a part of speech tagger and a parser. When it comes to the tagging process, the program's rules for identifying different parts of speech

are based on the *Natural Language Toolkit* (*NLTK*) Python library (Bird, Loper and Klein 2009). Once the words within the piece of news have been tokenized, they are analysed according to *NLTK*'s rules and are given a particular tag to each word. The parser's rules, however, are built based on *Spacy* (Honnibal and Montani 2017). It is a more modern python library that provides a straightforward parser that may be used to visualize some of the syntactic relations between words in a given sentence. The input of *Spacy*'s parser is the sentence itself, without needing any previous information, as the parsing function itself is already able to tag and parse words within a sentence. Figure 5 below shows an example of a sentence that has been parsed through the S*pacy* library. Instead of drawing a syntax tree, what *Spacy* does is show basic syntactic relations between words. It is based on a trained English corpus annotated with examples of syntactic relations within the sentences it contains.

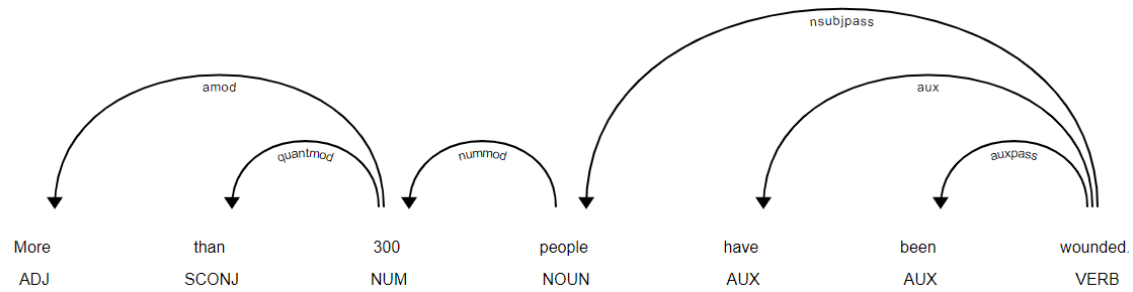


Figure 4. Results of a sentence parsed through *Spacy* from a piece of news. (The Guardian - Israel, May 12)

5.3 Adding analysis tools

Apart from the raw text data and the layers of part of speech tags and parsed sentences, this corpus also provides the user with various tools for analysing the texts within it. It contains all of the analysis tools previously described in section 3, namely the n-gram bundle searcher and the concordancer. They are both also built based on the *NLTK*

Python library (Bird et al. 2009), and allow the user to see the word searched in its immediate context, be it a lexical bundle or the sentences themselves. Lastly, this corpus also has the feature of displaying word occurrences within a given piece of news, ordered by their frequency of appearance. Similarly, the frequency counter is also based on *NLTK*, as the library provides a list of words that do not carry any lexical information, i.e., stop words. Such words would not be of any interest for the frequency counter, as stop words such as *a* or *the* would almost certainly always be on top of the list and provide the user with little to no useful information. Instead, the program only counts the appearances of lexical words to delve into the vocabulary usage of the text being processed.

As previously discussed in section 2.2, an analysis of the frequency of content words within texts may prove extremely useful for ELT, among other fields of study, as it is a discipline in which vocabulary learning is of utmost importance. Within an ELT classroom, the tool may be used to keep students motivated with the examples, as all text is retrieved from news articles and is therefore always recent. It would allow students to study the news in detail. A possible exercise would be for students to perform n-gram searches. Users can look for collocations within the text while discarding the lexical bundles which are of no interest. Another possible exercise is using the syntax parser to visualize the syntactic relations within sentences. Students may then draw their own trees and compare them to those provided in the program, for example. The tool has plenty of resources to be explored and may also be of use for teachers, were they to need new examples of sentences with their syntactic trees already drawn, among other possibilities.

6. Conclusion

The use of corpora in the ELT classroom has been proven to be highly beneficial for both teachers and students since it provides them examples of language used in real communication. In addition, corpus-based findings are helpful to build teaching materials or prioritize what to teach. In order to fully exploit all the information hidden in a corpus, we need to use some tools to facilitate the analysis of the vast amounts of data. Through these analyses, users can extract meaningful generalizations about language and hidden relationships established by words.

This paper has presented a tool built to help students and teachers of English with the creation and use of corpora. The tool presents them with linguistic data and several functionalities that will help them create, discover, and analyse texts. The corpus could be a significant part of an ELT lesson when properly used. Although the tool's functionalities work as intended, some minor errors on the front-end part of the project could be addressed in future development. Firstly, the clickable word spans on the scraped piece of news are not adequately divided into actual words, as non-alphabetical data has not been treated yet. This fact prevents the concordancer and the lexical bundles from working accurately when a word stands next to punctuation or numerals. Secondly, the concordancer and the n-gram searchers only take into consideration the text introduced by the user. In a future deployment, searches should be performed in all news articles scraped onto the corpus if required. This functionality would allow the tool to uncover and show more meaningful relations among words.

Lastly, other functionalities that would be interesting to implement in the future are a button for the client to download the scraped piece of news or the incorporation of other newspapers since currently, the only supported one is *The Guardian*. Another

possible improvement would be the evaluation of *Spacy*'s syntactic parser accuracy. The trees it provides could be tested against manually drawn ones to check the correctness of the analyses. Finally, I would like to conclude by mentioning that the program written for this thesis is user-friendly and has been made readily accessible on the Internet at Open Corpus.

References

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2), 141-161. DOI.

Baker, M. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*. In M., Baker, G., Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 233-252). John Benjamins Publishing Company.

Bird, S., Loper, E. & E. Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Cheng, W. (2012). *Exploring Corpus Linguistics: Language in Action*. Routledge.

Ghadessy, M., Henry, A. & R. L. Roseberry (Eds.). (2001). *Small Corpus Studies and ELT: Theory and practice*. John Benjamins Publishing Company.

Goulden, R., Nation, P. & J. Read (1990) *How Large Can a Receptive Vocabulary Be?* Applied Linguistics, 11, 341-363. DOI.

Honnibal, M., & I. Montani (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Pearson Education Limited.

McEnery, T., & A. Hardie. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Pollach, I. (2012). Taming Textual Data: The contribution of Corpus Linguistics to Computer-aided Text Analysis. *Organizational Research Methods*, 15(2), 263-287. DOI.

Sinclair, J. M. & A. Renouf (1988) *A lexical syllabus for language learning*.

Appendix

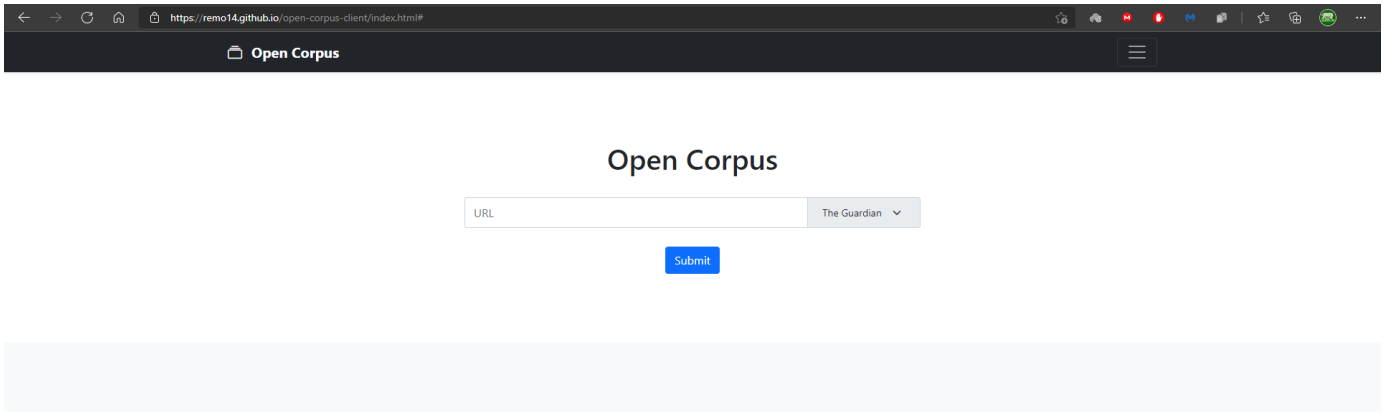
Annex 1. Front End Documentation

1. The programming languages which have been used are the following:

- JavaScript (JavaScript.com)
  JavaScript is a language which is used to make interactive webpages, modifying its structure and keeping it user-friendly.
- CSS (CSS Snapshot 2020 (w3.org))
  CSS is a language which is used to establish the style and visual structure of elements within the webpage.
- HTML (HTML)
  HTML is a language which structures the web page and the contents within it.

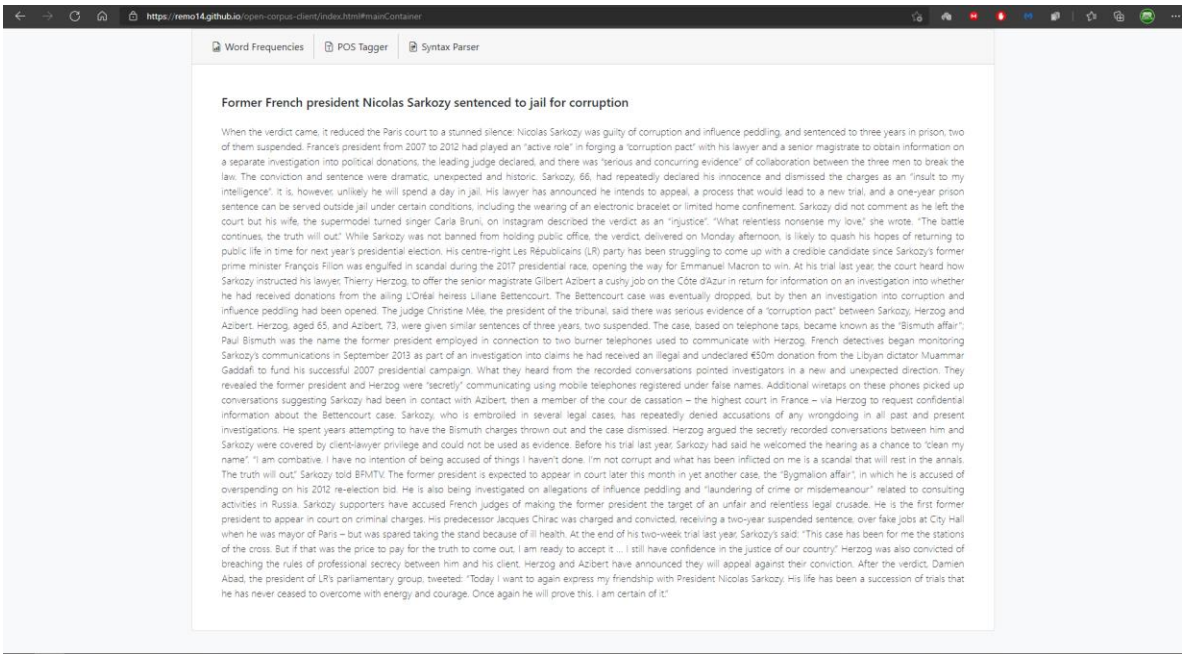
2. The following section includes screenshots of the various views within the web page, including a brief description of their functionality.

a) Landing page of the website on remote. The URL is the following:
Open Corpus.



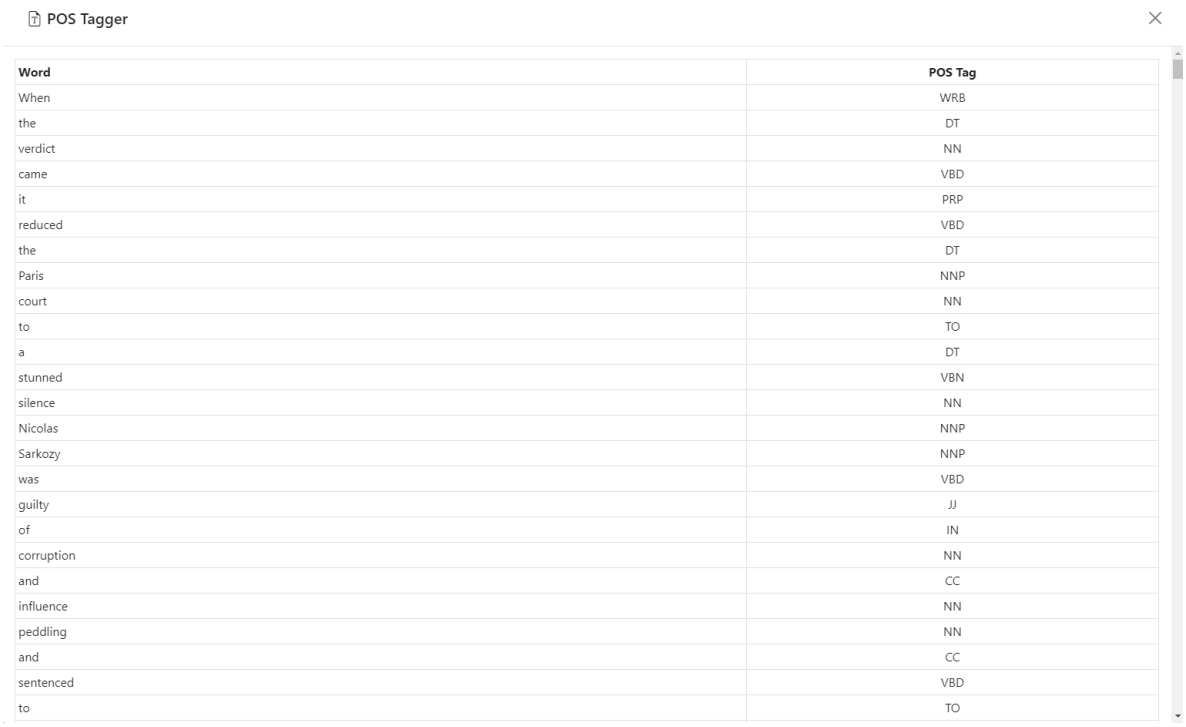
On the main page, the user may paste an URL from a news article (currently, the supported websites are: TheGuardian).

b) Once users submit1 the URL, it returns the title and body of the piece of news.



c) From there, the user may interact with either the words within the body itself or with 3 buttons which appear on top. The first button shows the frequency of words, ordered alphabetically:

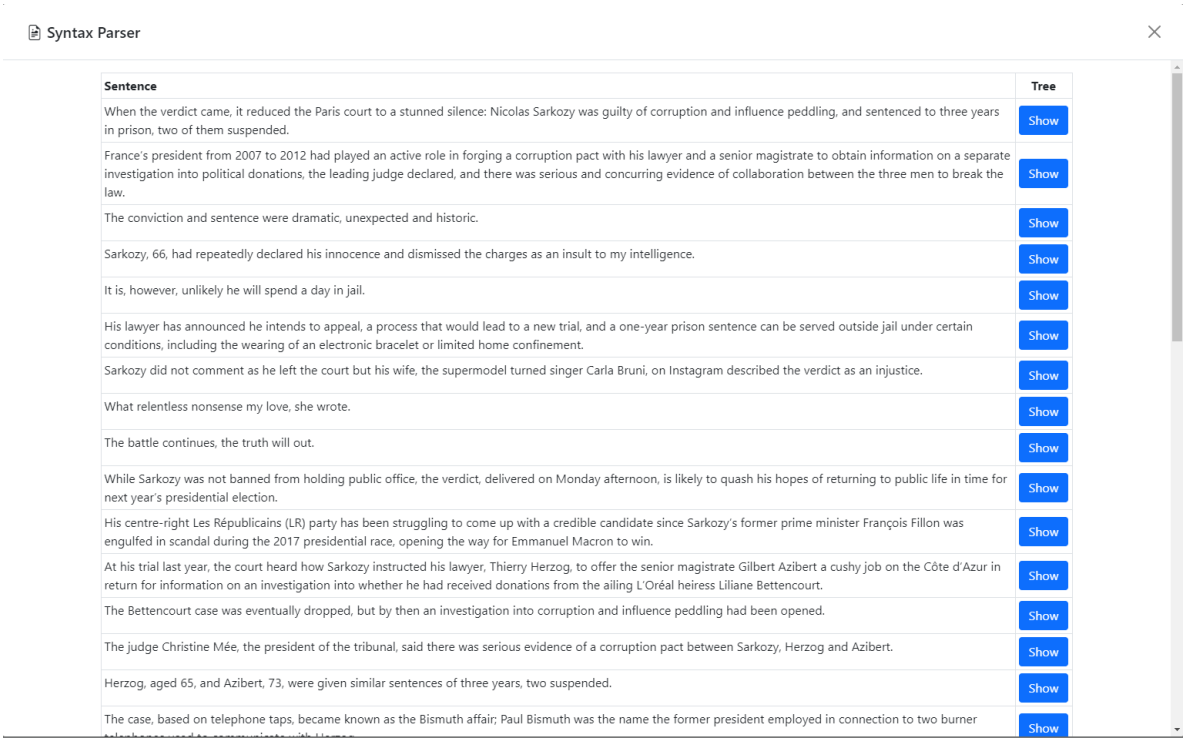
d) The second button displays a list of all words with their respective part of speech tag:



POS Tagger

Word	POS Tag
When	WRB
the	DT
verdict	NN
came	VBD
it	PRP
reduced	VBD
the	DT
Paris	NNP
court	NN
to	TO
a	DT
stunned	VBN
silence	NN
Nicolas	NNP
Sarkozy	NNP
was	VBD
guilty	JJ
of	IN
corruption	NN
and	CC
influence	NN
peddling	NN
and	CC
sentenced	VBD
to	TO

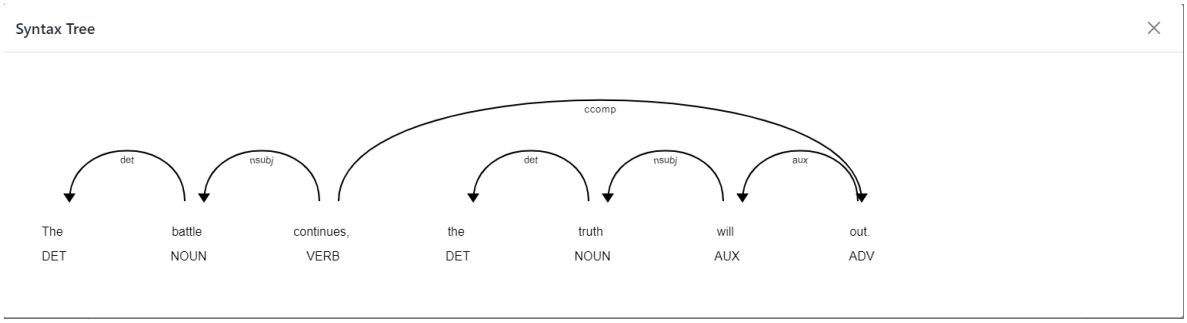
e) The third button displays a list of all sentences, each with a button of its own:



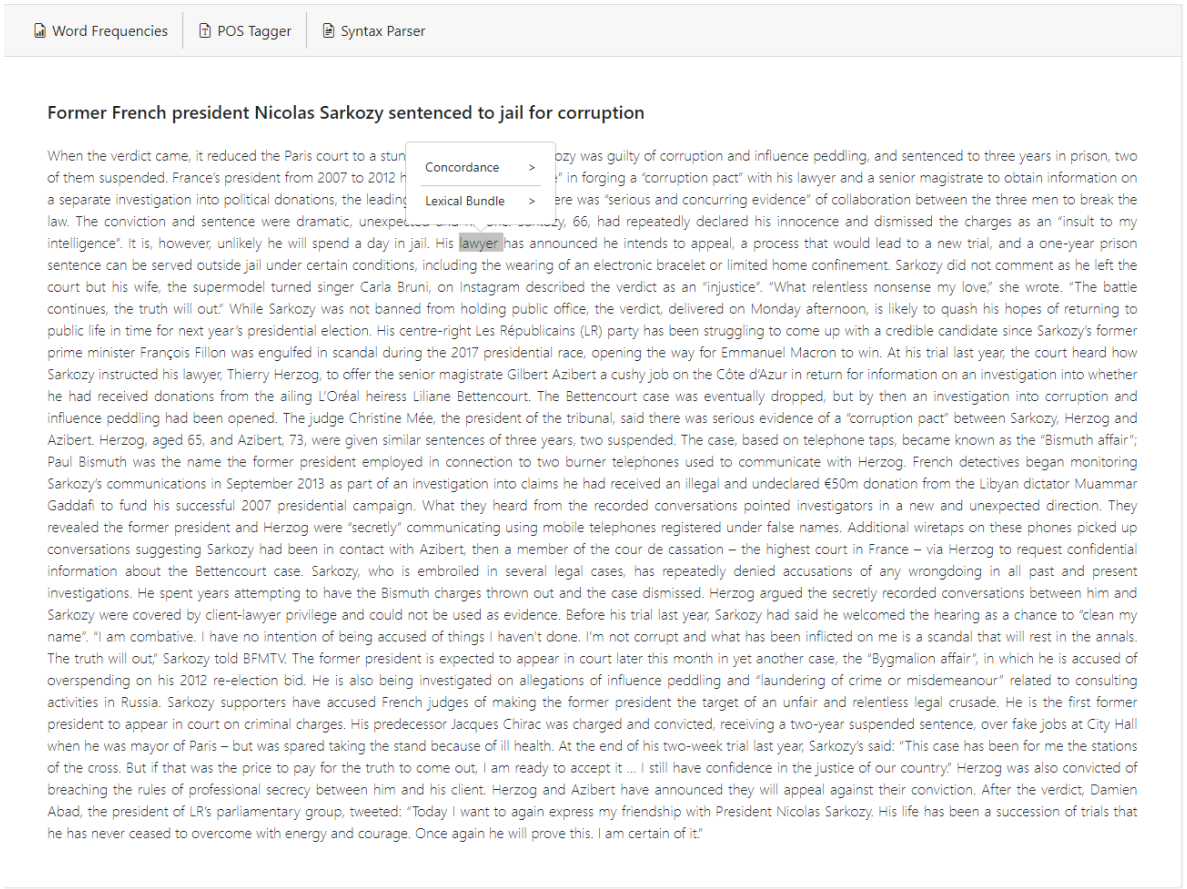
Syntax Parser

Sentence	Tree
When the verdict came, it reduced the Paris court to a stunned silence: Nicolas Sarkozy was guilty of corruption and influence peddling, and sentenced to three years in prison, two of them suspended.	Show
France's president from 2007 to 2012 had played an active role in forging a corruption pact with his lawyer and a senior magistrate to obtain information on a separate investigation into political donations, the leading judge declared, and there was serious and concurring evidence of collaboration between the three men to break the law.	Show
The conviction and sentence were dramatic, unexpected and historic.	Show
Sarkozy, 66, had repeatedly declared his innocence and dismissed the charges as an insult to my intelligence.	Show
It is, however, unlikely he will spend a day in jail.	Show
His lawyer has announced he intends to appeal, a process that would lead to a new trial, and a one-year prison sentence can be served outside jail under certain conditions, including the wearing of an electronic bracelet or limited home confinement.	Show
Sarkozy did not comment as he left the court but his wife, the supermodel turned singer Carla Bruni, on Instagram described the verdict as an injustice.	Show
What relentless nonsense my love, she wrote.	Show
The battle continues, the truth will out.	Show
While Sarkozy was not banned from holding public office, the verdict, delivered on Monday afternoon, is likely to quash his hopes of returning to public life in time for next year's presidential election.	Show
His centre-right Les Républicains (LR) party has been struggling to come up with a credible candidate since Sarkozy's former prime minister François Fillon was engulfed in scandal during the 2017 presidential race, opening the way for Emmanuel Macron to win.	Show
At his trial last year, the court heard how Sarkozy instructed his lawyer, Thierry Herzog, to offer the senior magistrate Gilbert Azibert a cushy job on the Côte d'Azur in return for information on an investigation into whether he had received donations from the ailing L'Oréal heiress Liliane Bettencourt.	Show
The Bettencourt case was eventually dropped, but by then an investigation into corruption and influence peddling had been opened.	Show
The judge Christine Mée, the president of the tribunal, said there was serious evidence of a corruption pact between Sarkozy, Herzog and Azibert.	Show
Herzog, aged 65, and Azibert, 73, were given similar sentences of three years, two suspended.	Show
The case, based on telephone taps, became known as the Bismuth affair; Paul Bismuth was the name the former president employed in connection to two burner	Show

f) When clicking on “Show” on a sentence, it draws a syntactic tree and displays it:



g) Back in the body of the text, the user may click on any word, which will give him two options of searches for that word:



h) The option “Concordance” displays all contexts in which said word appears:

Concordance > lawyer

Previous context	Word	Following context
them suspended France s president from to had played an active role in forging a corruption pact with his	lawyer	and a senior magistrate to obtain information on a separate investigation into political donations the leading judge declared
charges as an insult to my intelligence It is however unlikely he will spend a day in jail His	lawyer	has announced he intends to appeal a process that would lead to a new trial and a oneyear
the way for Emmanuel Macron to win At his trial last year the court heard how Sarkozy instructed his	lawyer	Thierry Herzog to offer the senior magistrate Gilbert Azibert a cushy job on the Côte d Azur in

i) The option “Lexical bundle” displays all bigrams of the word:

Bigram > lawyer 2 3 4

Lexical Bundle	Frequency
his, lawyer	2
lawyer, and	1
His, lawyer	1
lawyer, has	1
lawyer, Thierry	1

j) Lastly, from there, users may choose to display trigram or tetragrams instead by clicking either on numbers 3 or 4, respectively:

Trigram > lawyer 2 3 4

Lexical Bundle	Frequency
with, his, lawyer	1
his, lawyer, and	1
lawyer, and, a	1
jail, His, lawyer	1
His, lawyer, has	1
lawyer, has, announced	1
instructed, his, lawyer	1
his, lawyer, Thierry	1
lawyer, Thierry, Herzog	1

Tetragram > lawyer 2 3 4

Lexical Bundle	Frequency
pact, with, his, lawyer	1
with, his, lawyer, and	1
his, lawyer, and, a	1
lawyer, and, a, senior	1
in, jail, His, lawyer	1
jail, His, lawyer, has	1
His, lawyer, has, announced	1
lawyer, has, announced, he	1
Sarkozy, instructed, his, lawyer	1
instructed, his, lawyer, Thierry	1
his, lawyer, Thierry, Herzog	1
lawyer, Thierry, Herzog, to	1

3. Github (GitHub) has been used as a platform to save the code in repositories and to establish the webpage within their webpage service Github Pages (GitHub Pages).

4. The libraries which have been used are JQuery (jQuery) for better handling of the html document, Bootstrap (Bootstrap) to keep the webpage responsive and adding features such as popovers.

5. In order to deploy the code into the cloud server, it is synchronized with GitHub, which allows for an easy way of updating any changes made.

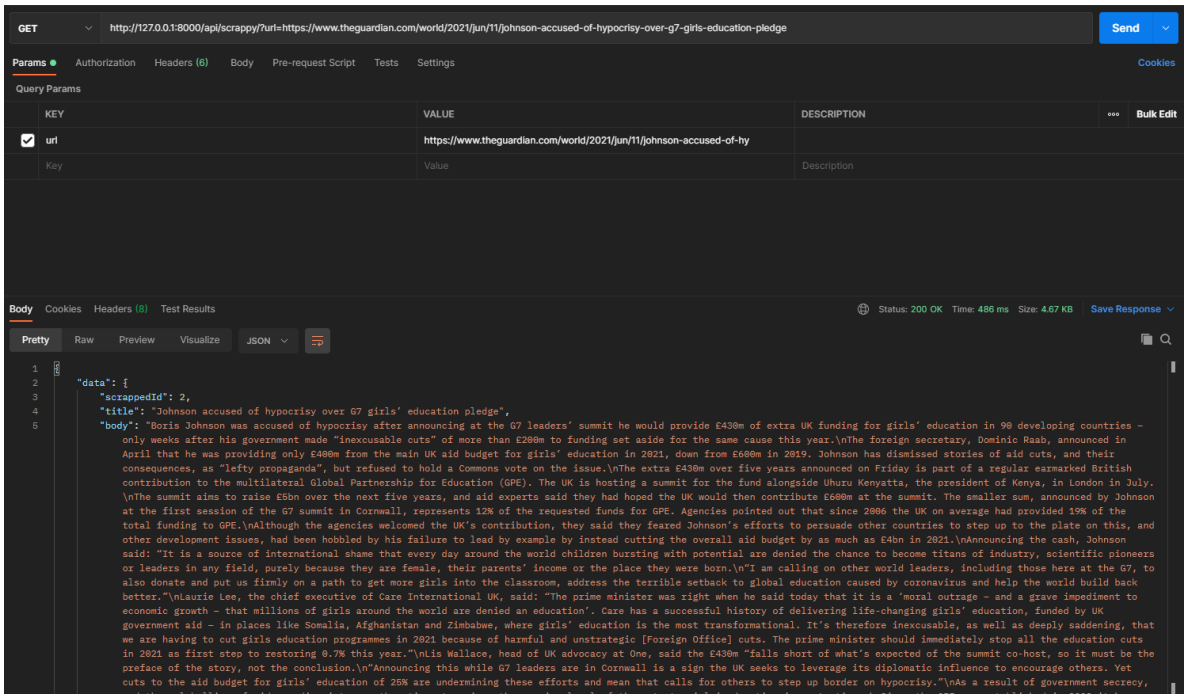
Annex 2. Back End Documentation

1. The programming languages which have been used are the following:

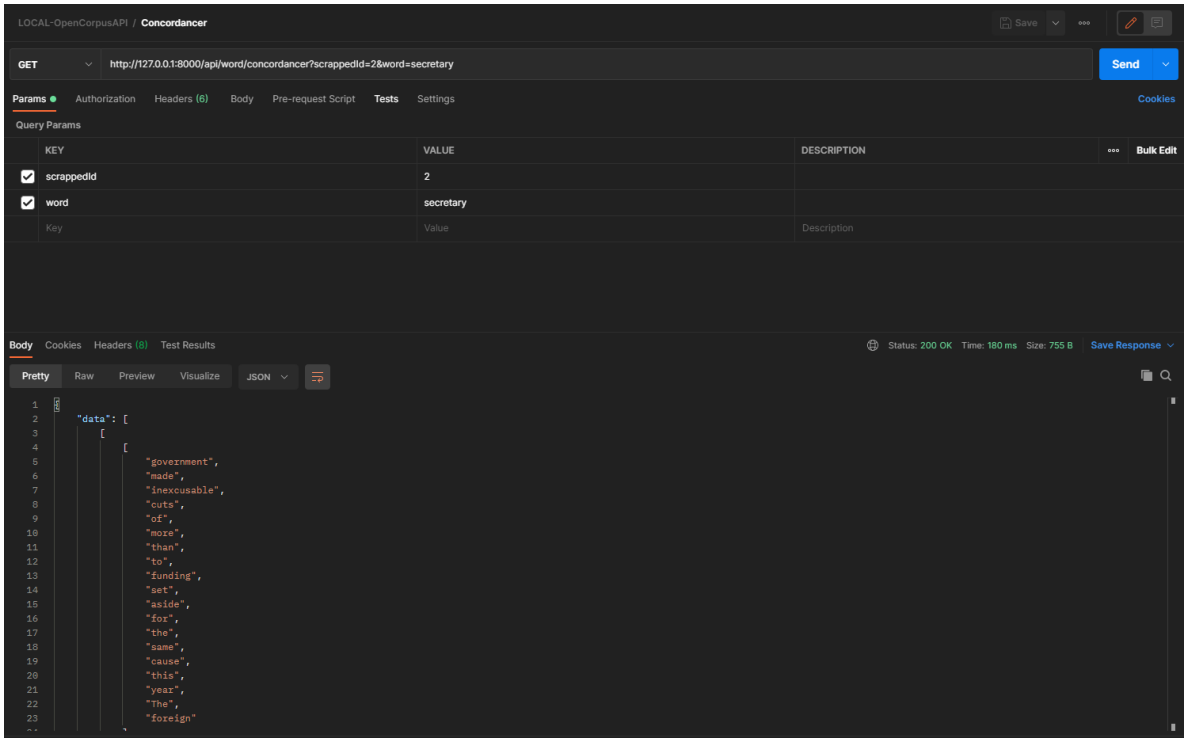
- Python (Welcome to Python.org)
  Python is a language which is often used for developing software, web pages or performing data analysis.

2. The following section includes screenshots of the various endpoints within the web page, both in local access and remote, by using Postman (Postman).

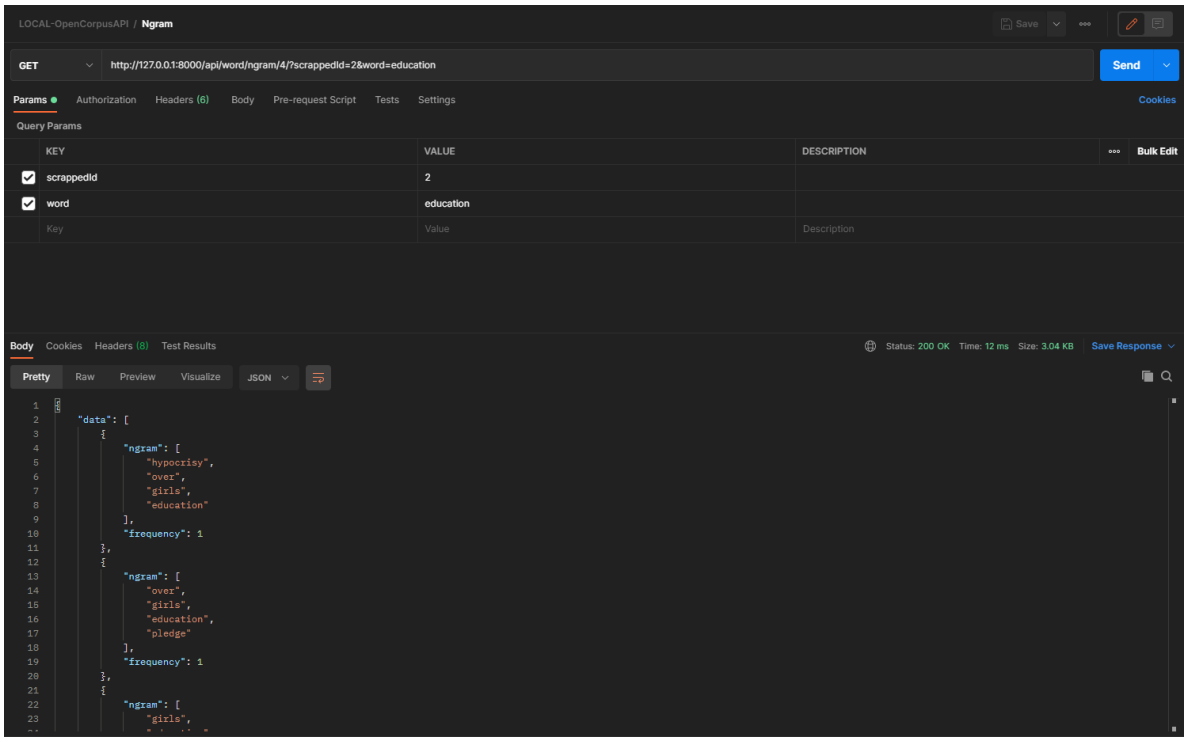
3. Local endpoints. I used them during the development phase of the project.
   a) Scrapped



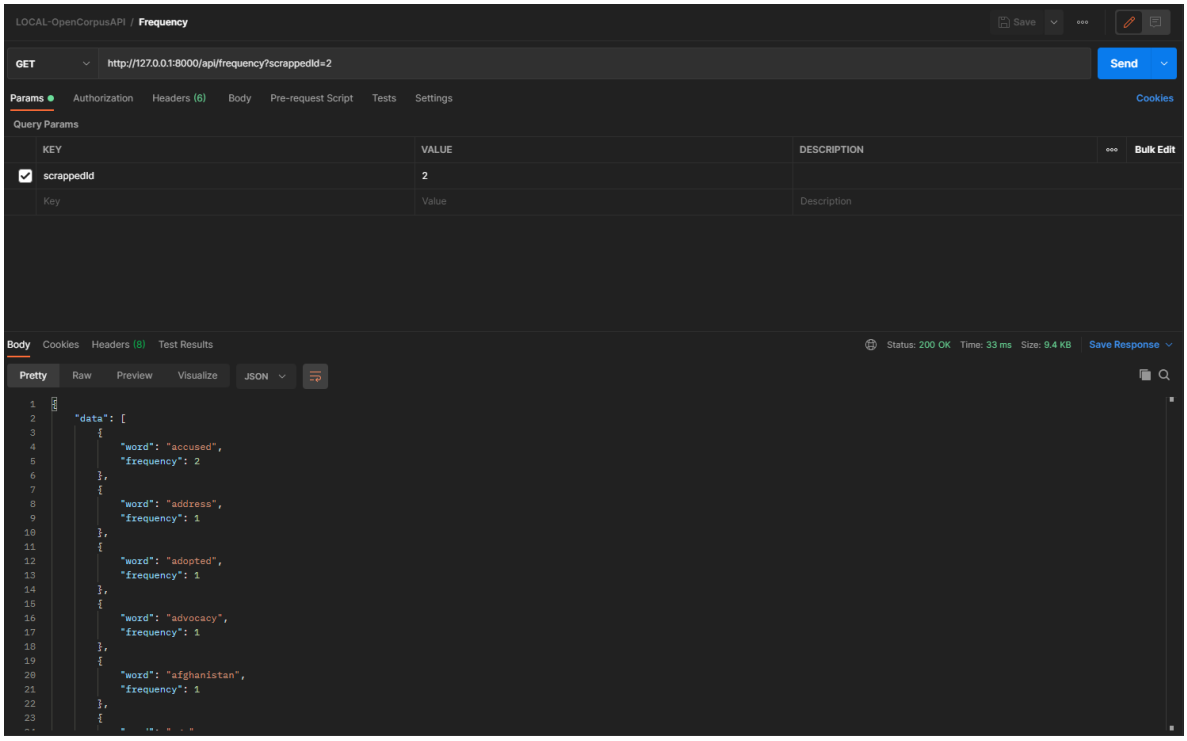
b) Concordancer



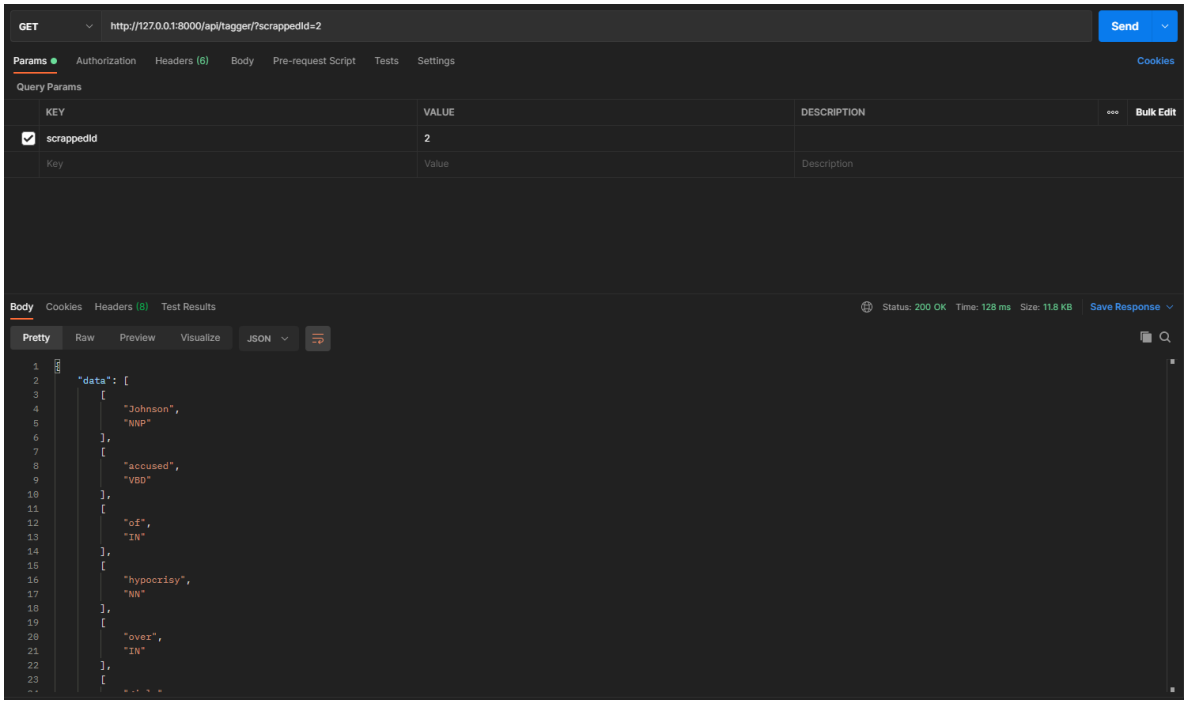
c) Ngram



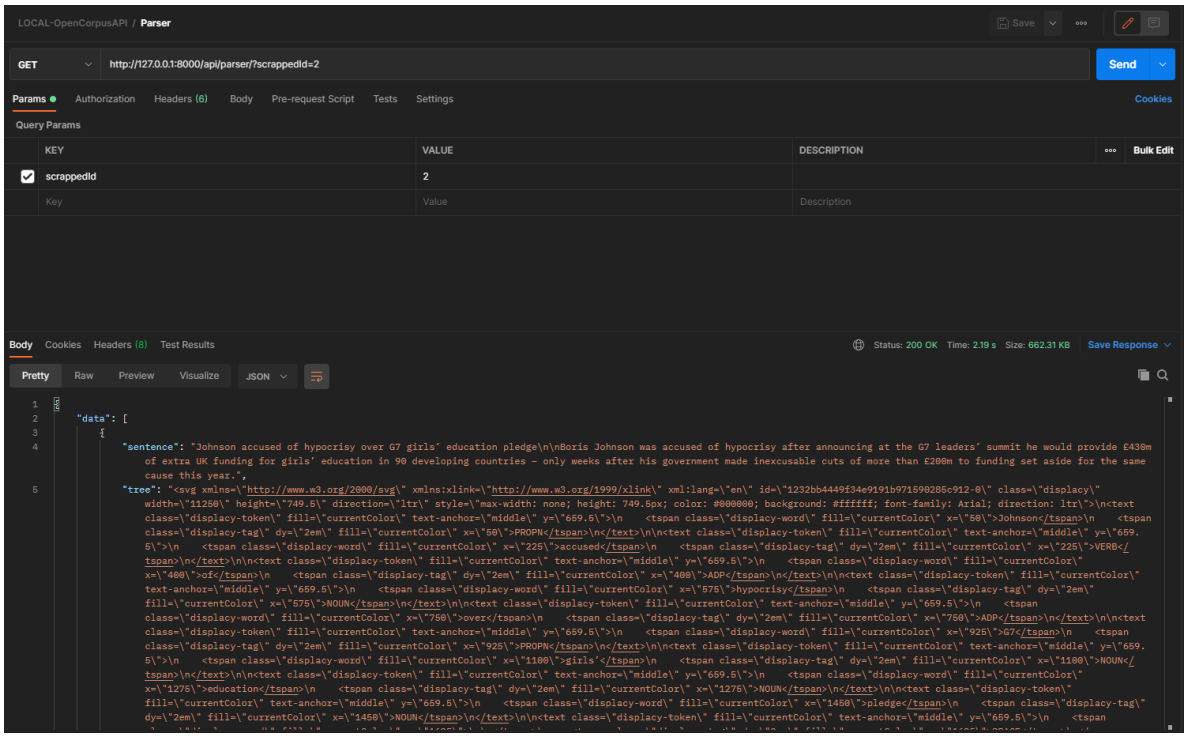
d) Frequency



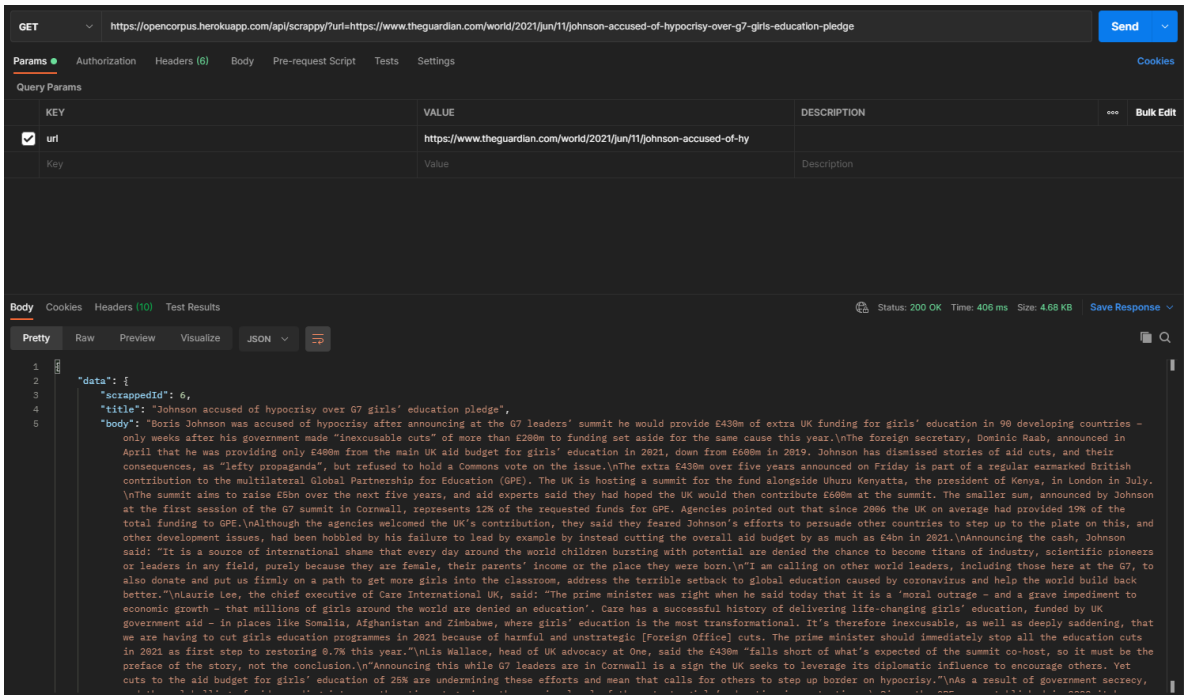
e) Tagger



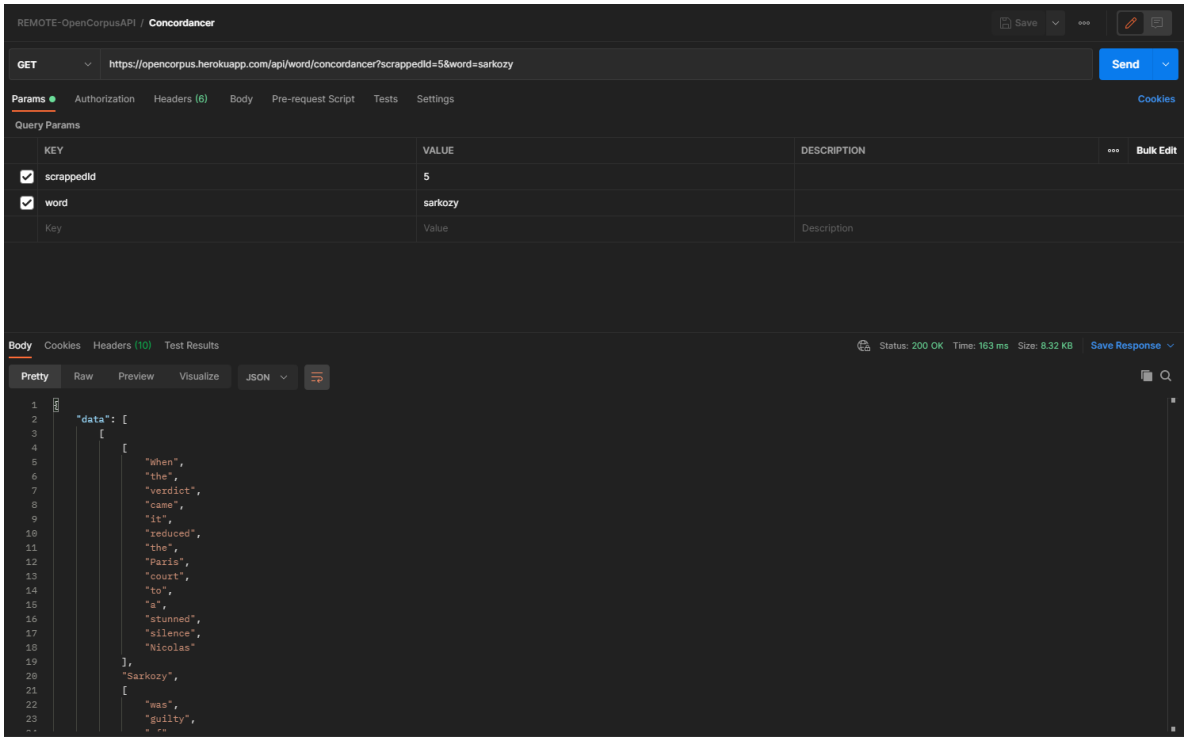
f) Parser



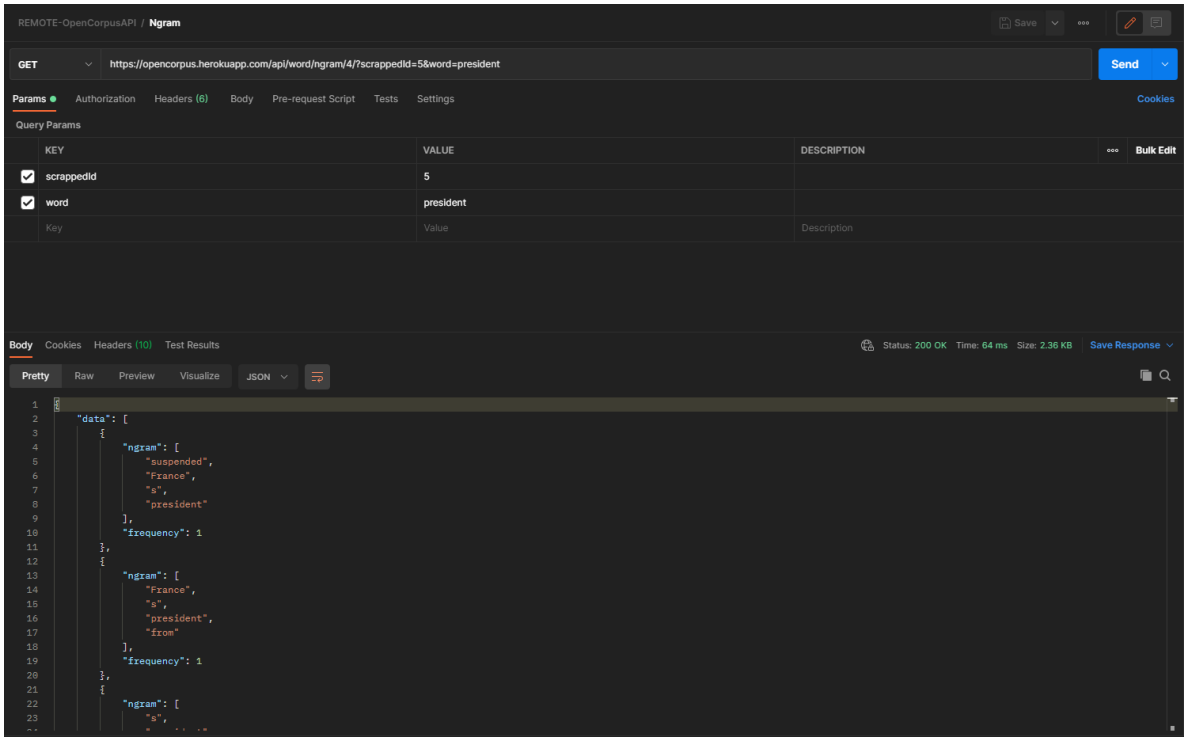
4. Remote endpoints. These are the endpoints used once the app has been deployed on remote.
   a) Scrapped



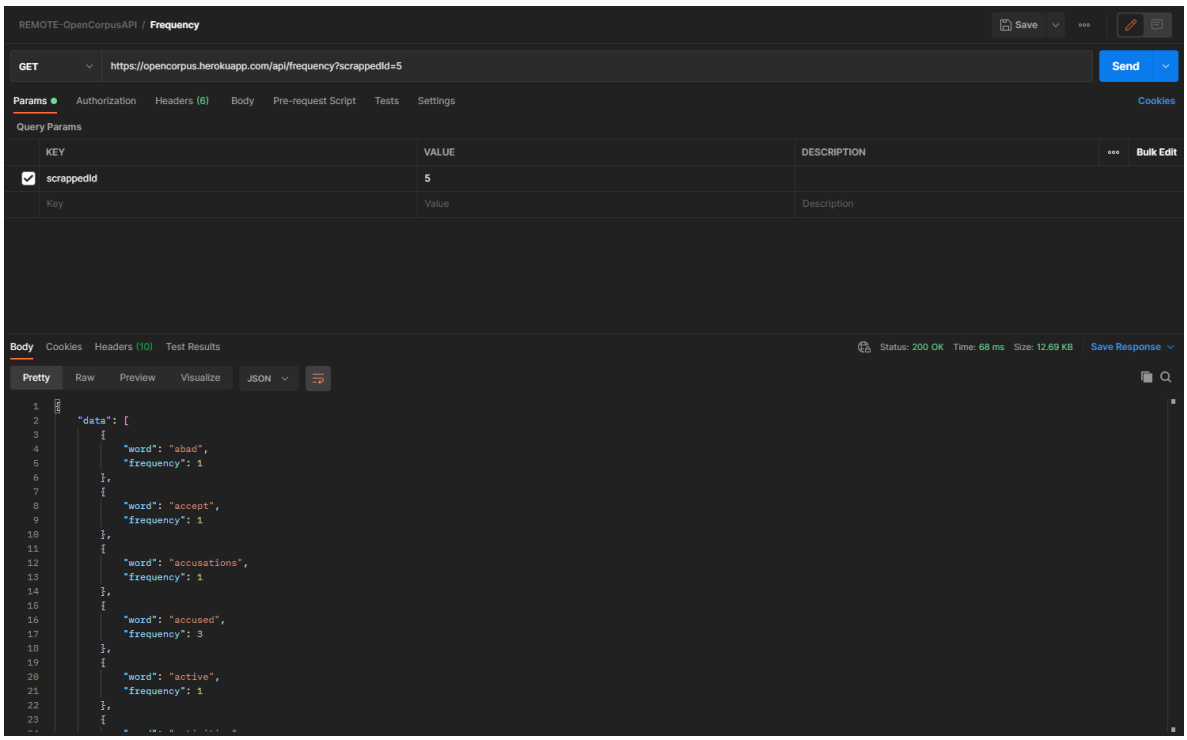
   b) Concordancer



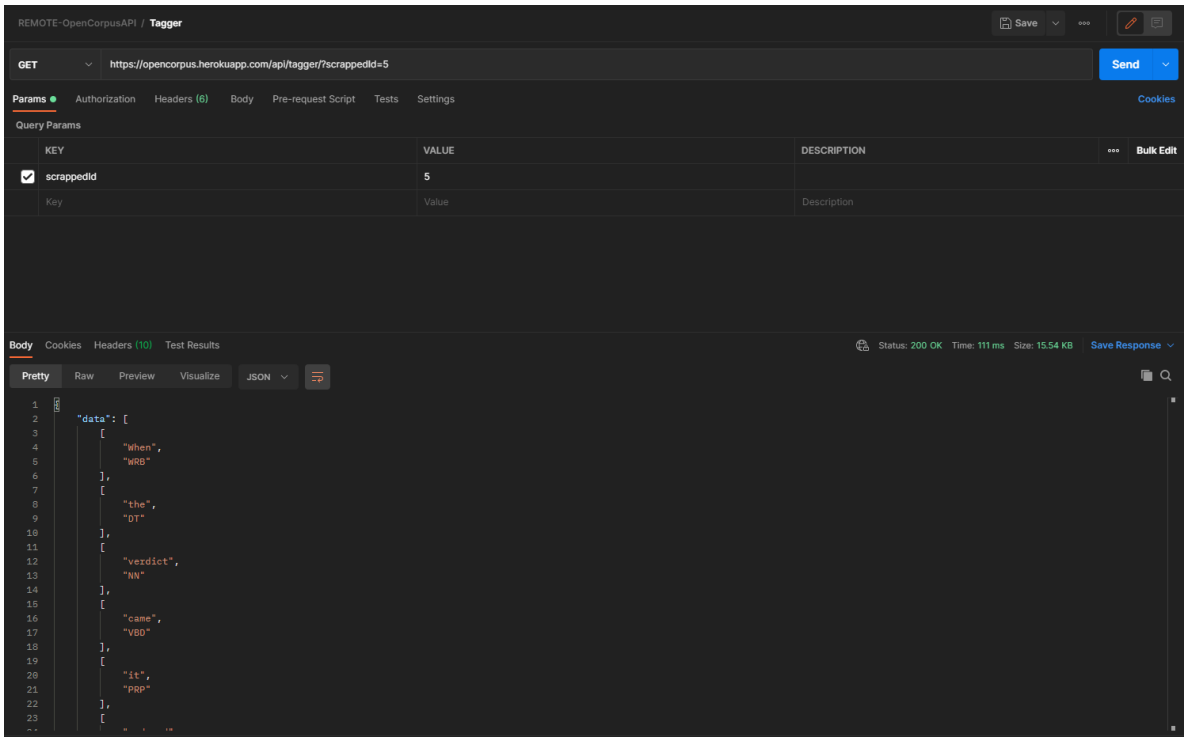
c) Ngram



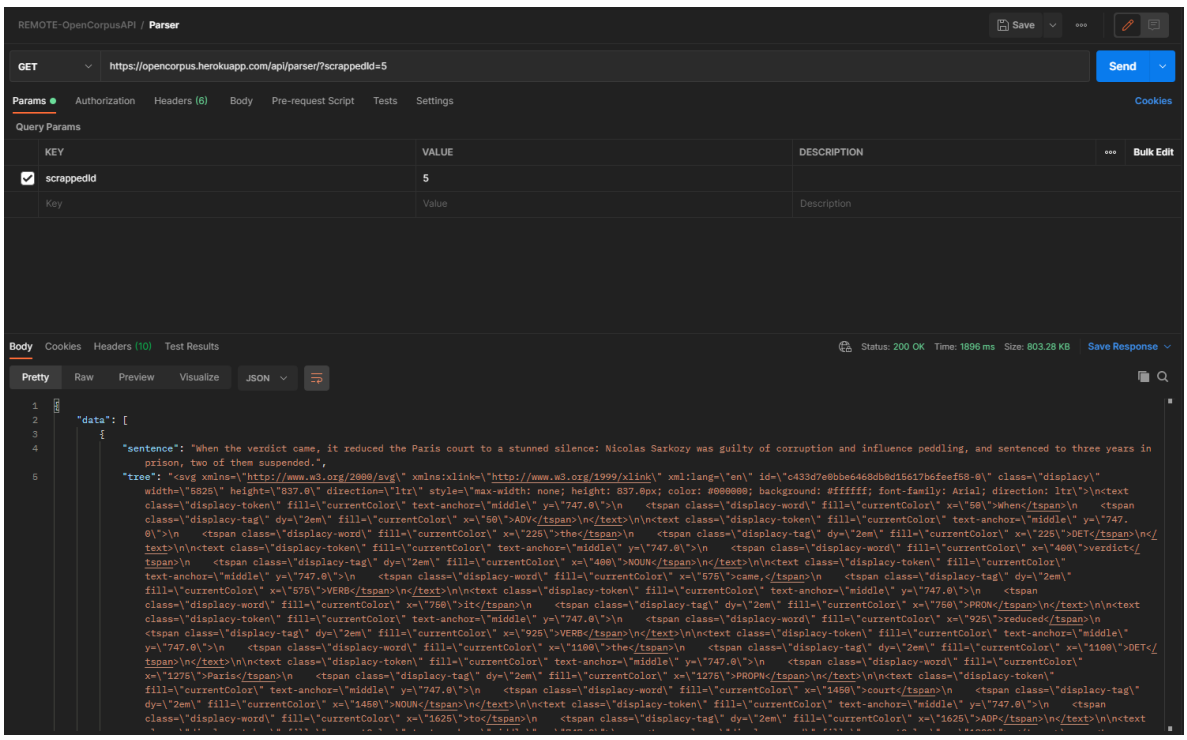
d) Frequency



e) Tagger



f) Parser



5. Django (Django) has been used as a framework. It provides an API structure to be used in the creation of a webpage. An API is an interface which allows for interaction between software.

6. Heroku (Heroku) is a platform which allows to implement code into the cloud, and it supports Python. Heroku is a free-to-use server in which to deploy the framework on remote.

7. In order to deploy the code into the cloud server, it is synchronized with GitHub, which allows for an easy way of updating any changes made.