Voice-text integrated system for interpreters

User Guide

Contenido

CORPUS

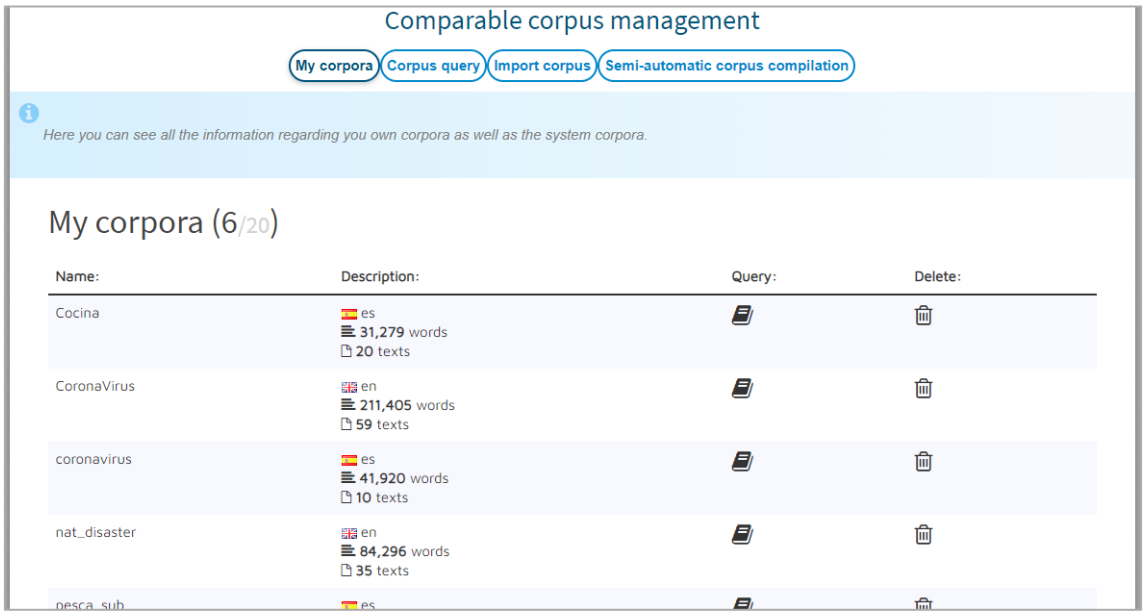
This module is designed to carry out the preparatory tasks for an interpretation, such as the creation and query of corpora, glossaries, etc. The different parts of this module are described below.

1. Comparable corpus management

This module offers different functionalities related to the corpora. You can create your own corpus, either using text files or from an Internet search. Different types of queries can also be performed on the created corpus.

1.1. My corpora

This section shows the basic information for the corpora that have been created: name, language, number of words and number of texts that comprise it. It also allows you to delete a corpus by clicking on the icon, or go to the corpus query by clicking on the icon.



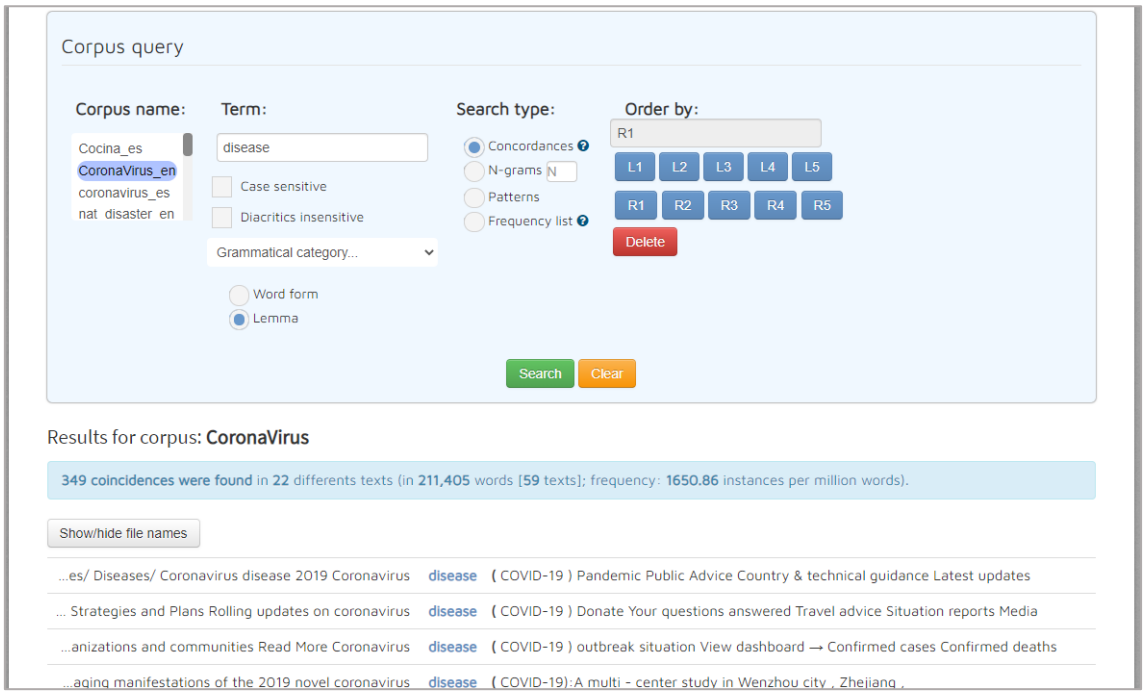
1.2. Corpus query

In this section, you can perform different queries over the corpora that you have created or imported. You just need to select one or several corpora from the list, type the search term and choose the type of query. To select several corpora, first select one and then press Ctrl+click on the next corpus that you would like to include in the selection.

In addition, several options can be set for the term, such as “Case sensitive” (makes the search case sensitive), “Diacritics insensitive” (ignore diacritics), “Grammatical category” (only terms matching the selected grammatical category will be shown). You can also choose if you want to perform the search by word form (as it appears in the corpus) or by lemma (lemmatized form of terms). The different types of queries that can be performed are described below.

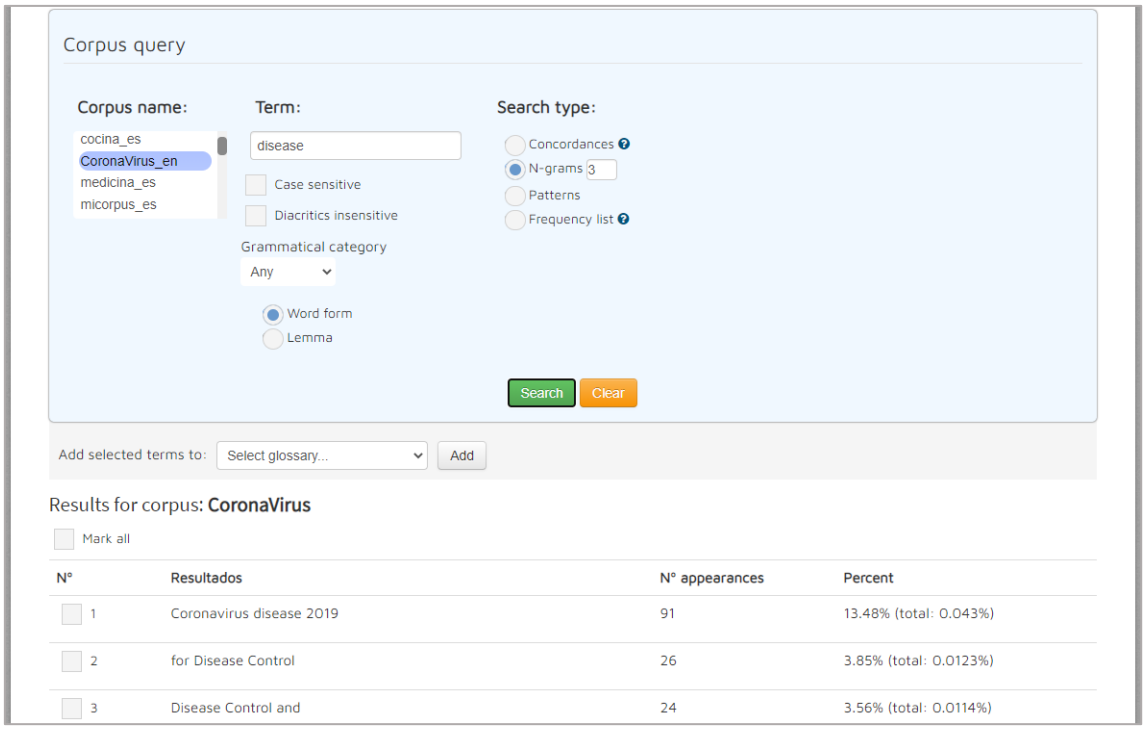
- Concordances

Concordances are shown in a small context containing the given term in the selected corpora. The results can be sorted alphabetically according to the first word from the left (L1), the first word from the right (R1), the second word from the left (L2), the second word from the right (R2), and so on. In addition, several ordering levels can be set.



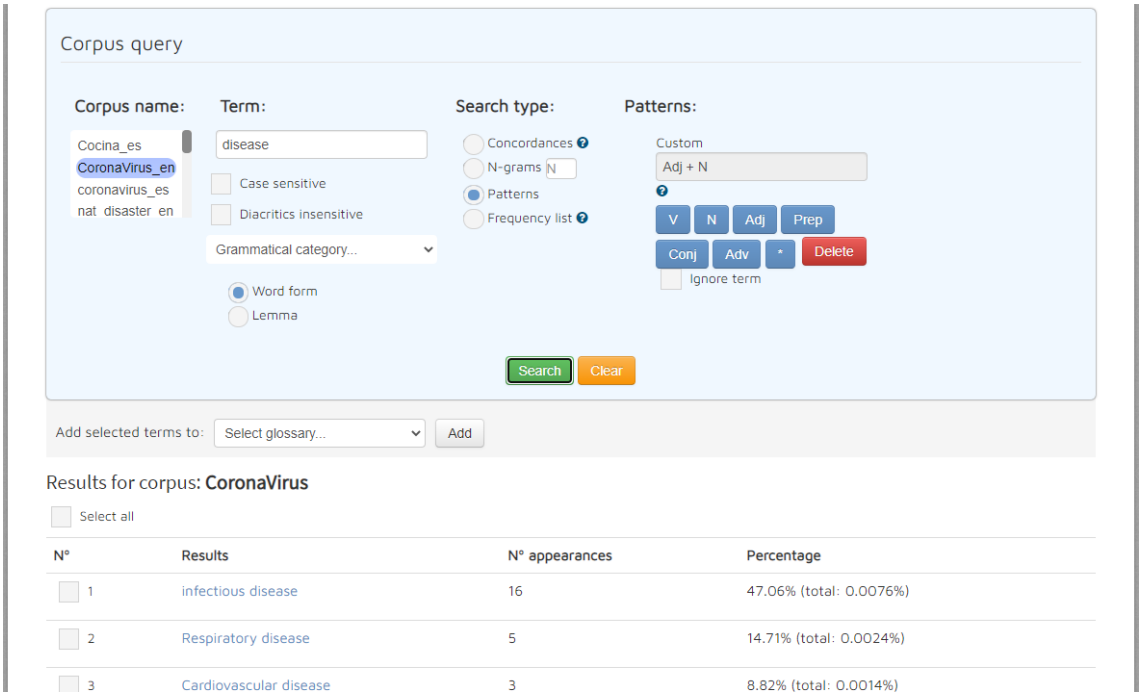
- N-grams

In this case you must assign a value to N (2 for bigrams, 3 for trigrams, etc.). The N-grams containing the given term in the selected corpora are shown for the chosen N value.



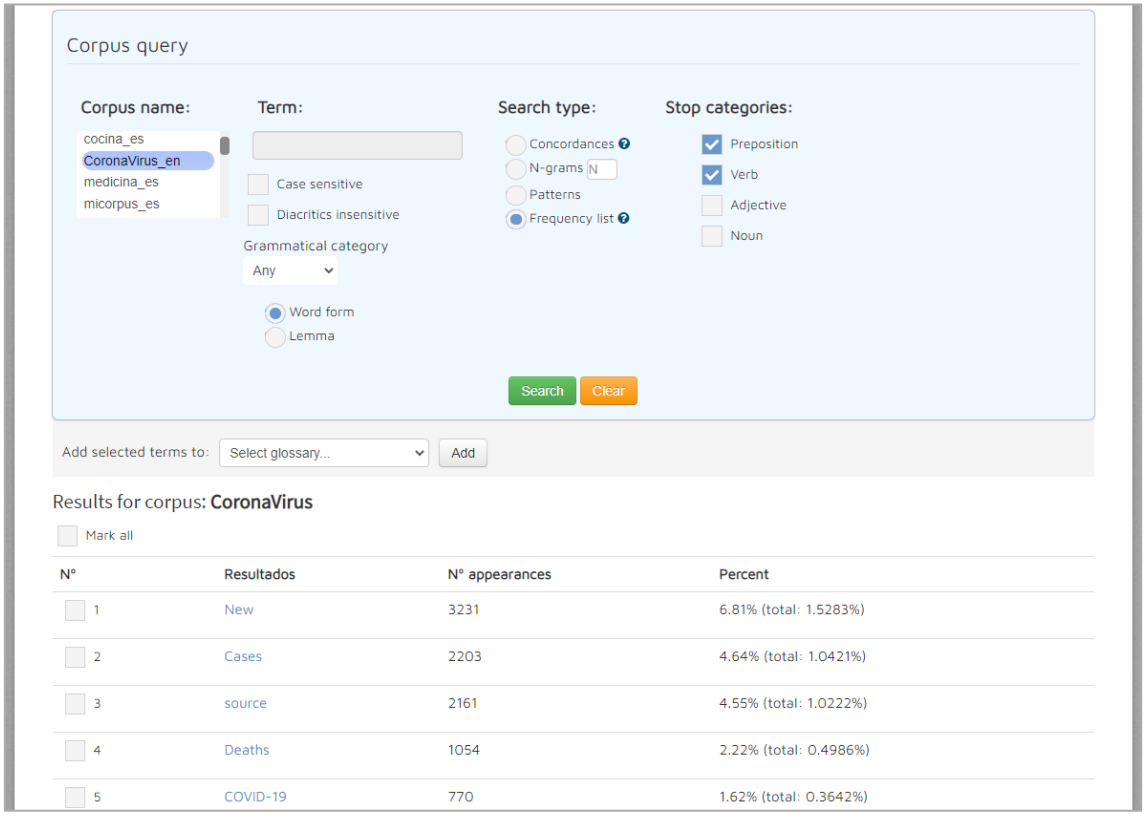
- Patterns

The terms or collocations that match the chosen pattern and contain the given term in the selected corpus are displayed. Custom patterns can also be created. If the “Ignore term” option is checked, the term will not be considered.



- Frequency list

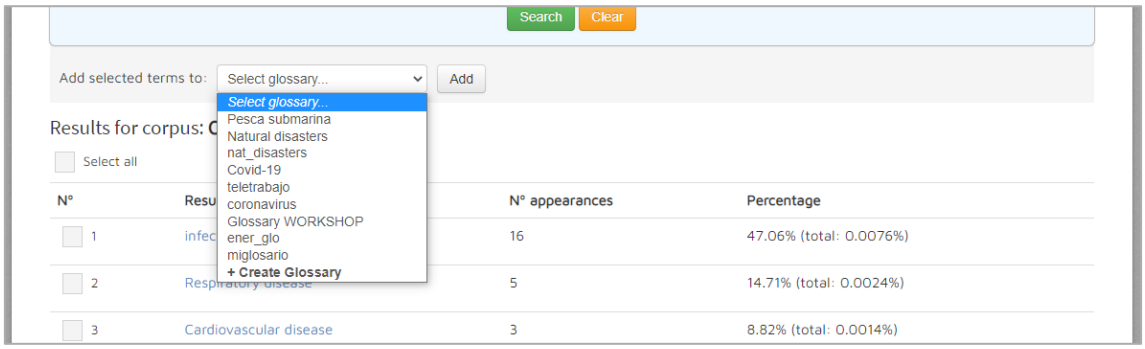
This functionality offers a list of the most frequent words that could be candidates for terms or phraseological units in the selected corpora. You do not need to enter any term, only the grammatical category you do not want to be shown ("Noun", "Adjective", "Verb", “Preposition”, “Conjunction”, “Adverb”.)



After performing the query, you can select (by checking the checkbox) the resulting terms (except for *Concordances*) and add them to a glossary that you can later consult and edit (see section *Glossary query*). To do this, select a glossary from the drop-down menu or choose a name to create a new one and press the "Add" button. In order to quickly select several terms, you have two options:

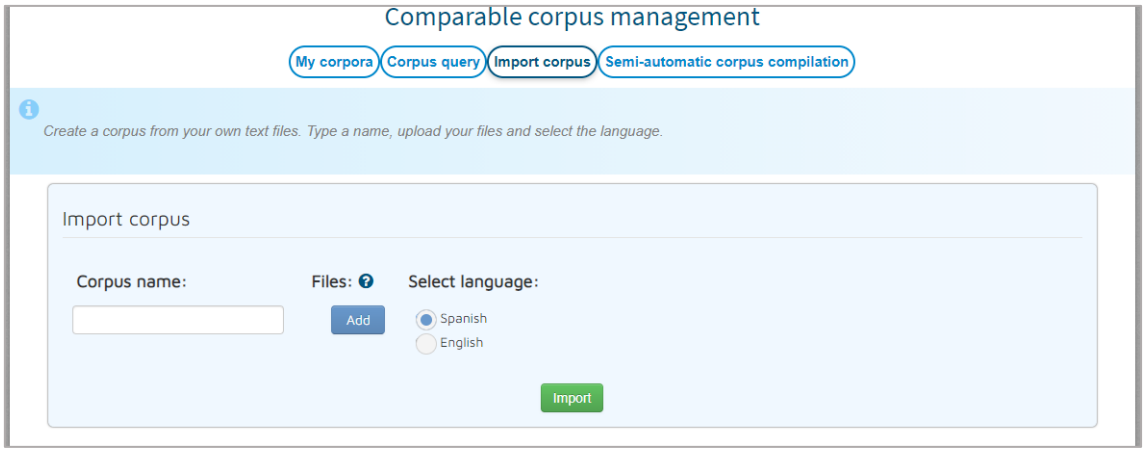
1) Check “Select all” to select all terms.

2) Select a range. To do this, check an initial checkbox and, while pressing the SHIFT button, check an ending checkbox. All checkboxes in between will be automatically checked.



1.3. Importing a corpus

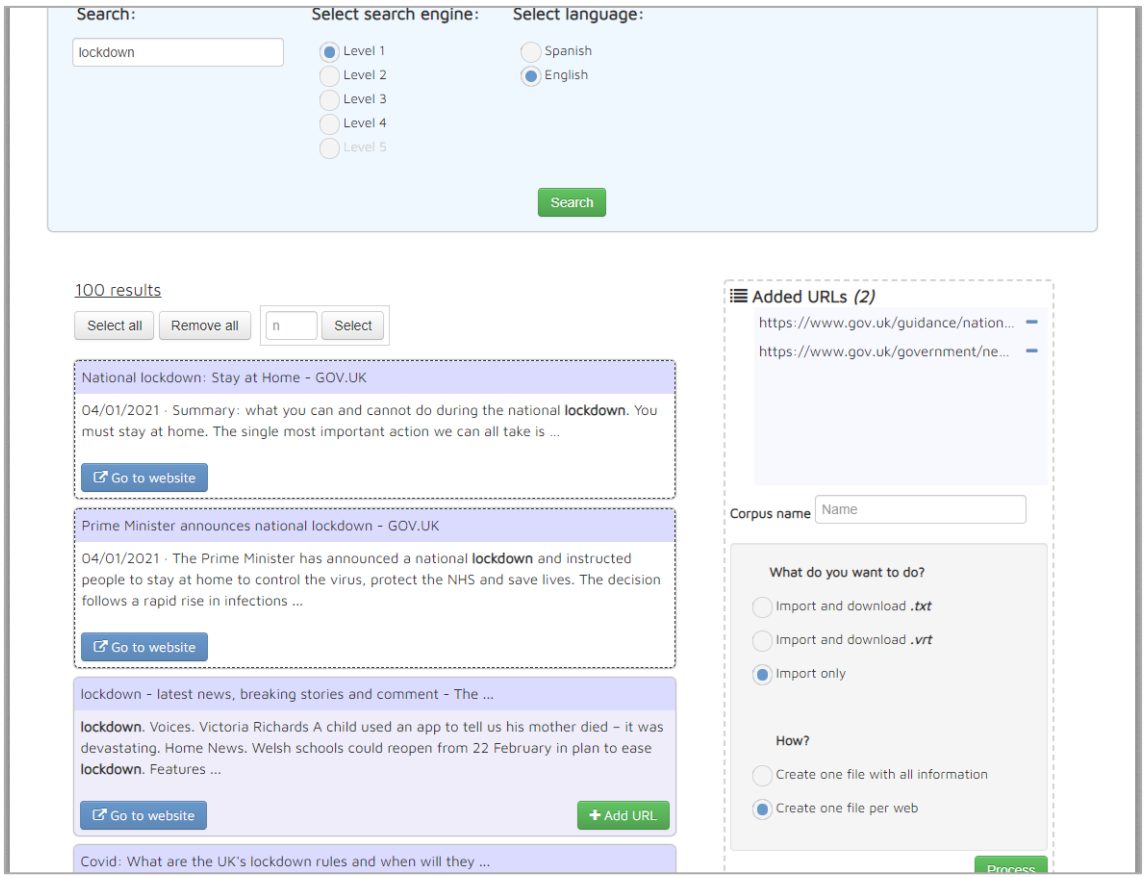
In this section you can create your own corpus from a set of texts. To do so, choose a name for the corpus and select the text files from your computer. These files can be in TXT or VRT[1] (VeRticalized Text) format. Select the language of the texts. After pressing the "Import" button, the corpus will be created, and you will be able to consult it in the section *Corpus query*.



[1] The columns order must be: token, part of speech, lemma. See *Appendix: Verticalized text*.

1.4. Semi-automated corpus compilation

In this section you will be able to create a corpus using texts from the Internet. You just need to type a search term or phrase (you can use the usual search techniques such as putting the phrase in quotation marks to search for an exact match or typing the "-" Boolean operator before a word to exclude it from the search and select a language. After pressing the "Search" button, a list of related websites will be displayed, from which you will be able to choose as many as you consider appropriate to add to your corpus. Then you only have to choose a name for the corpus and select one of the three following options: "Import and download .txt" (the corpus will be downloaded in plain text format), "Import and download .vrt" (the corpus will be downloaded in vertical text format with morphological labelling, see *Appendix: Verticalized text*) or "Import only". In any case, a corpus which can be consulted in the section *Corpus query* will be created. You can also choose whether to combine all the obtained texts into a single file (by selecting "Create a file with all the information") or create a text file for each chosen web resource in addition to a metadata file (by selecting "Create a file for each web").

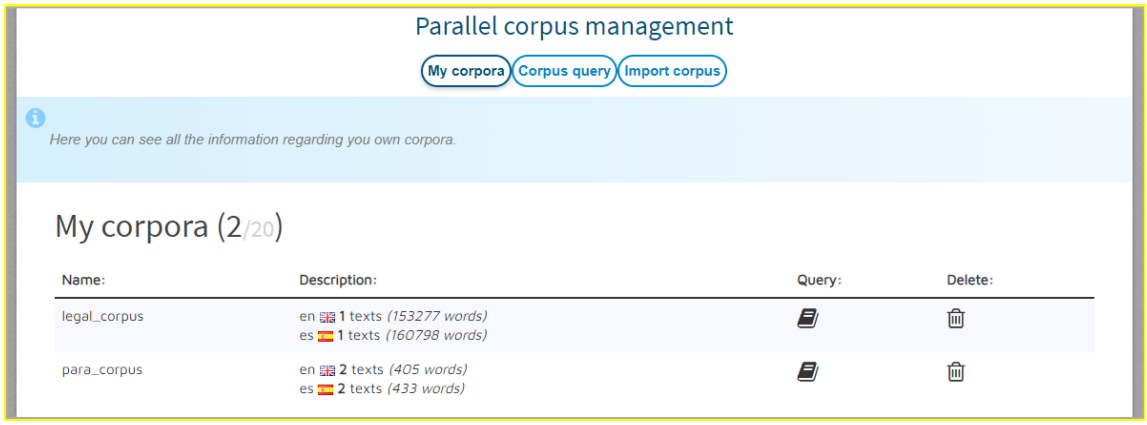


2. Parallel corpus management

In this module you can create your own parallel corpus using text files. Different types of queries can also be performed on the created corpus.

2.1 My corpora

This section shows the basic information for the parallel corpora that have been created: name, language, number of words and number of texts that comprise it. It also allows you to delete a corpus by clicking on the icon, or go to the corpus query by clicking on the icon.



2.2 Corpus query

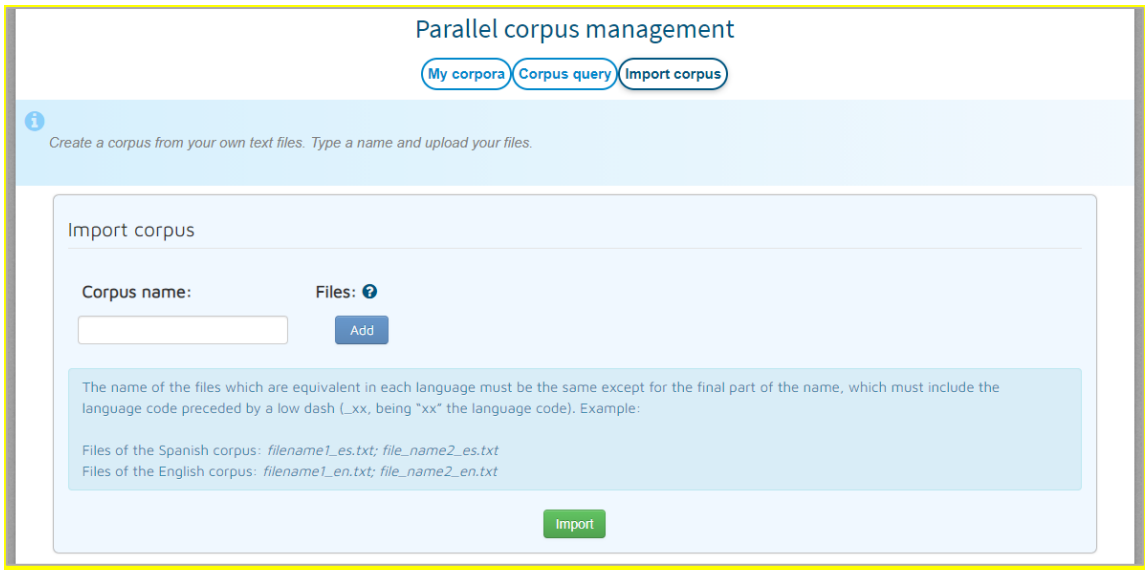
In this section, you can perform different queries over the parallel corpora that you have created or imported. You just need to select one or several corpora from the list and type the search term. To select several corpora, first select one and then press Ctrl+click on the next corpus that you would like to include in the selection.

In addition, several options can be set for the term, such as “Case sensitive” (makes the search case sensitive), “Diacritics insensitive” (ignore diacritics), “Grammatical category” (only terms matching the selected grammatical category will be shown). You can also choose if you want to perform the search by word form (as it appears in the corpus) or by lemma (lemmatized form of terms).

After clicking “Search” button, fragments matching the typed text in selected corpus will be shown next to the equivalent in the other language corpus fragments.

2.3 Corpus importation

In this section, you can create your own parallel corpus from a set of text files. The name of these files must be in a specific format: the name of the files which are equivalent in each language must be the same except for the final part of the name, which must include the language code preceded by a low dash (_xx, being “xx” the language code. For example, the Spanish file filename1_es.txt corresponds to the English file filename1_en.txt, the Spanish file file_name2_es.txt corresponds to the English file file_name2_en.txt, etc.). Then, choose a name for the corpus and select the text files from your computer. These files can be in TXT or VRT (VeRticalized Text) format. After pressing the "Import" button, the corpus will be created, and you will be able to consult it in the section Corpus query.



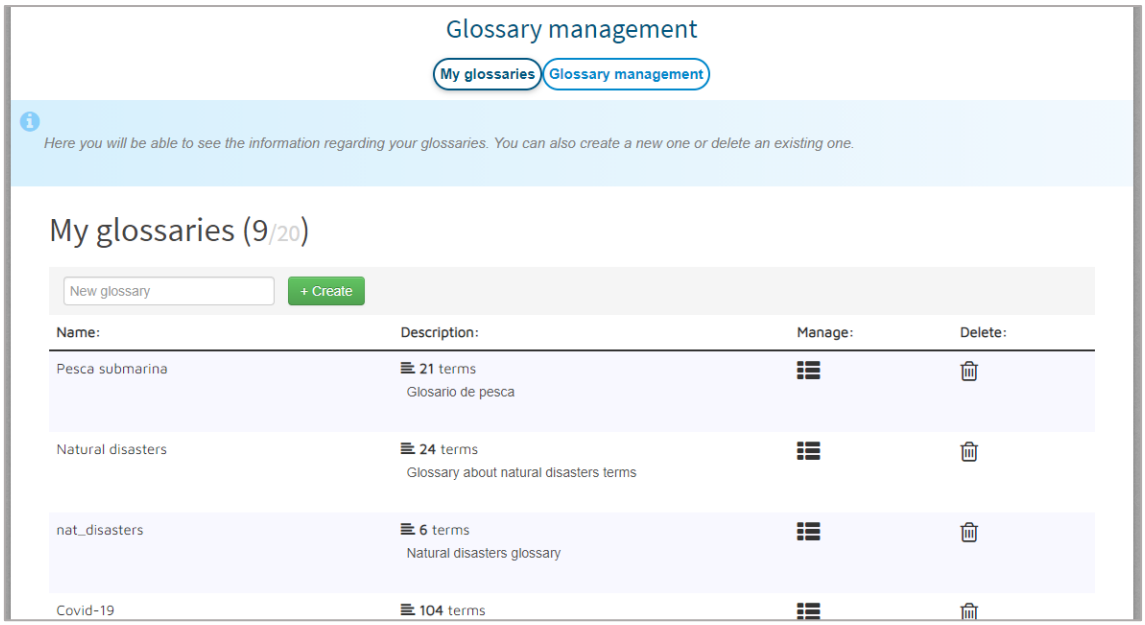
GLOSSARY

3. Dictionary and glossary management

In this section, you can see and manage the glossaries you have created.

3.1. My glossaries

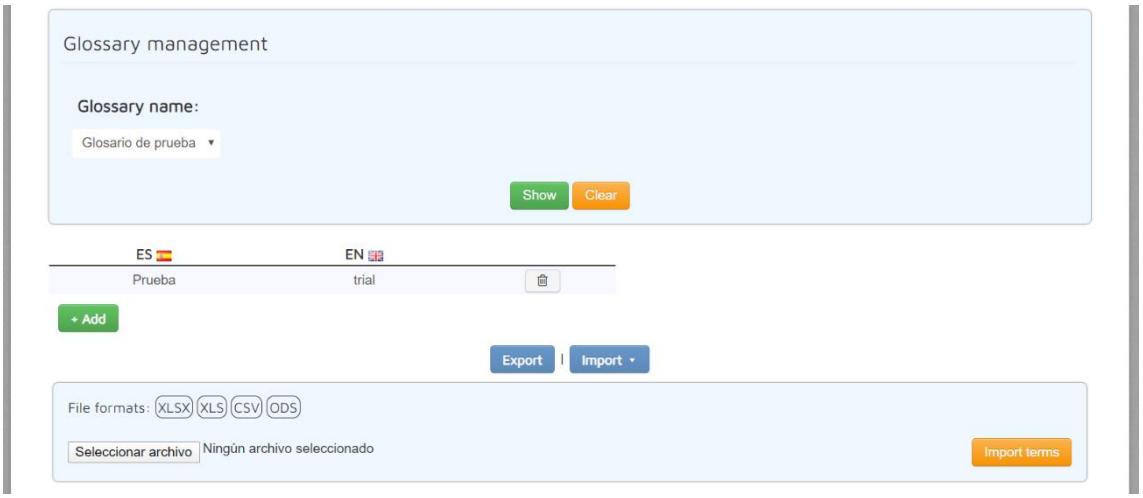
Here you can see basic information about the glossaries you have created (name, description, number of terms) as well as create and delete glossaries. In order to create a new glossary, you just need press the "Create" button and type the glossary information (name, languages and description). In doing so, an empty glossary will be created. You can change the description of any glossary by pressing on the edit icon which appears in the description when you hover your mouse over it. Then you can either confirm the change or discard it . To delete a glossary, press on the corresponding delete icon: . In order to add terms to a certain glossary you must go to *Glossary management* section (instructions are in the next section). If you click on the icon, you will be automatically redirected to the management of the selected glossary.



3.2. Glossary management

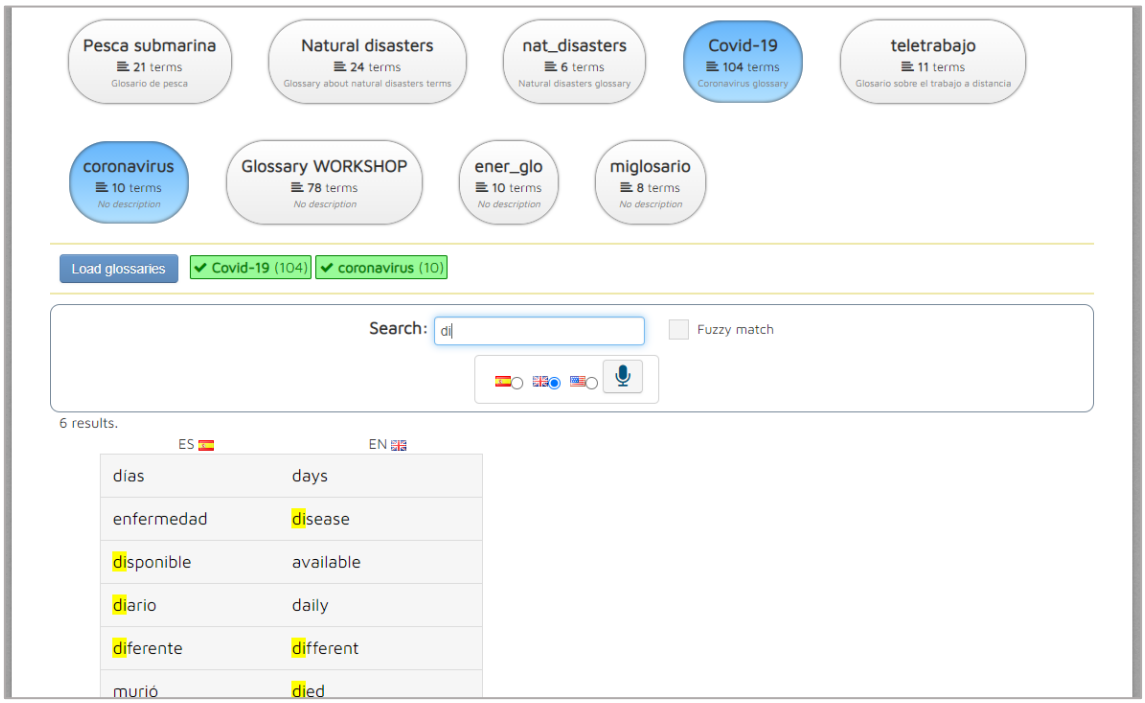
In this section you can edit the created glossaries (modify, add, etc.). Just select a glossary from the list and click on "Show". To edit any term in the glossary, press on it or on the edit icon which appears when you hover the mouse cursor over it. Then you can either confirm the change (pressing or ENTER) or discard it (pressing or ESC). If you click on the external resources search icon , a drop-down list with different resources (Google, Wikipedia, Linguee, etc.) will be displayed. When selecting any of these resources, a search for the selected term in that resource will be performed. You can also delete a glossary entry by pressing on the corresponding delete icon , or add a new blank one by pressing the "Add" button.

Moreover, the tool allows both importing terms into glossaries and exporting those glossaries. To import terms to a glossary, press the "Import" button and select the desired file (the allowed formats are XLSX, XLS, CSV and ODS). The first row of this file must include the language codes (en, es, fr…). This way all the terms included in your file will be copied to the current glossary. To export a glossary to an XLS file, just click the "Export" button and the file will be automatically downloaded.



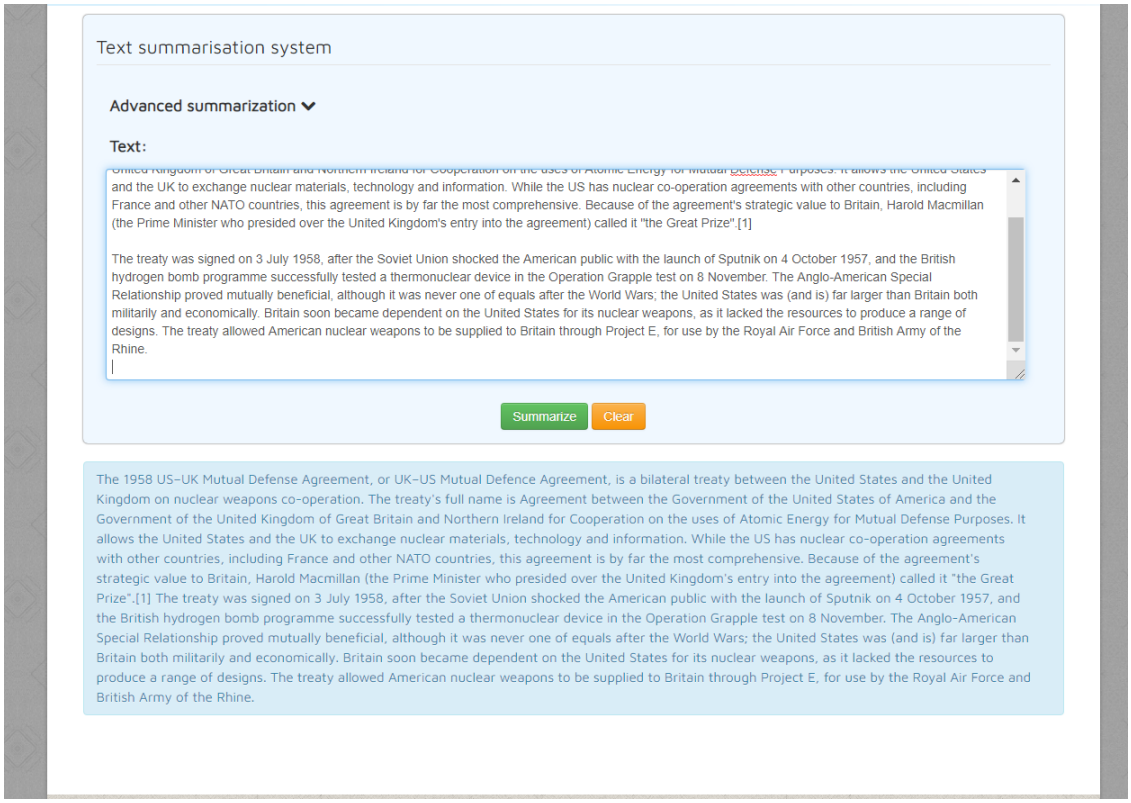
4. Glossary query

In this section you can use the glossaries previously created. You just have to select the desired glossaries, press on "Load Glossaries" and a search field will appear. When typing a term, the results will start appearing automatically in a quick and convenient way. If you check the "Fuzzy match" option, results with partial coincidences will be shown too. In addition, you can perform the search using your voice, by clicking the icon or pressing the "s" key.



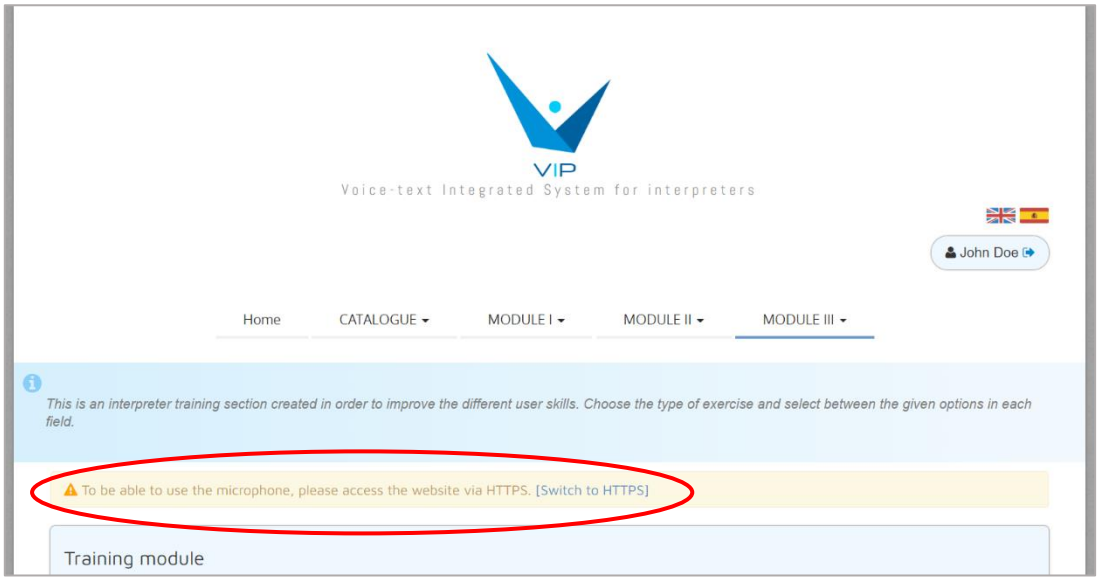
5. Text Summarisation system

In this section you can summarize a text. For this purpose, you can either load one or several plain text files (.txt), paste the text directly into the corresponding field, or use the text extraction tools to get the text from an URL or a PDF file. These tools will show the text in the corresponding field and allow the user to download this text in .txt format. Then, choose the type of summary by selecting the number of words or the percentage of the original text that it will take up. After pressing the "Summarize" button, the result will be shown next to the original text.

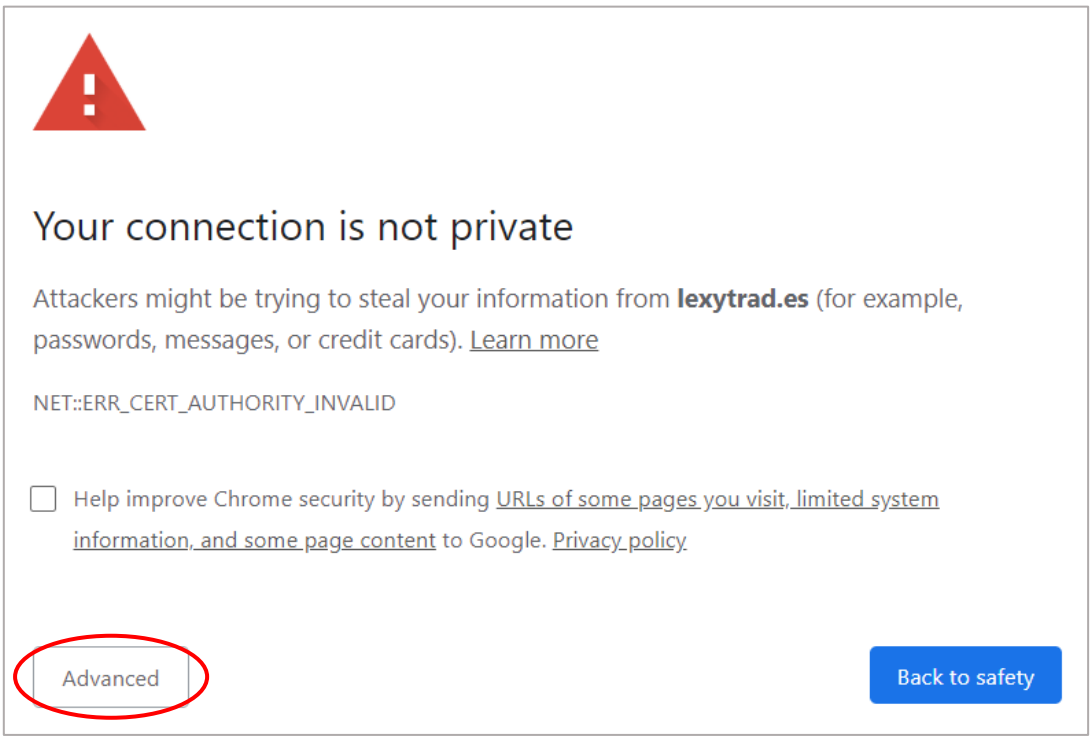


Appendix: Microphone configuration

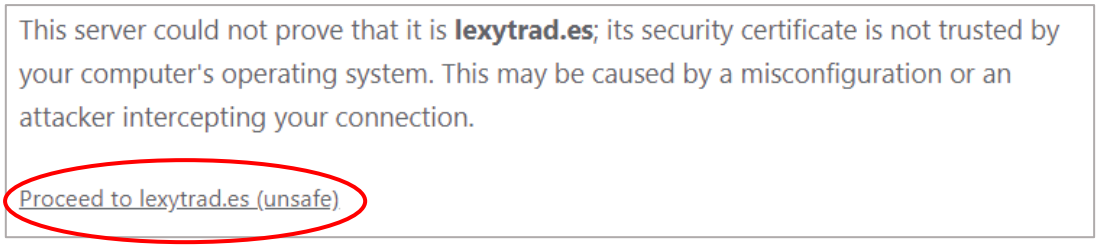
In order to use the microphone in the system, you must access the platform via the secure HTTPS protocol. Otherwise, if you are accessing via HTTP, the following warning message will appear in the sections where the use of the microphone is required:



To access via HTTPS, click on "Switch to HTTPS" and the browser will display the following warning message:



To continue, press the "Advanced" button and the following message will be displayed:



Finally, click on "Access lexytrad.es (non-secure site)" and you will return to the system, where you will be able to use the microphone.

Appendix: Verticalized text (.vrt)

Verticalized text (.vrt), also known as one-word-per-line, is a format in which each token is in a different line. It can contain additional information related to each token, as the grammatical category or lemma (separated by tabulations). Structures such as sentences or paragraphs can be set too using XML tags.

If we take this text as example:

```
This is the first sentence. This is the second.
```

Verticalized text would be:

```
This
is
the
first
sentence
.
This
is
the
second
.
```

We can include additional information (such as grammatical category or lemma) next to each token, leaving one tabulation separation:

```
This        DET        this
is          VER        be
the         DET        the
first       ADJ        first
sentence    NOU        sentence
.           PUN        .
This        DET        this
is          VER        be
the         DET        the
second      ADJ        second
.           PUN        .
```

In addition, we can set structures such as sentences (<s>) or paragraphs (<p>) by using XML tags:

```
<p>
<s>
This        DET        this
is          VER        be
the         DET        the
first       ADJ        first
sentence    NOU        sentence
.           PUN        .
</s>
<s>
This        DET        this
is          VER        be
the         DET        the
second      ADJ        second
.           PUN        .
</s>
</p>
```