Funded by the Erasmus+ Programme of the European Union
SHaping the Interpreters of the Future and of Today
www.shiftproject.eu

Using of CAT tools and term banks to collect terminological resources – a guide for interpreters

Claudia Lecci – University of Bologna

In collaboration with Félix San Vicente and Nicoletta Spinolo – University of Bologna

Introduction

This guide aims at providing an interpreter with all the skills and abilities needed to create a terminological project and detect the terminology belonging to specific domains, belonging to the orality genre.

The insights gained from this guide will help trainee and professional interpreters prepare terminological resources both for specific assignments and for more general topics. This will be particularly useful for trainees and professionals interested in the teaching materials developed within the *SHIFT in Orality* project, as remote interpreting assignments can include both booked assignments, in which the interpreter knows the topic of the interaction in advance, and last-minute assignments, for which the interpreter would benefit from a general and less specific terminological resource to use during the interaction.

The typical approach for detecting terminology belonging to a specific domain can be described through a workflow which combines different stages, starting from an **information mining stage** and ending with the creation of a **terminological resource**, namely a glossary or a terminology database.

The workflow starts with the definition of a domain and with the collection of reference materials on the Internet with the aim of acquiring the main concepts of the research topic. The second step consists in the construction of specialized comparable corpora from the web using a dedicated tool. The third stage is the corpus-based extraction of simple or complex terms with the help of a concordancing tool. The last step of the workflow is the creation of terminological entries organized in the form of glossaries and/or TermBases.

1. Domains and languages of the terminological project

For our terminological project, we have selected three specific domains, i.e. tourism, political asylum and health emergency. Our working languages will be Italian, English and Spanish.

The first step of the workflow, as mentioned above, will be the collection of reference materials in the three domains selected, that is to say the acquiring of a minimum amount of information about the field and the identification of the main concepts. This will be helpful in the following step, which is the creation of the first corpus.

2. Comparable corpora

A corpus is a large collection of authentic texts in electronic form, collected according to a specific set of criteria (Bowker and Pearson, 2002).

There are different types of corpora, e.g. reference corpora, parallel corpora, diachronic corpora, but for our purposes we will take into consideration specialized comparable corpora. Comparable corpora are texts originally produced (not translated) in the respective languages which consist of independent texts which are “similar” according to some pre-determined criteria (e.g. domain, text type, genre, publication span, topic).

The use of corpora can be applied in a wide range of disciplines, for example lexicography, language learning, socio-linguistic studies, computational linguistics, translation studies, interpreting studies (Bowker and Pearson, 2002). In this case we will use corpora for extracting specialized terminology belonging to some specific domains.

2.1. Building corpora with a specific tool for corpus construction - BootCaT front-end

The Internet represents a real source of corpora built for specific purposes (e.g. an interpreting or a translation task, the creation of a glossary or a terminological database, etc.). These corpora are extremely important resources for language professionals who work with specialized languages. Even if it is possible to construct a web-based corpus through manual queries and downloads, this process is more time-consuming, so a group of linguists from the University of Bologna developed the BootCaT front-end, which is a graphical interface for the BootCaT toolkit.

Three exists other similar tool, i.e. the corpus creator of the Sketch Engine tool[1], which is a commercial software, and the corpus tool[2] of Translator Bank, free and developed at the University of Mainz. Anyway, in this guide we will use BootCaT.

BootCaT automates the process of finding texts on the web and collating them in a single corpus. The pipeline allows varying levels of control. In the first step, users provide a list of single or multi-word keywords, called ***seeds***[3], to be used as for text collection. These are then combined into "tuples" (a variety of combinations of your seeds) and sent as queries to a search engine, which returns a list of potentially relevant URLs. At this point the user has the option of inspecting the URLs and trimming them; the actual web pages are then retrieved, converted to plain text and saved in txt format. The corpus can thus be interrogated using different concordancers.

2.1.1. Installing BootCaT

The software is free, open-source and multiplatform and installing it is very easy and fast. You just need to go the BootCaT installation webpage (http://bootcat.dipintra.it/?section=installation), download the version fit for your operating system (Windows, Mac OS X or Linux/Unix) and run the installer. Once installation is successfully completed, the "BootCaT front-end" icon will appear on your desktop.

2.2. Building comparable corpora: different stages

In this section we will guide you through the construction of three comparable corpora in Italian, English and Spanish in the domain of the health emergency. Let us start from the corpus in Italian.

[1] https://www.sketchengine.co.uk/user-guide/user-manual/corpora/create-from-files/
[2] http://www.academia.edu/12717759/TranslatorBank_-_Corpus_tool_for_translators_and_interpreters
[3] We call these keywords *seeds* because, thanks to a particular process, they give rise to corpora, just like real seeds give life to plants or trees.

The first thing to do is launching BootCaT simply by double-clicking on the icon which appeared on your desktop. This will start a wizard for the creation of our web corpus.

In the *Welcome* screen you will find some basic information about the BootCaT method, just read it and click on *Next*.

The second screen will be the *Project Definition* one. Here you have to choose a name and select a language for the corpus.

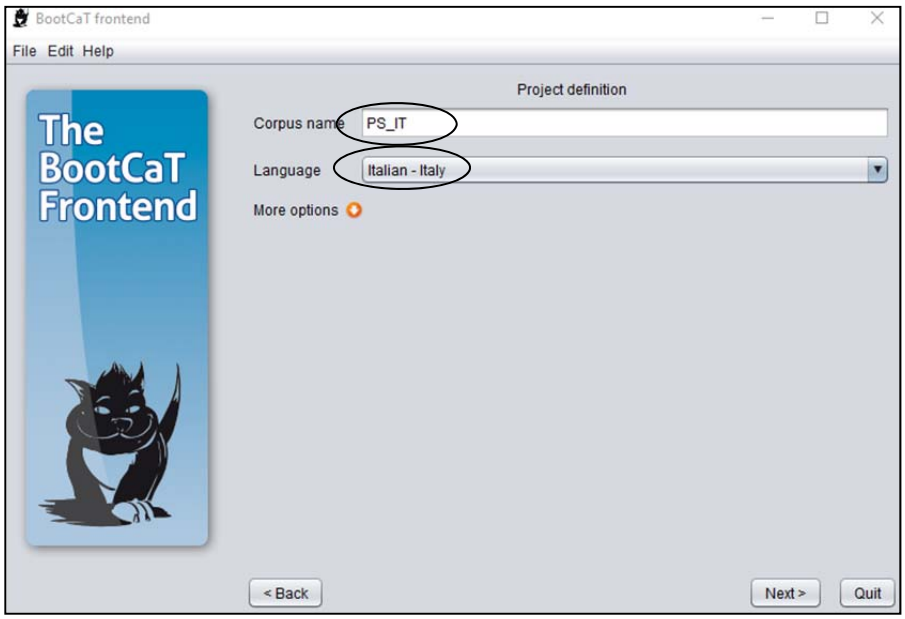

Figure 1 BootCaT project definition screen

In the third step, you have to choose the corpus creation mode among the four proposed. Choose the *Simple Mode* for this corpus and click on *Next*.

In the following screen you have to provide the seeds that BootCaT will use to generate the queries that will be submitted to the search engine. The minimum number of seeds you must provide is 5; so here we used 5.

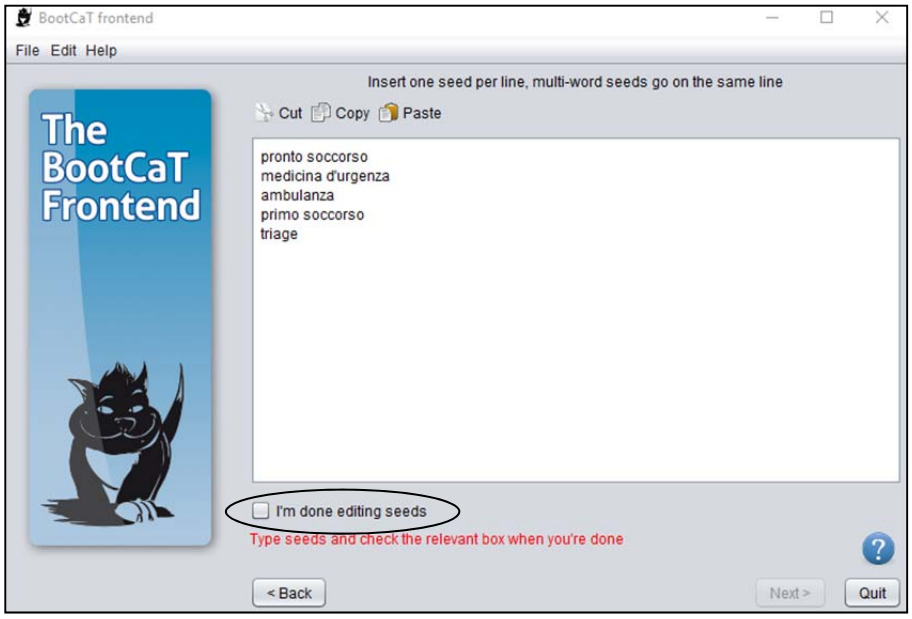

Figure 2 Inserting the seeds for the corpus creation

Once you have provided the seeds of your choice, check the *I'm done editing seeds* box and click on *Next*.

The seeds you provided in this step will be randomly grouped to form tuples, which will be submitted as queries to the search engine.

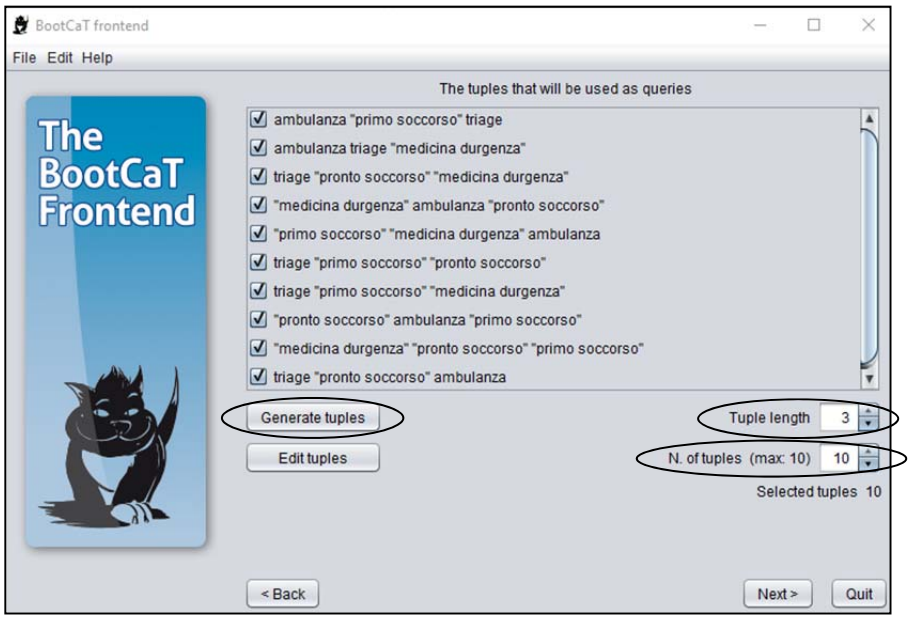

Figure 3 The tuples generated

You can choose the number of tuples to be generated; of course, the number of possible random combinations is finite and depends on how many seeds you provided. Since we provided 5 seeds, we can generate a maximum of 10 tuples.

You can also modify the length of the tuples (i.e. the number of seeds forming it); typical values for this option are:

- 3 if you want to build a specialized corpus (this is our case)
- 2 if you are creating a general language corpus

Now click on *Next* to proceed to the next step.

In order to query the search engine we need to provide BootCaT with the Bing Search Engine API[4]. Follow the instructions available here to obtain it: https://azure.microsoft.com/it-it/try/cognitive-services/?api=bing-web-search-api. Now paste your Search Engine Key in the relevant box, and click on *Next*.

Figure 4 BootCaT collects URLs

In this window BootCaT searches the web via the search engine, looking for pages that contain the tuples that were generated in the previous step. The search engine will return only a limited number of pages for each query (i.e. tuple) we submit; the default value is 10 URLs per query and we will not change it. This might take a while, depending on the number of tuples, Internet traffic and speed of your connection.

[4] Until the time of writing we are able to use the Bing Search API with the 30 days trial license, then BootCaT uses the Bing search engine to find web pages relevant to our domain. Anyway, since the aim is to ensure a completely free tool, BootCaT developers introduced the possibility to query a search engine via an external browser. Using this method, it is possible to circumvent search engine key limitations (http://docs.sslmit.unibo.it/doku.php?id=bootcat:help:external_browser).

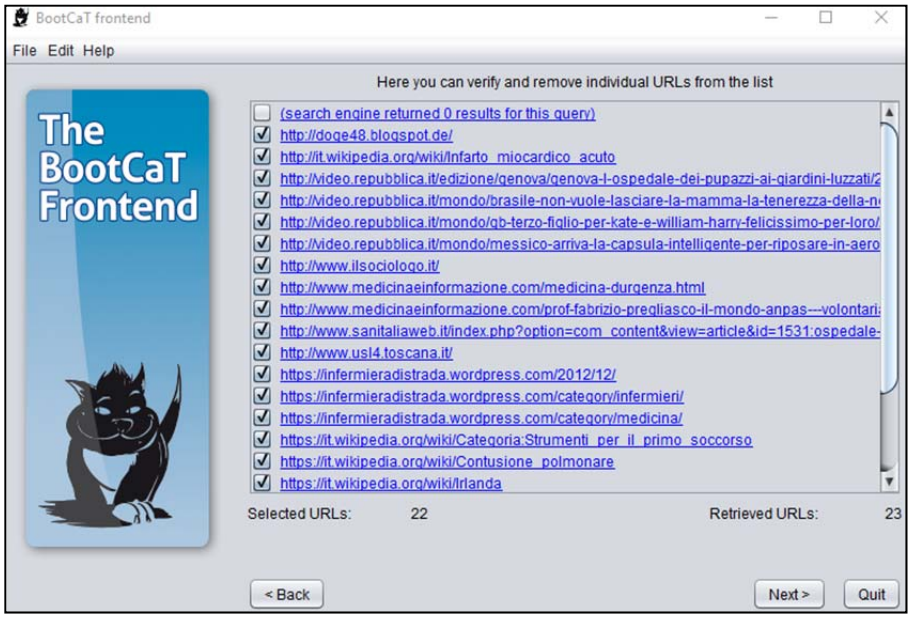

Figure 5 URLs collected

In this step you can choose to remove URLs you think might not be interesting. You can also click on the URLs to visit the web page and decide whether you want to include the page in your corpus or not. Once finished, click on *Next*.

In the final step BootCaT will download the web pages detected and will automatically clean them. In particular, HTML code will be removed and boilerplate (i.e. objects like menus, navigation bars, ads, disclaimers, automatic error messages, that are unlikely to be of interest for corpus users) will be stripped.

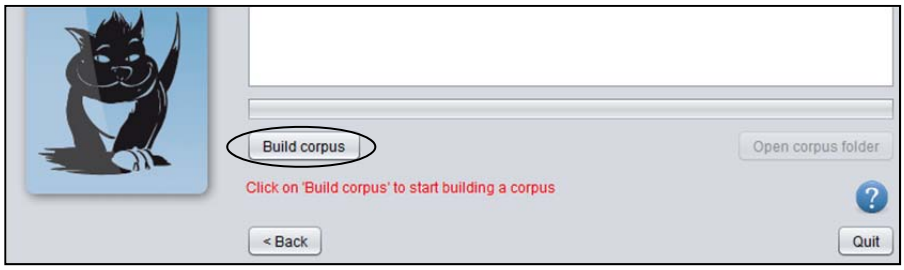

Figure 6 Building the corpus

In this last window, you just have to click on *Build Corpus* to start the corpus creation process. Once the download is complete click on *Open corpus folder* to open the folder containing the corpus and all the relevant files.

Now that we have created the corpus on health emergency in Italian, how do we create the comparable corpora in English and Spanish? We have to choose the "right" target language terms, that is to say to identify and reproduce the features of the specific domain in the target language(s). One method could be taking into account the seeds used for the source language and searching for the equivalent seeds in the target language(s).

Do it in English and Spanish and then create the two comparable corpora in the two target languages selected following the steps indicated for the corpus in Italian.

I attach to this guide three trilingual corpora in the three domains selected at the beginning. For the examples here, I used the health emergency corpus in Italian, but the workflow would have been the same also with all the other domains and corpora.

3. Extracting multilingual terminology from corpora

According to our workflow, after creating the three comparable corpora, we have to extract terminology from them with the help of a concordancing tool.

However, before starting with the extraction, we want to provide an overview of the main concepts about terminology and terms in general.

3.1. Brief introduction to terminology

Terminology is an interdisciplinary science applied to the systematization and standardization of specialized languages. Its object of study are terms and their use in real communicative situations. The three main functions of terminology are:

- systematically describing terms belonging to specialized fields;
- facilitating the transmission/dissemination of technical and scientific knowledge by creating terminological resources;
- standardizing lexicon to allow unambiguous communication of specialized contents and their interpretation/translation. (Cabré, 1998)

Words and phrases with a precise, specialized meaning are called terms. The analysis and systematization of terms aims at facilitating correct transfer of information at the "local", national and international level.

3.2. Extracting terminology from corpora

Before extracting terminology from corpora we have first of all to classify terms. The main distinction is between simple or complex terms or phraseologies, according to the number of words involved. One word is a simple term, two or more words are complex terms, two or more words within a syntactic structure are

phraseologies. Moreover, we have to take into account the function-meaning of words, and in this case the distinction is among nouns, verbs and adjectives.

But, in practice, how can we extract terminology from corpora? What corpus analysis tools are available?

There exists a number of concordancers, free or commercial, that we could use to analyze our corpora, for example AntConC, TextStat, Wordsmith Tools, and many others. In this guide we will use AntConC, which is a user-friendly concordancing tool with many functionalities which can be downloaded for free from this URL: www.laurenceanthony.net/software.html. Like BootCaT, AntConC is a multiplatform software, so you just have to choose your version according to your operating system, run the installer and install it on your computer.

3.2.1. Analyzing corpora with AntConc

Once installed, we can use our concordancer to analyze the corpora and to extract multilingual terminology.

The main operations that we can carry out with AntConC are:

- creating frequency lists;
- calculating keywords;
- creating concordances;
- making collocations;
- generating N-grams.

The **frequency list is** the most basic tool and it allows us to discover how often individual lexical items occur in an individual corpus. Thanks to a wordlist we can also learn about the size of our corpus, i.e. how many words it contains.

Keywords allow us to identify which words are particularly frequent in our corpus, hence words that are particularly typical of the domain in question. Keyword lists are obtained comparing specialized corpora with reference (general) corpora, containing different varieties of texts and domains.

The **concordance** tool allows us to search for a specific term (simple or complex) within the corpus. The result will be a list of examples taken from your corpus file(s) and presented together. Concordances are presented as KWIC (Key Word in Context) output, whereby the search term or phrase (the node) is positioned in the middle of the screen, with the co-textual elements positioned to the right and left of it (these words are known as the *span*). The span includes words that are positioned to the right of the node word.

For example, if we make a concordance search in our corpus on health emergency in Italian for the word *lesioni*, we could obtain the pattern *il paziente può presentare lesioni gravi su tutto il corpo*. In this case the node word is *lesioni* and the words on the left and on the right are called *span*. An interesting results in this case could be the complex term *lesioni gravi*, where *gravi* is located one word right (the span in this case is 1 right).

You are also able to create a concordance output from individual words listed in a wordlist or a keyword list, just clicking on them. This will automatically take you to the concordance outputs of this item or phrase.

Collocations are lexical items that regularly co-occur (Halliday and Hasan, 1976). AntConC uses different statistical measures, which help us define the strength of the pattern co-occurrences.

Finally, AntConc not only provides information on the most frequent words that occur in a text, but also on the most frequent groups of words. Among these you could also find significant complex terms, like for example *temperatura corporea* in our corpus in Italian on health emergency. The way to obtain lists of groups of words is generating lists of **n-grams**

3.2.2. Using AntConc with our corpora

The aim of this activity is creating a trilingual glossary (Italian, English, and Spanish) in the domain of health emergency. The basic principle is detecting first the terminology in the first language and then matching the terms inter-linguistically in the other languages.

Let us start with our corpus Italian corpus.

Launch AntConC and upload the corpus PS_IT[5] clicking on *File* and then on *Open Dir*[6]….

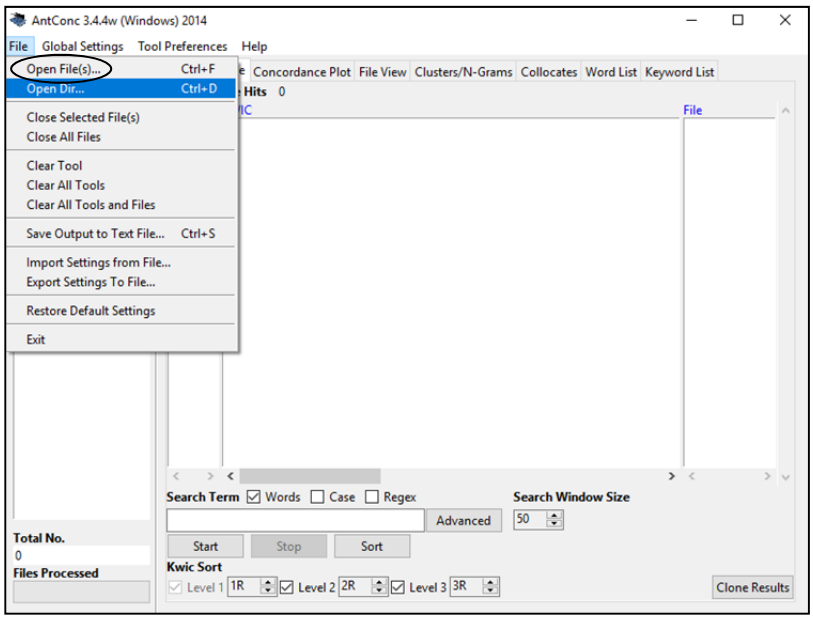

Figure 7 Opening the corpus

Now click on the *Wordlist* tab and then on *Start* to generate the wordlist of this corpus.

[5] The corpus is available among the materials attached to this guide
[6] BootCaT has stored in a folder called *corpus* all the .txt files which are part of the corpus.

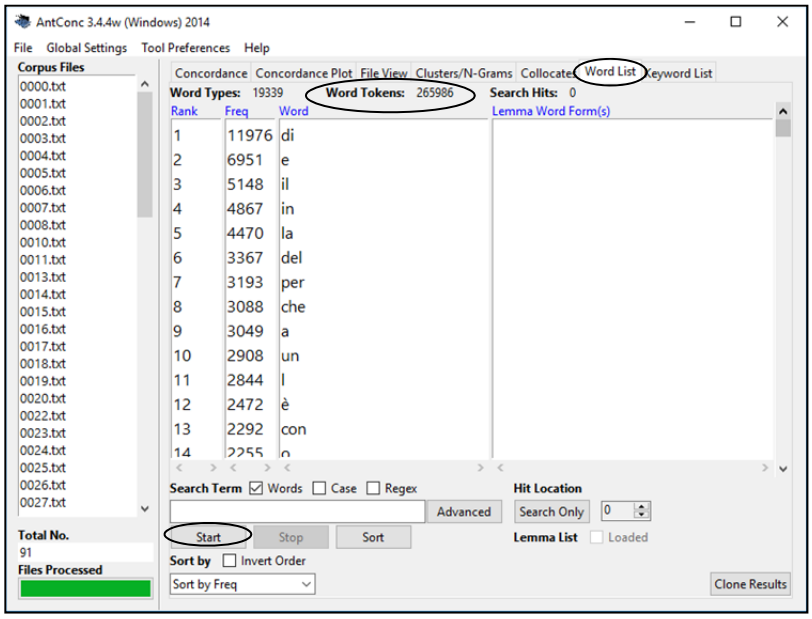

Figure 8 Generating a wordlist

As you can notice, in the first part of the wordlist we have a number of function words, which are of course the most frequent in a collection of texts. To find content words [7]we have to scroll down the list. Through the creation of a wordlist we can also know the size of the corpus, in this case we have 265.986 words.

Anyway, the most effective way to search for typical terminology in a specialized corpus is by generating a keyword list.

Click on the *Keyword List* tab, then on *Tool preferences* > *Keyword List*. Under *Reference Corpus* make sure that *Use raw files* is checked, then click on *Add Files* to select the reference corpus[8] and then on *Load*, to load it.

Then go back to the *Keyword List* tab and click on *Start*.

[7] Words that carry semantic content, e.g. nouns, adjectives, verbs, etc.

[8] For Italian and English we will use The WaCky corpora, a collection of very large general corpora generated from the web (Baroni, Bernardini, Ferraresi and Zanchetta. 2009). For Spanish we will use a large corpus created with BootCaT within the research activity of the Department of Interpreting and Translation at Forlì. The reference corpora will be attached to this guide.

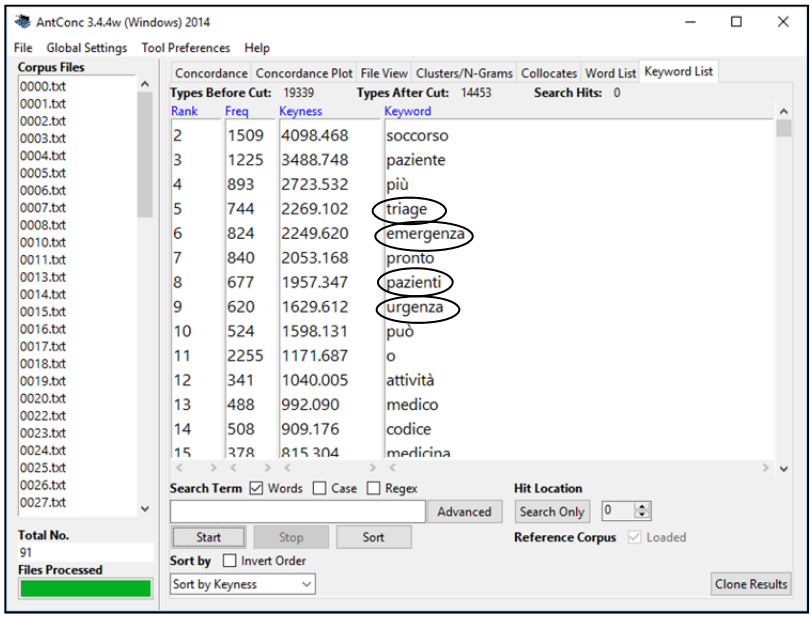

Figure 9 Generating a keyword list

This is the result of the operation and, as you can see, here we have a list of content words. Scrolling down the list, you will find a number of very interesting terms belonging to the domain in question, which could be potentially included in the glossary (e.g. *pazienti*, *triage*, *urgenza*, *emergenza*, etc…).

But, to learn more, we can continue analyzing the corpus by making some concordances. Suppose that we want to discover the reason why the apparently general word (for this domain) *pronto* has been positioned in the very first rows of the keyword list. Click on the word *pronto*, generate a concordance, and have a look at the results.

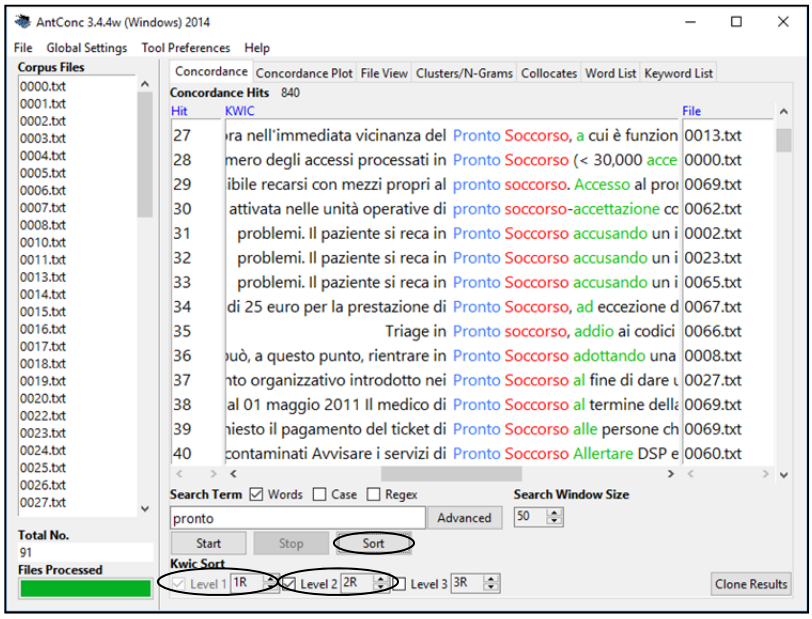

Figure 10 An example of concordance

In this case, we have highlighted the two words following *pronto*, positioned at the right hand of the node word, using the *Sort* button. As you can see, thanks to the concordance search, we have discovered that the significant word is *Pronto Soccorso*, and not just *pronto*. We could include this complex term in our glossary.

Now, we could try to make a concordance search for the term *lesioni*, which we can find if we scroll down the keyword list. Sorting the first word on the right side we have very interesting results, like for example the terms *lesioni gravi*, *lesioni superficiali*, *traumatiche*, *lesioni cutanee*, *lesioni interne*, *lesioni midollari*, and many others.

Let us try now to generate an n-gram list.

Click on the *Clusters/N-Grams* tab, make sure that the *N-grams* box is checked, do not change the n-gram size (the default value is 2) and then click on *Start*.

In the first part of the n-gram list we have the most frequent occurrences (like *pronto soccorso*, *centrale operativa*, *primo soccorso*), but if we scroll down the list we can also find some interesting occurrences, like for example *temperatura corporea*, *trattamento immediato*, *unità operativa*, etc.

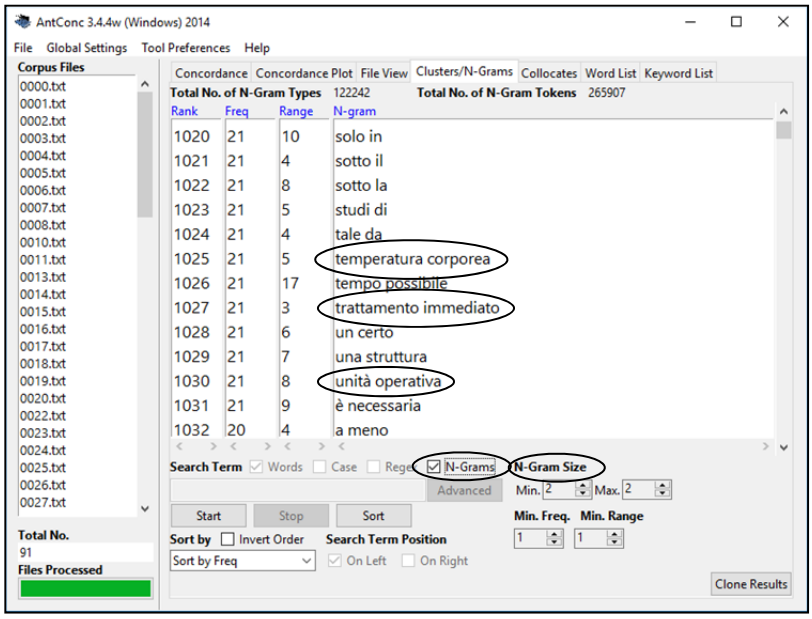

Figure 11 Generating an n-gram list

If we try to increase the N-gram size, for example at 3, we find complex terms like *medicina di emergenza*, *mezzi di soccorso*, *pericolo di vita*, etc, which we can consider as belonging to our domain.

Using AntConc is very easy and stimulating, because you can jump from one tab to another, searching for new terms and discovering new features and solutions.

Once finished the search in the source language corpus, it is time to analyze our comparable corpora. The best way to find equivalents for the terms detected in the source language is repeating for the target languages the same strategies used for analyzing the source corpus. You can do it separately for English and for Spanish, or you can do it in parallel, launching the software twice and loading both corpora.

Now generate a wordlist, a keyword list, an n-gram list (with n-gram size 2 and 3) and make concordances for the uncertain terms, as we did for Italian. With this method you should easily identify translations for the source terms detected. If you are in trouble because you are not able to find some equivalents, try to make hypotheses of possible translation equivalents for the terms in questions and then make a concordance search in order to check whether they are present and frequent in your corpora.

While searching for equivalents, take note of the terms detected, as the next step will be the creation of our glossary.

4. Cataloguing terminology

Let us now catalogue the multilingual terminology extracted from the comparable corpora. The most useful cataloguing method in this case is using an Excel file. In the Excel format you have pre-determined rows and columns and you can take advantage of this structure to create a well-ordered and organized glossary.

Here is a section of the trilingual glossary on health emergency that we created with the help of this guide.

IT	EN	ES
pronto soccorso	emergency department	servicio de urgencias
pazienti	patients	pacientes
triage	triage	triaje
emergenza	emergency	emergencia
lesioni gravi	seroius injuries	lesiones graves
lesioni superficiali	minor injuries	lesiones menores
temperatura corporea	body temperature	temperatura corporal

Figure 12 An extract of our Excel glossary

Another useful format could be a tab separated text file, i.e. a structure like this**:** *Term_L1 TAB Term_L2 TAB Term_L3*.

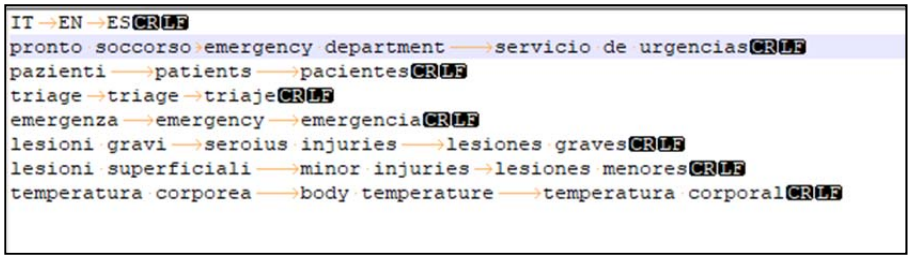

Figure 13 An extract of our tab separated glossary

This format is highly recommended in the event that your customer requires for example the translation of documentation relating the same matter and domain you are interpreting. The tab separated txt format indeed is fully compliant with the free and open-source Computer Assisted Translation tool[9] OmegaT (downloadable from http://omegat.org/) that you could use for your translations.

You can also obtain a tab separated txt file just copying and pasting the Excel glossary in an empty txt file.

Another option for cataloguing of your terms could be the conversion of the Excel file in a more complex structure, namely in a terminology database (or TermBase) containing source and target terms in a number of languages, but also other additional information, such as definitions of the terms, contexts, notes, etc. In fact, this kind of term cataloguing is not really intended for your communicative situation (dialogue interpreting), but is more suitable and appropriate for terminologists or specialized translators.

[9] A Computer Assisted Translation (CAT) tool is a software based on the principle of re-use of previously translated sentences and terms. CAT Tools are useful for accelerating the process of translation and for ensure consistency in terms of sentences and terms.

The terminology management tool which allows the construction of complex TermBases is *SDL MultiTerm*[10], a commercial tool developed by *SDL Trados*. One of the applications of *SDL MultiTerm* is *SDL MultiTerm Convert*, a wizard that enables you to convert Excel glossaries in .sdltb files.

5. Conclusion

Using the correct terminology when you communicate with people speaking other languages is a key issue. Right terms ensure a correct transfer of information avoiding misunderstandings and communication problems, a very important element when we deal with socio-cultural contexts like those considered in the SHIFT in Orality project.

[10] http://www.sdltrados.com/products/multiterm-desktop/

Bibliography

Baroni M., Bernardini S., Ferraresi A. and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43(3): 209-226

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms form the web. In Proceedings of LREC 2004 (pp. 1313-1316).

Bernardini, S. and Ferraresi, A. (2013). Old needs, new solutions: Comparable corpora for language professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum e P. Fung (edited by), Building and Using Comparable Corpora (pp. 303-319). Springer, Dordrecht.

Bowker, L. and Pearson, J. (2002). Working with Specialized Language: a Practical Guide to Using Corpora. Routledge, London and New York.

Cabré, M.T. (1998.) Terminology: Theory, methods, and applications. John Benjamins, Amsterdam and Philadelphia.

Fantinuoli, C. (2006). Specialized corpora from the web and term extraction for simultaneous interpreters. In Marco Baroni and Silvia Bernardini, editors, Wacky! Working Papers on the Web as Corpus, pages 173–190. GEDIT, Bologna.

Pearson, J. (1998). Terms in Context. John Benjamins, Amsterdam and Philadelphia.

Zanchetta, E. (2011). Corpora for the masses: The BootCaT front-end. Pecha Kucha presented at the Corpus Linguistics 2011 Conference, University of Birmingham, Birmingham.

Sitography

A simple guide to using AntConc:
http://www.laurenceanthony.net/software/antconc/resources/help_AntConc321_english.pdf

BootCaT front-end tutorial (using an external browser):
http://docs.sslmit.unibo.it/doku.php?id=bootcat:help:external_browser

BootCaT front-end tutorial:
http://docs.sslmit.unibo.it/doku.php?id=bootcat:tutorials

For more information on SDL MultiTerm:
http://www.sdltrados.com/products/multiterm-desktop/

For more information on the Sketch Engine tool:
https://www.sketchengine.co.uk/user-guide/user-manual/corpora/create-from-files/

To download and install AntConc:
www.laurenceanthony.net/software.html

To download and install BootCaT:
http://bootcat.dipintra.it/?section=installation

To download and install OmegaT:
http://omegat.org/

To download and install TranslatorBank:
http://www.academia.edu/12717759/TranslatorBank_-_Corpus_tool_for_translators_and_interpreters

To obtain a Bing Search API:
https://azure.microsoft.com/it-it/try/cognitive-services/?api=bing-web-search-api