Universidad de Valladolid

FACULTAD de FILOSOFÍA Y LETRAS
DEPARTAMENTO de FILOLOGÍA INGLESA
Grado en Estudios Ingleses

TRABAJO DE FIN DE GRADO

The Usefulness of Specialized Monolingual Corpora in Translation activity

Sandra Ivonne Silva Aedo

Tutor: Beatriz Méndez Cendón

Curso 2013/14

ABSTRACT

The aim of the project is to point out the great usefulness that specialized monolingual corpora can provide translators in their challenging task of medical translation. To carry out our purpose, a specialized corpus is built. We take into account the criteria of corpus compilation and corpus analysis. Once it is built, we compare the usage of our corpus with lexicographical resources, such as: specialized dictionaries and glossaries, in order to indicate the benefits that translators can obtain from corpora. Finally, our conclusion is reachedbe on the basis of the importance of using corpora in translation activity.

Keywords: medical translation, corpora, phraseology, KWIC concordance, collocations.

RESUMEN: El propósito del proyecto es señalar la gran utilidad que el corpus especializado monolingüe puede ofrecer a los traductores en la difícil tarea de la traducción médica. Para ello, compilamos un corpus especializado y, a continuación, analizamos los criterios que hemos tenido en su compilación y análisis. Más adelante, llevamos a cabo una comparación de su uso con recursos lexicográficos frecuentes, como los diccionarios o glosarios especializados, para indicar los beneficios que los traducotres pueden obtener de los corpus. Por último, ofrecemos una conclusión basada en la importancia de usar un corpus en la actividad de la traducción.

Palabras clave: traducción medica, corpus, fraseología, concordancia KWIC, colocaciones.

Index

List of tables

1. Introduction

The production of a text into another language, most specifically in a specialized text, covers a set of aspects regarding terminology and translation. Therefore, it does not only involve the fact of transferring words from one language into another, but the meaning of the source text must be expressed according to the rules of the target language in order to be accepted by the audience. In this regard, the translators play the role of intermediary between two languages and cultures.

Every translator follows a translation process to carry out this assignment. Following the classification of Sven Tarp (2007), the translation process can be classified into three phases: "preparation, translation and revision".
The preparation phase deals with the collection of data in order to acquire knowledge of the topic involved.

The second phase (the translation) refers to the understanding of the source language (SL). This phase translators may find difficulties since they find new terms. In order to solve their problems, they look for the equivalences of the terms in dictionaries or glossaries. The problem lies in the fact that these sources provide them with several options and they may select the wrong ones or the less appropriate ones. As a consequence their translation may not be accepted by the LSP target readers. It is due to the fact that they lack of information and therefore they do not satisfy the needs of translators.

The last phase is the production of the target text (TT), in which translators need, apart from the equivalence of the terms, grammatical and syntactic information about the terms. It is also difficult that translators find this information in the lexicographical resources mentioned above. In most cases, these resources are not helpful because they are not outdated.

To summarize, a good translation into L2 does not only involve using the appropriate terminology but also use the right grammatical structures in order to make the target text sounds naturally. In other words, translators have to achieve a translation that must be as close as possible to original texts produced within a particular field of LSP.

As it is well known, the Internet is a useful source that has changed translation activity together with the increasing availability of corpora. It is true that a corpus cannot respond to all answers of translators, but provides them with information that the previous resources do not offer. A corpus, therefore, provides several advantages for their users. For example, since it contains thousands of words, translators can identify and extract collocational, conceptual and terminological information using some software programs. The fact that the texts are gathered means that they can be accessed easily by translators.

For the reasons mentioned previously, the aim of our project is to compile a corpus that consists of original texts in a particular subject field in order to point out that it can become a considerable help for the translators. We also demonstrate how software programs can be used by the translators in order to solve their translation problems, and in this way, they can produce a high-quality translation.

Our project is divided into four parts: the first part deals with English specialized language and its main characteristics. We believe that the understanding of specialized language is essential for translation. In the second section, we introduce a definition of corpora and explain the importance of corpora in translation studies. We also describe different types of corpora. In the third chapter, we explain the characteristic of our specialized corpus and its compilation process, which is followed by the corpus analysis that consists of the retrieval of terms and collocations. Finally, the last part deals with, an analysis in which examine the role that a specialized monolingual corpus plays in translation activity by using a Spanish article. In this regard, our aim is to point out its usefulness, but also to highlight the differences that exist between using this tool and common resources (e.g. dictionaries and

glossaries). The project ends, with a conclusion in which we summarize the main points of our work.

2. English Specialized language

The translation of a specialized text is an activity that is totally different from a text belonging to general language. One of the reasons is that specialized language has its own terminology and its own phraseology and it makes that specialized language different from the general language, although they also share similar characteristics.

Before focusing on the main topic of our project, we would like to indicate what English specialized language is and to demonstrate the differences and similitudes between general and specialized languages since we consider that this information should be learnt by translators before they begin to translate a specialized text.

First, English specialized language is part of LSP (Language for Specific Purposes), which may be defined as a language used by a specific group in order to communicate and discuss within a particular field by using a specialized knowledge. This term comprises English specialized languages, French specialized languages, German specialized languages, Spanish specialized languages, etc. Therefore, it is said that LSP covers all global specialized languages (Méndez Cendón 2012-13).

Second, although LSP differs from LGP (Language for general purposes) there is an interrelation between them. It is due to the fact that they share the same grammatical structures. In other words, there aren't two different grammars and, therefore, we find that grammatical structures of LSP exist in LGP. They also share many words thanks to **de-terminologization**. It is a process by which specialized words start being used in general language, and it may be for several reasons such as the emergence of scientific knowledge in the society (Meyer and Mackintosh, 2000). For example, "virus" is a term from the medical field that is nowadays used in the computer field and by the people in general to refer to a program that affects their computers.

LSP has its own characteristics and these are (Méndez Cendón 2012-13):

- Different degrees of abstraction within LSP: topic, communicative situation and users.
- Monofunctional character because it used by a restricted number of users who learn specialized language voluntarily.
- Knowledge of specialized language is mandatory for members who use it in an expert community.
- Each subject field has its own specific terminology.

As mentioned above LSP and LGP share grammatical characteristics, but in LSP we find a high frequency of the following structures (Méndez Cendón 2012-13):

- Specific terminology
- Prefixes and suffixes
- Premodifiers (complex noun phrases)
- Compound nouns
- Passive constructions
- Reduce relative clauses
- Nominalizations

Third, LSP is classified according to specialized fields, communicative context and users or participants. Pearson (2002:27) classifies LSP according to users: expert, semi–experts and non – experts. As there are three types of users, also there are different levels of LSP communication and these are the following:

Expert- to expert communication: The writer and the reader share a high knowledge of specialized language, and therefore they understand the terminology and complex structures that are used in texts.

Expert- to semi–experts communication: The writer and the reader do not share the same level of knowledge. However, the reader may be familiar with some of the terms used within the field. In this type of communication, the writer provides the users with some explanations in order to make them understand the terms.

Expert- to non–expert communication: The situation in which the reader is not familiar with terminology. For this reason, the author uses words belonging to general language in order to describe the concepts in an easier way.

Finally, it is important to say that if the user is interested in learning LSP he needs to acquire two types of knowledge: linguistic and conceptual as Bowker and Pearson (2002: 30) point out. Linguistic knowledge refers to terms, collocations, grammatical structures and stylistic features, whereas conceptual knowledge deals with the information about the specialized concepts.

3. Corpora in translation

In this section, we provide an overview about the concept of corpora, followed by their importance in translation and finally their types.

3.1. Definition of corpora

There are several definitions for a corpus and maybe one of the easier to understand is the one that Bowker and Pearson formulate in their book *Working with Specialized Language: A Practical Guide to Using Corpora* (2002:9). They state that a corpus can be described as "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria".

Therefore, the texts that a corpus comprises have to be selected according to a specific purpose to be used as a representative sample of language, or a particular aspect of language.

3.2. Importance of corpora

The usage of corpora has been introduced in translation studies recently. Electronic corpora have changed translation training and, scholars have published several studies about this. Mona Baker (1993) was one of the first to discuss the importance of corpora in translation. She says that they can help translators to investigate translations. In other words, translators would analyze the texts and extract relevant information for their translations.

Baker (1996) also indicates that comparable corpora are very useful for translators because they allow them to extract valuable information when translating, due to the fact that this type of corpora contains original texts and not translations. On the other hand, Maeve Olohan points out:

> Pearson (2003) and Bowker (1999) state that comparable corpora can be a useful resource for students and teachers involved in translator training, used for checking terminology and collocates,

> identifying text-type-specific formulations, validating intuitions and providing explanations for appropriateness of certain solutions to problems. (2004: 172).

Therefore, this type of corpora provides translators with different types of information which help in their translation assignment.

Silvia Bernardini is another of the scholars who gives importance to the usage of corpora in translation. She indicates that the corpus helps translation students to obtain a better translation. She says "The result is a better – documented, more accurate, as well as more fluent translation" (Bernardini 2004:20).

Moreover, the development of corpus analysis tools such as Wordsmith and AntConc have also led to several translation scholars publish studies based on corpora in translation. Some of them are Belinda Maia (2005) and Natalie Kübler (2003).

Maia (2005) argues that the use of general corpora allows translation students to go beyond the dictionary because they can observe words in context. Additionally, Kübler (2003) focuses especially on technical translations and indicates: "Learning to use corpora and corpus analysis tools can give future translators the technical skills that were usually not associated with translation".

Therefore, there is no doubt that the usage of corpora has played an important role in the translation field and continues to do so. Thanks to them, the translators can investigate style and usage, as well as search for information about contexts in which a word occurs. As Bowker and Pearson (2002) already mentionated, a corpus can be used as a translation resource.

However there is no just one type of corpus, but there are different types of corpora and translators should determine the type of corpus they will compile.

3.3. Types of corpora

Corpora are usually classified according to their purpose. The types of corpora are "general corpora", "specialized corpora", "parallel corpora", "comparable corpora" and "monitor corpora".

As for "general and specialized corpora", according to Baker (1995), a general corpus is used for a wide range of purposes, giving all the relevant varieties of the language to the researchers. An example of this type is British National Corpus (BNC) for English Language and Real Academia Española (RAE) for Spanish language. In contrast, a specialized corpus is smaller and it is due to the fact that it focuses on a specific aspect of language.

Other types of corpora are "parallel" and "comparable". They are different because a parallel corpus is "composed of a set of the source texts and their translations" (Bravo Gozalo, 1998: 225), whereas a comparable corpus only contains original texts in two or more languages. In order to be a comparable corpus, both subcorpora have to share common features such as size, date of publication, genre, subject – field, etc. If not, they are just a collection of texts.

A "monitor corpus" is one that increases in size regularly in order to provide the users with the changes in language. In the words of Sinclair (1987:21), it is "a dynamic rather than a static phenomenon, consisting of very large amounts of electronically-held texts which will pass through the computer". An example of this type is COCA (Corpus of Contemporary American English).

Apart from the types of corpora listed above, there are other types in terms of mode, language and time.

Corpora can be classified according to mode, which deals with the way in which we want to recollect the information and there are two types: "written corpora" and "spoken corpora". Atkins et al. (1992) point out that whether mode is not specified, it will be written by default.

According to time, corpora may be "synchronic" and "diachronic". It is a synchronic corpus when the purpose of the study focuses on one particular period of time whereas it is diachronic when its aim is related with the development of time.

Other factor involved in the classification of corpora is language: monolingual, bilingual or multilingual. A "monolingual corpus" refers to the type of corpus that includes only one language whereas "bilingual corpus" deals with two languages. A "multilingual corpus" comprises at least three languages.

In translation studies, "comparable" and "parallel corpora" are often the most useful tool for translators. "Parallel corpora" are used as a bilingual dictionary because the translator can identify terms and phrases with their equivalences into another language. And although, comparable corpora contain only original texts, they can also be used by translators to look for equivalent terms and phraseology and also by those who are interested in the topic that they deal with.

It is not an easy assignment to design "parallel corpora" since it is difficult to find the translations of source texts to compile a good parallel corpus.

As for the corpus designed and compiled for this project, it is monolingual. The aim is showing that the usage of this corpus can help translators in their translation assignment.

4. Corpus compilation and analysis

4.1. Corpus design and compilation

Once the translators decide the type of corpora that they will create, they must consider different criteria for their corpus. These criteria are important because they may influence the results of their analysis. According to Bowker and Pearson (2002: 45-52), the criteria are size, medium, subject, text type, authorship, language and date of publication.

Size is one of factors that the translators must take into account. Sinclair (1991:18) argues that it must be "as large as possible, and should keep on growing" while Bowker and Pearson (2002:45) point out that there are no hard and fast rules that can be followed to determine the ideal size of a corpus. And Biber (1993) says that size depends on the purpose for which the corpus is intended. Therefore, the translators are the one who have to determine how big their corpus is going to be, and their decision will be based mainly on the aim of their study.

The size of the corpus of this project is 231, 989 words. It is not big if we compare it to a general corpus. Obviously, there are reasons that justify this difference. First, we found some difficulties to collect the samples since most the texts are not available on the Web. And second, this type of corpora is usually smaller than general corpora since it analyzes a particular aspect of language. Although the size of the corpus is small, it includes a reasonable number of texts in order to achieve the expected results. For example, if the aim of the translators is to find synonyms of specified terms and the size of their corpus is too small, this would be a problem in their analysis.

However, the size of a corpus is not the only relevant aspect but also the **size of the texts**. As well as the size of the corpus is based on the purpose of the study, the size of the texts also depends on this. If the translators only select chunk texts, they then delete part of the information which may be important for their study. For this reason, full texts are the best option rather than chunk texts. We would also like to mention the question of the **copyright**

in building a corpus. If the compiler is going to include whole texts he/she should ask permission to the editorial to use those texts for research or translation purposes. In this project we are just using the four typical sections of a medical research paper to compile the corpus: Introduction, Materials and Methods, Results and Discussion (IMRAD superstructure). This means that we are omitting the References.

The total **number of texts** is another criterion related to the size of the corpus. Whether the translators are interested in identifying the way in which the authors use specialized language, it would be advisable that he will include a great number of texts from a variety of authors in their corpus. This way, they will get information about what terms and concepts are typically used in a specific subject field. The texts included in our corpus of the project are written by different authors since the purpose is to identify terminology and grammatical structures commonly used by experts.

It is also important to determine whether the corpus will contain only **written material** or **spoken material**, or both. It also depends on the study that the translators will do. The medium chosen for the corpus that we built is written material and it is due to the main two reasons: first, the purpose is to identify terms and grammatical structures that experts use in scientific texts and second, it is much easier to collect this type of material than the other option.

It is said that all the texts included in a corpus have to deal with the specialized **subject** that the user will analyze. However, Bowker and Pearson (2002:50) point out that if your project sets out to study particular features of specialized language, it may not be necessary for all your texts to be about the same subject. Our purpose is to analyze the specialized subject of Oncology and thus all our texts deal with this subject.

It is also necessary to include texts in the corpus that belong to the same **text type**. The problem lies in the fact that the texts are written in the different style. For example, if we read a text from a newspaper we notice that the language style is totally different from a

text written in a magazine. As this study is based on a specific aspect of language, it was necessary to compile a corpus that included texts that share the same characteristics in their structure. The samples of language for the corpus are research articles, written by experts and addressed to experts.

Moreover, **language** is an issue that the translators take into account when they are compiling their corpus. The texts that they are going to include in their corpus must be original texts and not translations. The reason why original texts are the best option is that they contain expressions written by the experts in the source language. This information is essential for the translators, who aim to produce a translation that be read as an original text written in the target language.

And finally, **date of publication** of the texts also depends on the aim of the study. The translators are thus who determine whether they want to include up-to-date texts in their corpus or not. As stated above, the purpose of creating a corpus is to help the translators to be faced with problems that arise when they are translating, and we consider that up- to-date texts are essential since terms become obsolete in scientific fields.

4.1.1. Selection of suitable material

The translator is required to include appropriate information in his corpus to obtain a helpful analysis. No doubt that one of the easiest ways of searching for material is on the Internet since it contains an amount of information on almost any subject. However, this information given is not always suitable.

Bowker and Pearson (2002:61) state two main tools that the user can use for searching his texts and are: **search engines** and **subject directories**. The former are tools that allow them to enter the search term on the Web, for example, Google, Bing, and Yahoo search; whereas the latter are web-sites organized by people, who classify them according to subject field.

We consider that subject directories are better than search engines because the former contain a selective list according to the field that you are interested. For the selection of the texts of our corpus, we used the database of the University of Valladolid's library because thanks to the information retrieved we could make sure the quality of the texts.

This database has a list of several sections such as Science and Technology, Legal Sciences, ProQuest Database, Journal Citations Report, etc. The section chosen was Journal Citations Report, which gives the users two types of selection: by categories and by journal or category title. Due to the fact that we did not know the titles of journals, we chose the first option by selecting the field of oncology. After our selection, we were given a detailed list indicating the main journals frequently read in the field of oncology. We selected the journals paying attention to their impact factor and to their total cites. We also took into account those that came from USA because our concern was to design a corpus that only included American English texts since the main oncology journals are North American.

The journals were: *Neoplasia*, *Cancer Journal*, *Oncotarget*, *Lancet Oncology*, and *Radiation Oncology*. Once the journals were selected, the next step was to download their articles from the Internet. The number of articles collected was limited by copyright restrictions in the case of *Oncotarget*, *Cancer Journal*, *Lancet Oncology* and *Radiation Oncology*. However, *Neoplasia* was the only journal in which their articles were immediately downloaded thanks to its accessibility.

In the image below there is a table that illustrates the medical journals we used for our corpus and the number of articles collected from each journal:

Medical journals	**Number of texts**
Neoplasia	17
Oncotarget	9
Cancer Journal	7
Lancet Oncology	6
Radiation Oncology	4

Image 1: Medical Journals

The corpus thus involved 43 texts collected from 5 medical journals.

Moreover, most the texts downloaded from the online journals had PDF format and it was necessary to copy them into a word processor. Subsequently, they were saved as a plain text file due to the fact it is the only file format type that AntConc program allows us to use for the analysis of the corpus.

The following image shows the main characteristic of the **monolingual corpus**:

Type of corpus	Specialized
Size	231.989 words
Number of texts	43 texts written by different authors
Medium	Written
Subject	Oncology
Authorship	Texts written by experts
Language	Texts written in American English by experts
Publication date	Texts from 2010 to 2014

Image 2: Characteristics of our corpus

4.1.2. Selection of computer software

As we already know, a corpus can be used as a valuable aid for translators; the question is how to obtain the information from corpora. Once the translators have determined the criteria for the design of their corpus, they now have to decide the computer software that they will use.

There are different software program packages available on Internet, for example: **WordSmith Tools** and **AntConc program**. The former is developed by Mike Scott and the latter by Laurence Anthony. They both provide their users with similar tools to work as researchers of language. However, they differ from each other because the former is not totally free; you just have an online demo version. AntConc is a program that does not require any payment and is downloaded easily. This program offers seven tools, such as: *Word List, Keyword list, Collocates, Clusters, Concordance, Concordance Plot* and *File*

View. Each of them has menu preferences available. The user can use them in a simple way thanks to the tutorials and information that the program provides[1]. These tools help the user with the analysis of his corpus.

For all these reasons we thought that AntConc was the appropriate software package chosen for the project.

[1] More information is available at http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3.2.4.pdf (accessed date 09/03/2014).

4.2. Corpus analysis

In this section, we discuss the procedure involved in the analysis of our specialized monolingual corpora. For the analysis, we have used the following tools:

- Word List
- Word Clusters
- Concordances

The first step was to identify the terms that frequently occur in the specialized subject of Oncology. Our goal was to use the terms as "the search node" (Pearson 1998: 191) in order to observe their usage in context. The question that immediately arose was how to extract them. On the Web, there are online tools called **term extractors** that may help the translator with this task. Some of them are free and do not require installation. However, the results that these term extractors provide are not totally suitable. Their disadvantage is the high percentage of noise and silence. We can confirm this drawback because we used online extractors, such as: "Termine" and "Maui-indexer" in our Applied Linguistics classes (Pizarro Sánchez 2013-14). We learned that some of these term extractors presented these previous problems. Bowker and Pearson (2002: 169) state that "noise refers to unwanted items that are erroneously retrieved (i.e. patterns that are *not* terms), while silence refers to cases where patterns that *are* terms do not get retrieved".

This question may be solved easily if the user knows the usages of corpora, since nowadays they can be used as a **term extraction tool**.

4.2.1. Term extraction

We are going to explain in this section how the translator can extract terms from the corpus. Terms are not extracted automatically from the corpus, this process is **semi-automatic**. After the extraction, some candidate terms may not be terms in the end so it is necessary that these term candidates are verified by the user. We speak about term candidates, when

we refer to those "words or phrases that appear to be terms" (Bowker and Pearson 2002:145).

As we already know, a corpus contains thousands of words; this means that it contains an extensive amount of information. However, it may hinder our process of identifying term candidates since it may be a more complex and time-consuming process. A good way of reducing the information of the corpus is to use a stop list. It is "a list of items that you would like the computer to ignore" (Pearson 2002:172). In other words, it is a list that includes those words that the user does not want to appear in the frequency lists, such as prepositions, articles, or conjunctions. So we decided to use a stop list[2] to delete those words that we considered useless to extract.

In specialized language, a term may be a single- word (e.g. 'cell') or multiple words (e.g. 'cancer stem cell'). Pearson (1998: 125) states that "single word terms are the most common, followed by two word noun + noun and adj. + noun combinations". We decided that our analysis is based on her approach. Therefore, we selected the nouns as single candidate terms and noun + noun and adjective + noun combination as multiword term candidates. They were extracted through the tools that were mentioned before.

Pearson (1998:123) points out that "frequency is an important criterion for assessing the eligibility of a term candidate". It is thus advisable that the user establishes the minimum number of times that a lexical unit may occur in the corpus in order for it to be considered as a candidate term. However, Pearson (1998:123) also argues that "low frequency should not prelude a term candidate from being considered". The problem lies in the fact that some terms can be used by their abbreviated form. On the other hand, acronyms are commonly used in medical texts. For that reason, the user is required to pay attention to both aspects.

[2] The stop list was obtained from http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop (accessed date 13/03/2014)

The easiest way to identify single - word term candidates is by using a word list, which shows the terms with their frequency of occurrence. It counts how many times each lexical unit appears in the corpus.

According to what is said before, we established that the minimum number of occurrences to be 8 and we used a concordance tool to find those terms that could be omitted. These terms could not be extracted because they are used by their abbreviated form rather than by their full name. To find those terms, we entered as keywords “tomography”, “imaging”, “survival” and “disease”. The following acronyms were found:

- PET (positron emission tomography)
- MRI (magnetic resonance imaging)
- OS (overall survival)
- SD (stable disease)

4.2.1.1. Simple term candidates

The following table shows term candidates that meet the minimum frequency stated above and acronyms found in the corpus that we have compiled. We extracted this list of terms using a word list software program:

SIMPLE TERM CANDIDATES	FREQUENCY
Cell	3332
Tumor	1552
Cancer	1375
Gene	635
Therapy	464
Drugs	459
Tissue	380
Line	369

Radiation	326
Protein	309
Pathway	299
Lung	282
Factor	255
Mutation	242
Melanoma	239
Imaging	235
Human	231
Chemotherapy	217
DNA	195
Apoptosis	185
Carcinoma	173
Antibody	163
Proliferation	148
Metastasis	145
Stage	145
PET	131
Toxicity	129
MRI	126
Stem	126
Surgery	123
Signals	98
Biopsy	91
Recurrence	87

Screenings	86
Docetaxel	85
Osteosarcoma	79
Blood	77
Lymphoma	74
Phenotype	69
Genome	60
Prognosis	59
Angiogenesis	57
Bone	56
Methylation	52
Adenocarcinoma	50
Marrow	49
Transplantation	49
Enzyme	48
Skin	46
Cisplatin	45
Mortality	43
Antigen	40
Oncogene	40
Staging	40
Leukemia	38
Differentiation	36
Layer	35
Locoregional	33

Biomarker	32
Translocation	31
Transfer	28
Immunohistochemistry	27
Myeloma	25
Nodule	25
Mastectomy	25
Resection	24
Brain	24
Infusion	23
Dissection	23
Chromosome	18
Necrosis	18
Carcinogenesis	18
Brachytherapy	18
Immunotherapy	17
OS	15
SD	14
Carcinogen	14
Lumpectomy	14
Prostatectomy	13
Epithelium	13
Nucleus	12
Macrophage	12

Chemoprevention	11
Colonoscopy	11
X-ray	10
Implantation	10

Image 3: Single term candidates extracted from a word list

4.2.1.2. Multiword term candidates

On the other hand, word cluster software programs enable the user to detect multiword term candidates. The selection of multiword term candidates may be based on a statistical approach (Bowker and Pearson 2002:170) which consist of establishing the minimum number of times that a series of lexical units must be repeated in the corpus. This minimum value should be indicated by the user, as well as the minimum and maximum length of lexical units in order to indicate multiword term candidates.

Our selection was based on this approach and we established that the minimum number will be 8 for the times that a series of lexical units are repeated and the minimum and maximum will be 2, 3, and 4 for the length of words. We indicated several values for the length of words in order to identify more term candidates.

The image below shows the cluster list:

MULTIWORD TERM CANDIDATES	FREQUENCY
Cell line	292
Breast cancer	205
Lung cancer	171
Radiation therapy	159
Tumor cells	156
In vivo	140
In vitro	123
Prostate cancer	111
Lymph node	108
Melanoma cell	99
Stem cell	79

Primary tumor	65
Signaling pathway	65
Clinical trials	62
Cell proliferation	53
Pancreatic cancer	50
Bone marrow	47
Cell cycle	45
Tumor tissue	42
Tumor volume	36
Drug resistance	35
Stem cell transplantation	34
Cell type	32
Risk factor	32
Cancer stem cell	31
DNA repair	31
Squamous cell carcinoma	30
DNA methylation	27
Partial response	26
Cell migration	24
Colon cancer	23
progression-free survival	22
Cell culture	21
Blood vessels	20
Survival rate	20
Local recurrence	20

Locoregional recurrence	19
Colorectal cancer	17
Multiple myeloma	16
Side effects	15
In situ	14
Non – small cell lung cancer	12
Tumor suppressor gene	10
Carcinoma in situ	9
CT scan	8

Image 4: Multiword term candidates extracted from a cluster list

There is another slower way to identify more term candidates. The user just takes the single – word candidates or multiword candidates and then uses a concordance tool. This tool allows him to identify lexical units associated to other lexical units, which may constitute a term.

4.2.1.3. Identification of key terms

As the process is semi- automatic, the user has the task to indicate which term candidates are the key terms and eliminate those lexical units that are not. **Specialized glossaries** are essential for him/her to achieve this assignment.

We searched specialized glossaries on the Web and selected those that are related with the field of oncology. They helped us to verify which term candidates were the key terms. The glossaries were found on these websites: Genomic Health (www.genomichealth.com), Cancer Monthly (www.cancermonthly.com), American Cancer Society (www.cancer.org) and American Society for Cancer Research (www.aacr.org)[3]. They are official medical centers and institutions from the US which give information about causes, treatment, prevention of cancer, etc. to people who suffer this disease or those who are interested in the topic. The information included in these sources comes from experts and it is therefore reliable. It was the main reason why they were selected as reference for indicating the key terms.

The following image shows our term candidates that occur in these previous glossaries

Term candidates	Genomic Health	Cancer Monthly	American Cancer Society	American Society for Cancer Research
Cell				√
Tumor				√
Cancer		√		√
Gene				√
Drugs				
Stem				
Human				
Tissue				
Therapy				
Line				

[3] Accessed date: 27/03/2014

Protein				
Pathway				
Lung				
Radiation				
Mutation				√
Melanoma				
Factor				
Imaging				√
Chemotherapy				√
DNA		√		
Apoptosis				√
Carcinoma		√		√
Antibody				√
Proliferation				
Metastasis			√	
Stage				√
PET			√	
Toxicity				√
MRI			√	
Surgery			√	
CT scan			√	
Signals				
Recurrence	√			√
Biopsy	√	√	√	
Docetaxel				
Lymphoma			√	
Blood				
Screenings	√			
Phenotype				√
Clone				
Bone				
Prognosis			√	
Genome			√	
Angiogenesis		√	√	√
Marrow				
Transplantation				
Methylation				
Skin				
Adenocarcinoma	√	√		
Enzyme			√	
Cisplatin				
Antigen	√	√	√	√
Oncogene				√
Staging	√	√	√	

Differentiation		√		
Layer				
Locoregional				
Biomarker		√		
Translocation			√	√
Mortality			√	
Transfer				
Brain				
Immunohistochemistry			√	
Myeloma				
Nodule			√	
Mastectomy		√	√	
Chromosome		√		√
Infusion		√		
Necrosis		√	√	
Carcinogenesis		√		
Brachytherapy		√		
OS				√
SD		√		
Carcinogen			√	√
Prostatectomy			√	
Epithelium				√
Immunotherapy			√	√
Dissection			√	
Nucleus			√	√
Macrophage			√	
Lumpectomy			√	
Chemoprevention			√	√
Colonoscopy	√	√	√	
X-ray			√	
Resection	√		√	
Implantation				
Cell line				
Tumor cells				
Lung cancer				
In vivo				
Breast cancer				√
In vitro				
Prostate cancer				
Lymph node			√	
Melanoma cell				
Radiation therapy		√		
Stem cell			√	
Primary tumor		√		

ÇÇ

Signaling pathway				
Clinical trials	√	√	√	√
Cell proliferation		√		
Bone marrow		√		
Pancreatic cancer				
Cell cycle			√	√
Tumor tissue				
Multiple myeloma			√	
Tumor volume			√	
Drug resistance			√	√
Cell type				
Stem cell transplantation			√	
Tumor suppressor gene		√	√	
Cancer stem cell				√
DNA repair			√	
Squamous cell carcinoma			√	√
DNA methylation				√
Partial response		√		√
Risk factor			√	
Colon cancer		√		
Cell migration				
Progression-free survival				√
Cell culture				
Blood vessels				
Survival rate		√	√	
Locoregional recurrence				
Local recurrence	√			√
Colorectal cancer			√	√
Non – small cell lung cancer		√	√	√
Osteosarcoma		√		
Side effects			√	√
In situ	√		√	
Carcinoma in situ	√	√	√	√

Image 5: Term candidates extracted from online medical glossaries

Therefore, the term candidates that we considered key terms are “cell”, “tumor”, “cancer”, “gene”, “mutation”, “imaging”, “chemotherapy”, “DNA”, “apoptosis”, “carcinoma”, “antibody”, “metastasis”, “stage”, “PET”, “toxicity”, “MRI”, “surgery”, “recurrence”, “biopsy”, “lymphoma”, “screenings”, “phenotype”, “prognosis”, “genome”, “angiogenesis”, “adenocarcinoma”, “enzyme”, “antigen”, “oncogene”, “staging”, “leukemia”, “differentiation”, “biomarker”, “translocation”, “mortality”, “immunohistochemistry”, “nodule”, “mastectomy”, “chromosome”, “infusion”, “necrosis”, “carcinogenesis”, “brachytherapy”, “OS”, “SD”, “carcinogen”, “prostatectomy”, “epithelium”, “immunotherapy”, “dissection”, “nucleus”, “macrophage”, “lumpectomy”, “chemoprevention”, “colonoscopy”, “X-ray”, “resection”, “osteosarcoma”, “breast cancer”, “lymph node”, “radiation therapy”, “stem cell”, “primary tumor”, “clinical trials”, “cell proliferation”, “bone marrow”, “cell cycle”, “multiple myeloma”, “tumor volume”, “drug resistance”, “stem cell transplantation”, “tumor suppressor gene”, “cancer stem cell”, “DNA repair”, “squamous cell carcinoma”, “DNA methylation”, “partial response”, “risk factor”, “colon cancer”, “progression-free survival”, “survival rate”, “local recurrence”, “colorectal cancer”, “non-small cell lung cancer”, “side effects”, “in situ”, “carcinoma in situ” and “CT scan”.

4.2.2. Collocations

Olohan (2004:63) states that a concordance software program is "the most common tool for data extraction". She also points that "each instance is displayed with its immediate co-text" (2004:63). As we indicated previously, the terms comprised in image 5 will be used as the keywords. Therefore, the next step in our analysis was to use the terms as the search pattern to extract the information about their usage in context. We used a concordance software program for our analysis.

We believe that this tool has two main advantages: The first one is that it is a helpful tool for the translators because they may verify whether the chosen translation equivalent that they have found in dictionaries or glossaries is the right one or not. And second, this tool allows the users to acquire knowledge of target language (TL) such as: new expressions.
Our analysis focused on the combinations formed by **verb + noun**, **noun+ verb**, **adjective+ noun** and **noun + noun**, in which the noun is referred to as the base and the verb, the adjective and the noun as the collocates. Similarly, we established a minimum of frequency to select collocations. This means that the candidate collocates should have a have a minimum of frequency of 3 occurrences to be considered as such (Méndez Cendón, 2002).

Collocates can be defined as those words that go along with the search term, for instance:

"Mammographic **screening**"

This example was extracted entering the search term "screening".

A table shows our analysis, which is composed of three sections: the first section is called "term" in which the single-word terms and multiword terms are included, the second one comprises their collocations, and the last section is called notes. This section was created to

introduce all information that we considered necessary for the translation task, such as: synonyms, abbreviated form and compound terms.

Term	Collocation				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	↵ Abbreviated form √ Synonym * Compound terms

Image 6: Table of information regarding the terms

As shown in Image 6, we used three different symbols to distinguish the information that will be included in Notes. The first symbol is ↵ to indicate abbreviated form, the second one is √ to point synonyms and the last one is * to show compound terms.

To begin with our analysis, we first use single - word terms as the search pattern and then we obtain their KWIC concordances lines.

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Cell	Induce, affect, treat, control, promote, inhibit, decrease, cause, use, enhance, initiate, produce	Treat, use consist, produce, undergo, increase, contain, activate	Leukemic endothelial, epithelial, basal, blast, primary, apoptotic, stromal*, myeloid, cancerous, dentritic*, malignant, clonogenic, mononuclear, mesenchymal,	line, type, division, death, in vitro, invasion, culture, carcinoma, differentiation, growth, adhesion, interaction, migration, apoptosis, viability, membrane, resistance, phenotype, generation, lymphoma, sensitivity	*Bone marrow stromal cells (BMSCs) *Tumor dendritic cells (TUDCs) *Melanoma initiating cells (MIC) *Senescent melanoma cells (SSMC) *Human umbilical cord endothelial cells

					(HUVECs)

Table 1: Collocates of the search term *cell*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Tumor	Generate, form, treat reduce, cause, inhibit, promote, initiate	Arise, use produce, reach,	gastric, recurrent, circulating*, solid, colonic, colorectal, epithelial, human, aggressive, immunosuppressive, metastatic, oral, subcutaneous, pancreatic, original, orthotopic, intestinal,	Cell, vessels, recurrence, DNA, clone, tissue, mass, malignancy, angiogenesis, invasiveness, stroma, stage, grown, nodule, environment, formation, development, necrosis, resection, expansion, killing, inoculation	*Circulating tumor cells (CTCs)

Table 2: Collocates of the search term *tumor*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Cancer	Develop, prevent, identify, detect, increase, inhibit	Cause, occur, include, increase	Esophageal, epithelial, gastric, hematologic, oesophagogastric, pancreatic, rectal, ovarian, invasive, human, malignant,	Cell, tissue, resistance, treatment, death, tumor, therapy, development, screening, growth, detection, xenograft, progression,	

				mortality	

Table 3: Collocates of the search term *cancer*

Term	**Collocations**				**Notes**
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Gene	Use, identify,	Contain	Circadian, housekeeping	Mutation, fusion, ontology*, expression*, regulation	*Gene ontology (GO) *Gene expression profiling (GEP)

Table 4: Collocates of the search term *gene*

Term	**Collocations**				**Notes**
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Mutation	Identify, active,	Identify, occur	Somatic, oncogenic, intronic, genetic	Detection	

Table 5: Collocates of the search term *mutation*

Term	**Collocations**				**Notes**
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Imaging	Use	Play, detect, perform	Spectroscopic, photoacoustic, multiparametric, intravital, optical, molecular, morphologic	Test, agent, modality	

Table 6: Collocates of the search term *imaging*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Chemotherapy	Reduce, resist,	Provide	Neoadjuvant, neurotoxic	Treatment, drug, agent, resistance	

Table 7: Collocates of the search term *chemotherapy*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
DNA	use, increase,		Genomic	Hypermethylation, sample, damage, fragmentation microarray	

Table 8: Collocates of the search term *DNA*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Apoptosis	Induce, trigger, mediate		Induced, endothetial	Assay	

Table 9: Collocates of the search term *apoptosis*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Carcinoma			lobular, gastric ovarian, invasive, mammary, hepatocellular,	Cell	*Invasive ductal carcinoma (IDC) *Basal-cell carcinoma

Table 10: Collocates of the search term *carcinoma*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Antibody	Control, use	Use	Monoclonal, polyclonal, secondary primary		

Table 11: Collocates of the search term *antibody*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Metastasis					

Table 12: Collocates of the search term *metastasis*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Stage			Clinical, early, unresectable, advanced	Melanoma, disease, III, IV, I	

Table 13: Collocates of the search term *stage*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
PET			Interim, early	Scan, imaging, finding	√Positron emission tomography

Table 14: Collocates of the search term *PET*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Toxicity	Reduce, minimize		Cardiac, systemic, haematological		

Table 15: Collocates of the search term *toxicity*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
MRI	Undergo, use	Use	Enhanced	Screening	√Magnetic resonance imaging

Table 16: Collocates of the search term *MRI*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Surgery	Undergo		Excisional		*breast-conserving surgery

Table 17: Collocates of the search term *surgery*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Recurrence			Local, biochemical, locoregional,		

Table 18: Collocates of the search term *recurrence*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Biopsy	Use	Yield, control	Human	specimens	

Table 19: Collocates of the search term *biopsy*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Lymphoma			Anaplastic, lymphocytic	Cell	

Table 20: Collocates of the search term *lymphoma*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Screenings			Mammographic	Mammography	

Table 21: Collocates of the search term *screenings*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Phenotype	Rescue		Tumorigenic, invasive, antiapoptotic, stemness, mesenchymal, epithelial,		

Table 22: Collocates of the search term *phenotype*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Prognosis	Confer		Dismal		

Table 23: Collocates of the search term *prognosis*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Genome	Perform, use		Human, whole*		*Whole genome sequencing (WES)

Table 24: Collocates of the search term *genome*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Angiogenesis	Induce, inhibit				

Table 25: Collocates of the search term *angiogenesis*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Adenocarcinoma			Pancreatic, ductal, gastric, oesophagogastric, gastrooesophageal,		*Pancreatic ductal adenocarcinoma (PDA) *Adenocarcinoma in situ (AIS)

Table 26: Collocates of the search term *adenocarcinoma*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Enzyme					

Table 27: Collocates of the search term *enzyme*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Antigen	Make		specific*	Retrieval	*Prostate – specific antigen (PSA)

Table 28: Collocates of the search term *antigen*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Oncogene					

Table 29: Collocates of the search term *oncogene*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Staging			Accurate		

Table 30: Collocates of the search term *staging*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Leukemia			Lymphoblastic*, myeloid,		*Chronic lymphoblastic leukemia (CLL) *Acute lymphoblastic leukemia (ALL)

Table 31: Collocates of the search term *leukemia*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Differentiation			Adipogenic		

Table 32: Collocates of the search term *differentiation*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Biomarker					

Table 33: Collocates of the search term *biomarker*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Translocation			Nuclear, chromosomal		

Table 34: Collocates of the search term *translocation*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Mortality					

Table 35: Collocates of the search term *mortality*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Immunohistochemistry					↵IHC

Table 36: Collocates of the search term *immunohistochemistry*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Nodule					

Table 37: Collocates of the search term *nodule*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Mastectomy	Undergo				

Table 38: Collocates of the search term *mastectomy*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Chromosome					

Table 39: Collocates of the search term *chromosome*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Infusion			intravenous		

Table 40: Collocates of the search term *infusion*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Necrosis					

Table 41: Collocates of the search term *necrosis*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Carcinogenesis					

Table 42: Collocates of the search term *carcinogenesis*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Brachytherapy					

Table 43: Collocates of the search term *brachytherapy*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
OS					√Overall survival

Table 44: Collocates of the search term *OS*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
SD					√Stable disease

Table 45: Collocates of the search term *SD*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Carcinogen					

Table 46: Collocates of the search term *carcinogen*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Prostatectomy			Radical		

Table 47: Collocates of the search term *prostatectomy*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Epithelium					

Table 48: Collocates of the search term *gene*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Immunotherapy					

Table 49: Collocates of the search term *immunotherapy*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Dissection	Undergo				

Table 50: Collocates of the search term *dissection*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Nucleus					

Table 51: Collocates of the search term *nucleus*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Macrophage			associated*		*Tumor-associated macrophages (TAMs)

Table 52: Collocates of the search term *macrophage*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Lumpectomy					

Table 53: Collocates of the search term *lumpectomy*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Chemoprevention					

Table 54: Collocates of the search term *chemoprevention*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Colonoscopy					

Table 55: Collocates of the search term *colonoscopy*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
X-ray					

Table 56: Collocates of the search term *X-ray*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Resection	Undergo		Surgical		

Table 57: Collocates of the search term *resection*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Osteosarcoma	Isolate		Human, primary	Cell, stem cell	

Table 58: Collocates of the search term *osteosarcoma*

The following tables illustrate that complex noun groups[4] are used as the search pattern.

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Breast cancer	Increase		invasive, human,	Cell, risk mortality, screening, xenograft,	

Table 59: Collocates of the search term *breast cancer*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Lymph node			Axillary, sentinel, regional, mammary, metastatic,	Metastasis, biopsy, dissection, invasion,	

Table 60: Collocates of the search term *lymph node*

[4] We use the term 'complex noun group' to designate not only compound terms but also those multiword combinations which can designate more than one concept in the medical field.

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Radiation therapy	Receive		Optimized, modulated, postoperative	regimen	↵RT √Radiotherapy, *Stereotactic ablative radiation therapy (SABR) *Intensity-modulated radiation therapy (IMRT)

Table 61: Collocates of the search term *radiation therapy*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Stem cell			Embryonic*, mesenchymal*	Transplant, markers	*Embryonic stem cells (ES cells) *Mesenchymal stem cells (MSCs)

Table 62: Collocates of the search term *stem cell*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Primary tumor				Cell, sample,	

Table 63: Collocates of the search term *primary tumor*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Clinical trials			Randomized, early		

Table 64: Collocates of the search term *clinical trials*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Cell proliferation	Inhibit				

Table 65: Collocates of the search term *cell proliferation*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Bone marrow			Normal, healthy	Microenvironment*, biopsy	*Bone marrow microenvironment (BMME)

Table 66: Collocates of the search term *bone marrow*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Cell cycle	Control			Regulation, distribution, progression,	

Table 67: Collocates of the search term *cell cycle*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Tumor volume		Calculate			

Table 68: Collocates of the search term *tumor volume*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Drug resistance			Mediated*,		*Environment - mediated drug resistance (EMDR)

Table 69: Collocates of the search term *drug resistance*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Stem cell transplantation	Undergo		Haemotopoietic, autologous*		*Autologous stem cell transplantation (ASCT)

Table 70: Collocates of the search term *stem cell transplantation*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Cancer stem cell					↵CSC

Table 71: Collocates of the search term *cancer stem cell*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
DNA repair	Increase		Enhanced	Efficiency	

Table 72: Collocates of the search term *DNA repair*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Squamous cell carcinoma			Oropharyngeal, oral		↵SCC *Oropharyngeal squamous cell carcinoma (OSCC) *Oral squamous cell carcinoma (OSCC)

Table 73: Collocates of the search term *squamous cell carcinoma*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
DNA methylation					

Table 74: Collocates of the search term *DNA methylation*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Partial response	Achieve				↵PR

Table 75: Collocates of the search term *partial response*

Term	Collocations				Notes
	Verb+ term	**Term+ verb**	**Adj.+ term**	**Term+ noun**	
Risk factor			Cardiac		

Table 76: Collocates of the search term *risk factor*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Colon cancer					

Table 77: Collocates of the search term *colon cancer*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Progression-free survival			Median		↲PFS

Table 78: Collocates of the search term *progression-free survival*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Local recurrence					

Table 79: Collocates of the search term *local recurrence*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Survival rate		Improve	Overall		

Table 80: Collocates of the search term *survival rate*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Colorectal cancer					↲CRC

Table 81: Collocates of the search term *colorectal cancer*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Multiple myeloma					↵MM

Table 82: Collocates of the search term *multiple myeloma*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Non-small cell lung cancer					↵NSCLC

Table 83: Collocates of the search term *non-small cell lung cancer*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Side effects			Severe		

Table 84: Collocates of the search term *side effects*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
In situ					

Table 85: Collocates of the search term *in situ*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
CT scan		Allow			

Table 86: Collocates of the search term *CT scan*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Tumor suppressor gene					

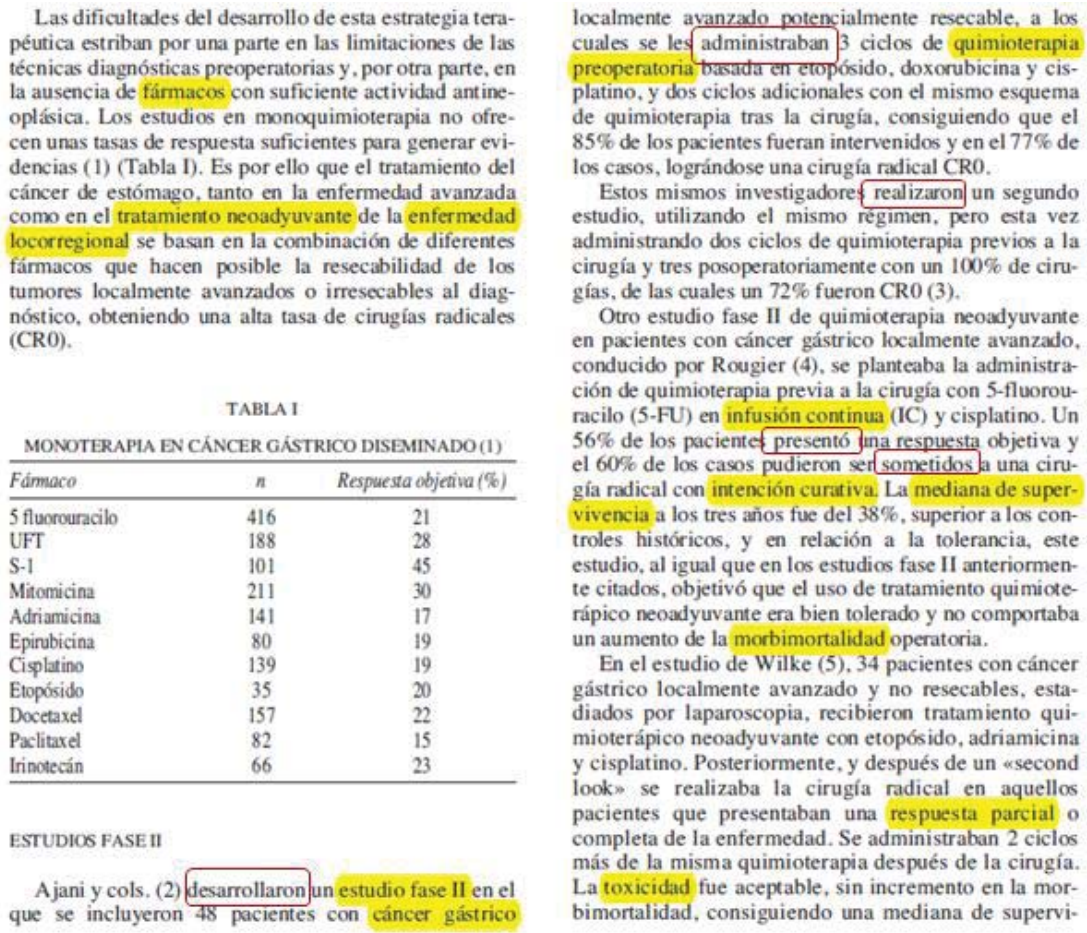
Table 87: Collocates of the search term *tumor suppressor gene*

Term	Collocations				Notes
	Verb+ term	Term+ verb	Adj.+ term	Term+ noun	
Carcinoma in situ			Ductal*, lobular		*Ductal carcinoma in situ (DCIS)

Table 88: Collocates of the search term *carcinoma in situ*

5. Specialized monolingual corpora in translation activity

In this section we are focusing on demonstrating the role that played the information extracted from KWIC concordance lines in translation activity. Therefore, the importance of compiling a specialized monolingual corpus is described. In order to indicate the usefulness of the corpus as a tool for translators, we selected a sample of the article entitled "*Tratamiento perioperatorio del cáncer gástrico"* published in the journal of *Revisiones en Cáncer* (Madrid 2013).



Las dificultades del desarrollo de esta estrategia terapéutica estriban por una parte en las limitaciones de las técnicas diagnósticas preoperatorias y, por otra parte, en la ausencia de fármacos con suficiente actividad antineoplásica. Los estudios en monoquimioterapia no ofrecen unas tasas de respuesta suficientes para generar evidencias (1) (Tabla I). Es por ello que el tratamiento del cáncer de estómago, tanto en la enfermedad avanzada como en el tratamiento neoadyuvante de la enfermedad locorregional se basan en la combinación de diferentes fármacos que hacen posible la resecabilidad de los tumores localmente avanzados o irresecables al diagnóstico, obteniendo una alta tasa de cirugías radicales (CR0).

TABLA I

MONOTERAPIA EN CÁNCER GÁSTRICO DISEMINADO (1)

Fármaco	*n*	*Respuesta objetiva (%)*
5 fluorouracilo	416	21
UFT	188	28
S-1	101	45
Mitomicina	211	30
Adriamicina	141	17
Epirubicina	80	19
Cisplatino	139	19
Etopósido	35	20
Docetaxel	157	22
Paclitaxel	82	15
Irinotecán	66	23

ESTUDIOS FASE II

Ajani y cols. (2) desarrollaron un estudio fase II en el que se incluyeron 48 pacientes con cáncer gástrico localmente avanzado potencialmente resecable, a los cuales se les administraban 3 ciclos de quimioterapia preoperatoria basada en etopósido, doxorubicina y cisplatino, y dos ciclos adicionales con el mismo esquema de quimioterapia tras la cirugía, consiguiendo que el 85% de los pacientes fueran intervenidos y en el 77% de los casos, lográndose una cirugía radical CR0.

Estos mismos investigadores realizaron un segundo estudio, utilizando el mismo régimen, pero esta vez administrando dos ciclos de quimioterapia previos a la cirugía y tres posoperatoriamente con un 100% de cirugías, de las cuales un 72% fueron CR0 (3).

Otro estudio fase II de quimioterapia neoadyuvante en pacientes con cáncer gástrico localmente avanzado, conducido por Rougier (4), se planteaba la administración de quimioterapia previa a la cirugía con 5-fluorouracilo (5-FU) en infusión continua (IC) y cisplatino. Un 56% de los pacientes presentó una respuesta objetiva y el 60% de los casos pudieron ser sometidos a una cirugía radical con intención curativa. La mediana de supervivencia a los tres años fue del 38%, superior a los controles históricos, y en relación a la tolerancia, este estudio, al igual que en los estudios fase II anteriormente citados, objetivó que el uso de tratamiento quimioterápico neoadyuvante era bien tolerado y no comportaba un aumento de la morbimortalidad operatoria.

En el estudio de Wilke (5), 34 pacientes con cáncer gástrico localmente avanzado y no resecables, estadiados por laparoscopia, recibieron tratamiento quimioterápico neoadyuvante con etopósido, adriamicina y cisplatino. Posteriormente, y después de un «second look» se realizaba la cirugía radical en aquellos pacientes que presentaban una respuesta parcial o completa de la enfermedad. Se administraban 2 ciclos más de la misma quimioterapia después de la cirugía. La toxicidad fue aceptable, sin incremento en la morbimortalidad, consiguiendo una mediana de supervi-

Image 7: Spanish article

As shown in Image 2, we selected those words and verbs that may present a problem for the translator. The words are highlighted in yellow while the verbs are underlined in red color.

Once we chose what we considered translation difficulties, our purpose was to reach a solution. Some words underlined are terms that we had already found in our previous analysis. It is the case of "respuesta parcial" (partial response), "toxicidad" (toxicity), "fármacos" (drugs), "infusión" (infusion), "quimioterapia" (chemotherapy) and "cáncer" (cancer).

However, our task was to translate "infusión continua," "quimioterapia preoperatoria" and "cáncer gástrico." What we did was to analyze their collocations on the right to find the appropriate adjective. The adjective "gastric" was found for cancer and therefore we considered that the translation equivalence was "gastric cancer".

But we did not find any adjective for "preoperatoria" and "continua". We trusted our intuition since we had studied these two adjectives in the subject "Specialized Translation". Therefore, our guess was that the equivalences might be "preoperative" and "continuous". We extracted concordance lines to confirm our hypothesis.

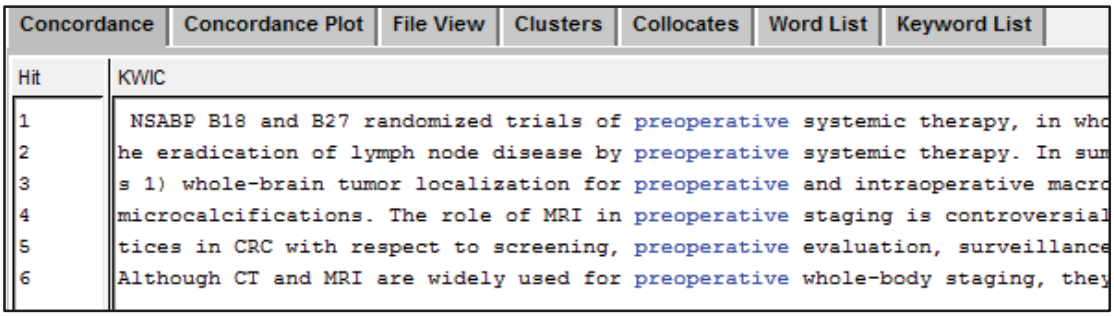


Image 8: Concordance for the search term *preoperative*

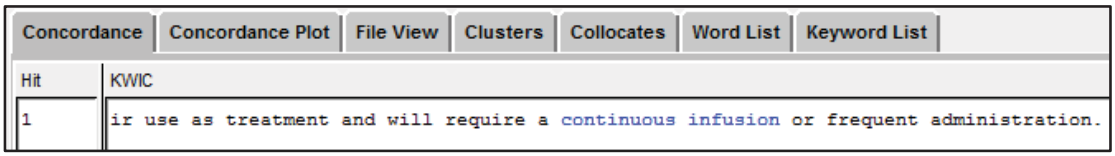
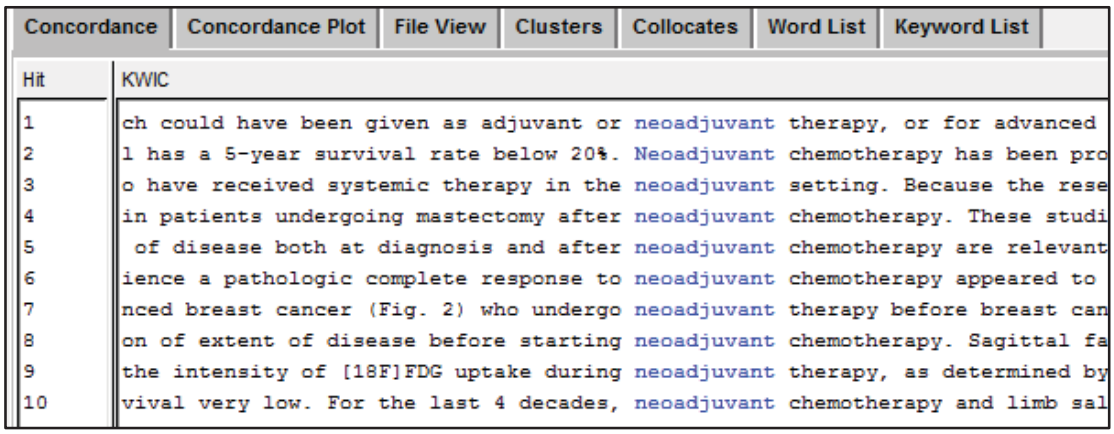


Image 9: Concordance for the search term *continuous infusion*

The images above illustrate that these adjectives are used in research articles. Therefore, "preoperative chemotherapy" and "continuous infusion" were the right equivalents. Now,

our task was to translate “tratamiento neoadjuvante”. Our intuition told us that its equivalence could be “neoadjuvant treatment”. In order make sure that it was right, we extracted concordance lines with “neoadjuvant” as the keyword but there were no results.

As we can observe the image below, this compound noun in English is “neoadjuvant therapy”.

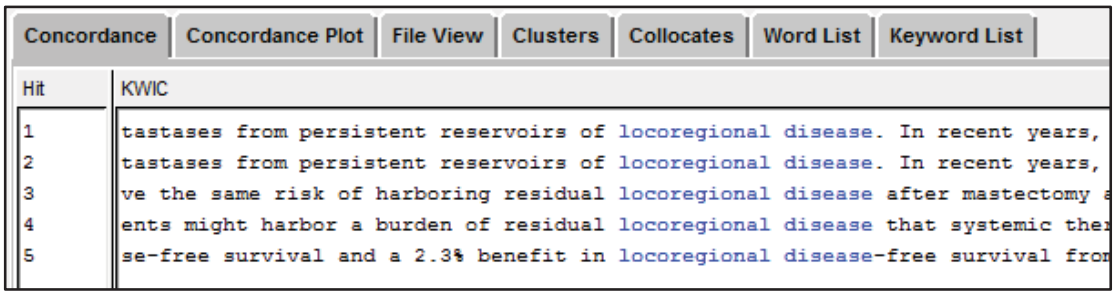


Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC
1	ch could have been given as adjuvant or neoadjuvant therapy, or for advanced
2	l has a 5-year survival rate below 20%. Neoadjuvant chemotherapy has been pro
3	o have received systemic therapy in the neoadjuvant setting. Because the rese
4	in patients undergoing mastectomy after neoadjuvant chemotherapy. These studi
5	of disease both at diagnosis and after neoadjuvant chemotherapy are relevant
6	ience a pathologic complete response to neoadjuvant chemotherapy appeared to
7	nced breast cancer (Fig. 2) who undergo neoadjuvant therapy before breast can
8	on of extent of disease before starting neoadjuvant chemotherapy. Sagittal fa
9	the intensity of [18F]FDG uptake during neoadjuvant therapy, as determined by
10	vival very low. For the last 4 decades, neoadjuvant chemotherapy and limb sal

Image 10: Concordance for the search term *neoadjuvant*

Moreover, “enfermedad” may be a problem for the translators since it can be translated into English as “illness, “sickness” or “disease.” In order to avoid mistakes, the KWIC concordance showed that “disease” had a higher frequency of occurrence than the others. Thus, it is term used by experts in research articles. The following image shows that “locoregional disease is the appropriate equivalent term for “enfermedad locorregional”.



Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC
1	tastases from persistent reservoirs of locoregional disease. In recent years,
2	tastases from persistent reservoirs of locoregional disease. In recent years,
3	ve the same risk of harboring residual locoregional disease after mastectomy a
4	ents might harbor a burden of residual locoregional disease that systemic ther
5	se-free survival and a 2.3% benefit in locoregional disease-free survival from

Image 11: Concordance for the search term *locoregional disease*

As for, “intención” it has two English equivalents: “intent” or “intention”. The task was to translate “intención curativa” but we did not know what equivalent to choose. We thought of two possibilities of translation and used them to generate concordances. “Curative intention” was the first option but we did not find results, whereas an occurrence was found for “curative intent as the image below shows.

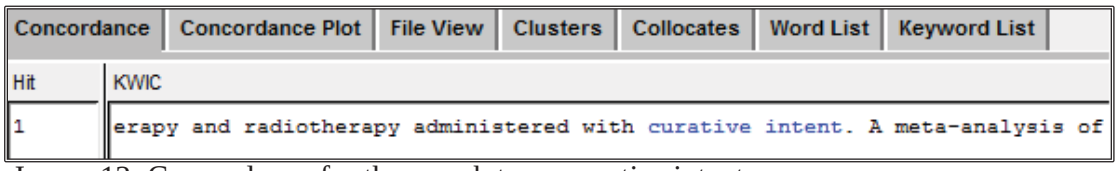


Image 12: Concordance for the search term *curative intent*

The need of knowing the translation equivalents into English leads the translators to consult a glossary or a dictionary. For that reason, we used Linguee[5] that is a tool consisting of a dictionary and a search engine. As it gives examples of usage, we think that the translators may consider it essential for their translation task.

We entered the search term “estudio fase II” and as the image 2 illustrates, it offers two possible equivalents: “phase II trial” and “study phase II”. Our doubt was whether the right equivalent was for a translation. Therefore, Linguee did not help us to solve our problems. So, we had to generate concordances with the possible translation equivalents until we finally found the right one. After this process, we observed that the right equivalent is “phase II trial”.

[5] http://www.linguee.es (accessed date 24/04/2014)

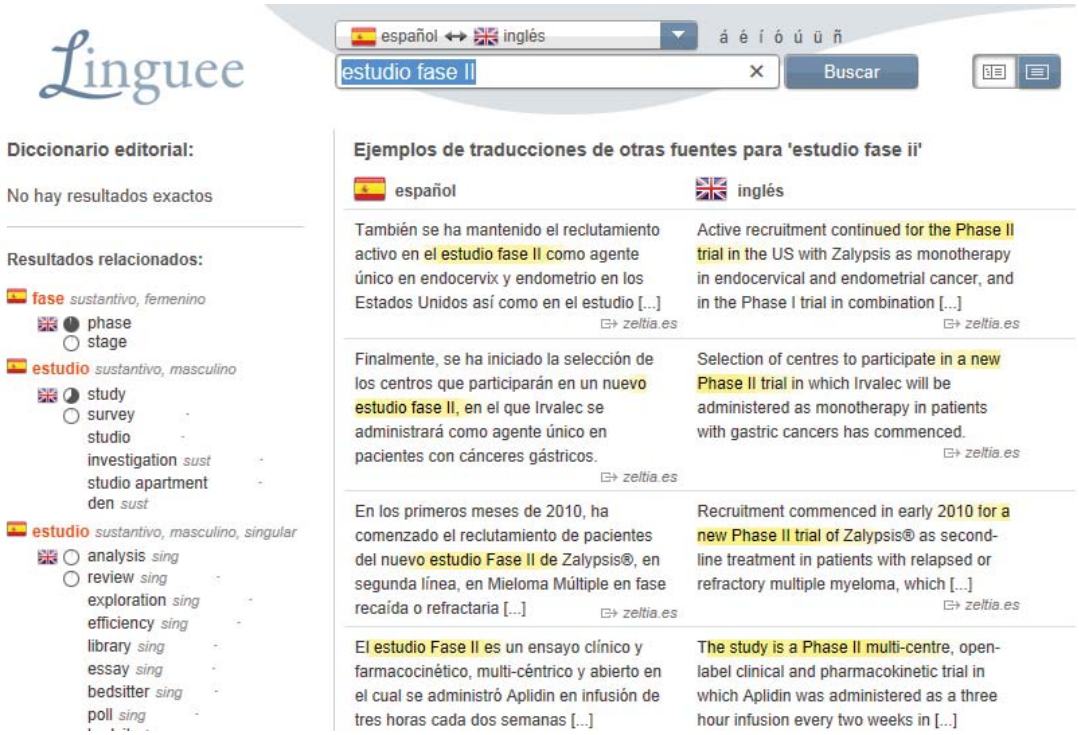
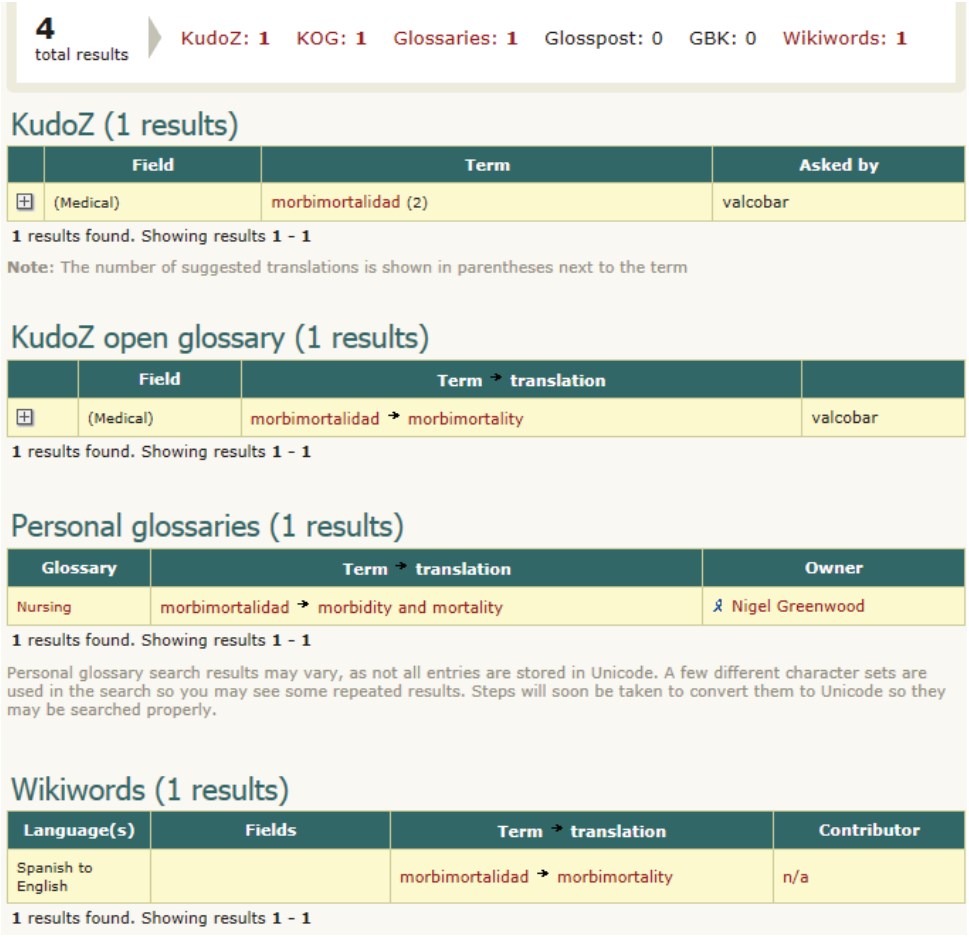


Image 13: Results for the search estudio fase II

Another online translation aid tool is the Proz glossary[6]. Our search term was "morbimortalidad". As we can see in the image 13 it offers two possibilities but we did not know which equivalent should be used. Then, we decided to extract KWIC concordance lines for both "morbidity and mortality" and "morbimortality". There were two results for the first possibility whereas there were no results for the second one. Thus, our conclusion was that the translator should use "morbidity and mortality" in his translation.

[6] http://www.proz.com/search/ (accessed date 25/04/2014)



4 total results | KudoZ: 1 | KOG: 1 | Glossaries: 1 | Glosspost: 0 | GBK: 0 | Wikiwords: 1

KudoZ (1 results)

	Field	Term	Asked by
⊞	(Medical)	morbimortalidad (2)	valcobar

1 results found. Showing results 1 - 1

Note: The number of suggested translations is shown in parentheses next to the term

KudoZ open glossary (1 results)

	Field	Term → translation	
⊞	(Medical)	morbimortalidad → morbimortality	valcobar

1 results found. Showing results 1 - 1

Personal glossaries (1 results)

Glossary	Term → translation	Owner
Nursing	morbimortalidad → morbidity and mortality	Nigel Greenwood

1 results found. Showing results 1 - 1

Personal glossary search results may vary, as not all entries are stored in Unicode. A few different character sets are used in the search so you may see some repeated results. Steps will soon be taken to convert them to Unicode so they may be searched properly.

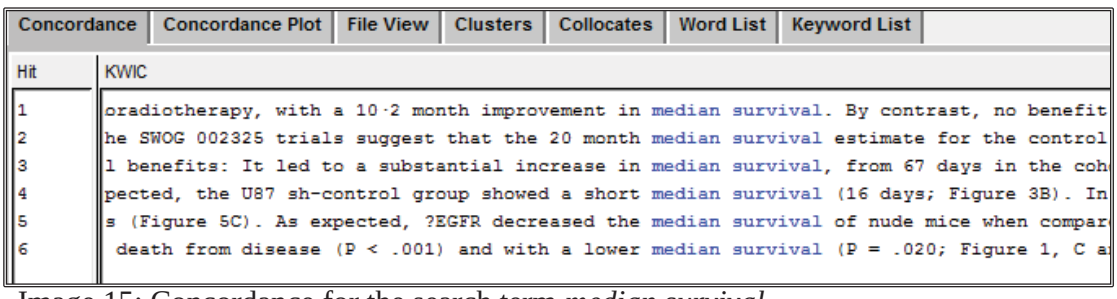
Wikiwords (1 results)

Language(s)	Fields	Term → translation	Contributor
Spanish to English		morbimortalidad → morbimortality	n/a

1 results found. Showing results 1 - 1

Image 14: Results for the search *morbimortalidad*

The last compound noun to translate was "mediana de supervivencia". In this case, we tried to make a guess again and think about "median survival" as the equivalent term. The following concordance lines shows that our guess was right.



Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC
1	oradiotherapy, with a 10·2 month improvement in median survival. By contrast, no benefit
2	he SWOG 002325 trials suggest that the 20 month median survival estimate for the control
3	l benefits: It led to a substantial increase in median survival, from 67 days in the coh
4	pected, the U87 sh-control group showed a short median survival (16 days; Figure 3B). In
5	s (Figure 5C). As expected, ?EGFR decreased the median survival of nude mice when compar
6	death from disease (P < .001) and with a lower median survival (P = .020; Figure 1, C a

Image 15: Concordance for the search term *median survival*

After we have solved the problems regarding terms and compound nouns previously, our task was to translate the verbs that are highlighted. However, we thought that it was not only necessary to look for their equivalents, but also to look for words that co-occur with them since the translators take into account the whole sentence in order to make their translation sound naturally.

The verbs we were going to analyze are "desarrollar", "administrar", "realizar", "presentar" and "someter". The problem lied in the fact that there were several options of translation. Furthermore, their equivalents had to be selected according to the words that co-occurred with them, as mentioned above. It was the reason why we considered that it was a good idea to extract some KWIC concordance lines, using as the search term the word associated with the verb.

First, we began with "desarrollar" and the search term was "phase". The concordance lines showed that "perform" could be used as the translation equivalent. On the other hand, we observed that "realizar" is a synonym of this verb in the Spanish text and therefore the translator may use the same translation equivalent.

As for "administrar", we made a guess and extracted KWIC concordance lines with the key term "administer" to validate this. As the image below shows, it is the right verb.

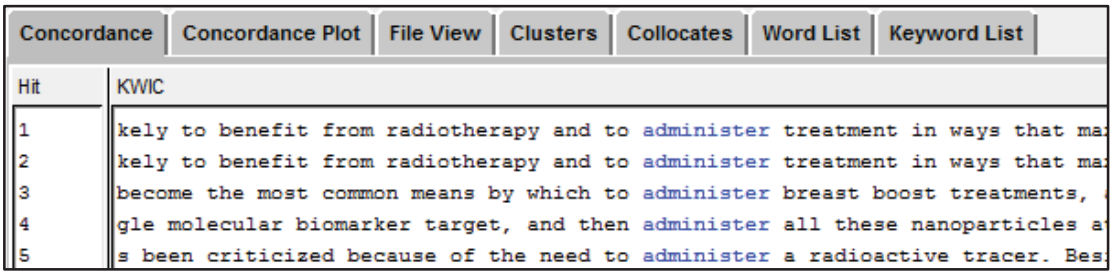


Image 16: Concordance for the search term *administer*

The next verb we analyzed was "presentar" and the search term was "partial response". We observed that "achieve" was the appropriate verb since occurred together with the term.

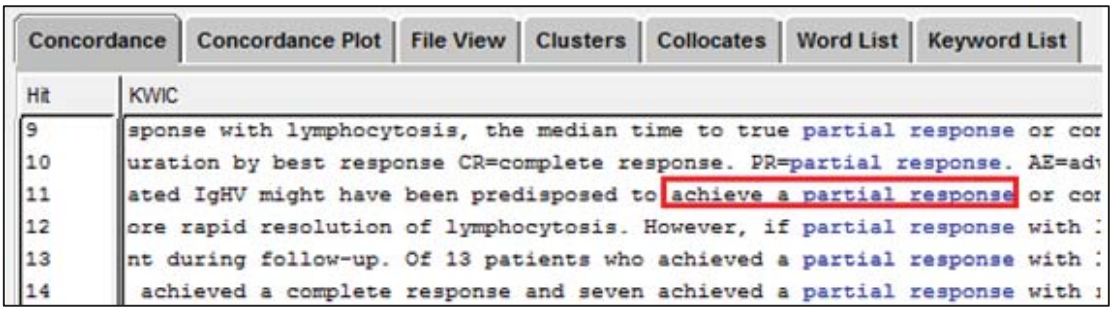


Image 17: Concordance for the search term *partial response*

Finally, "someter" was also analyzed in the corpus. We extracted KWIC concordance lines with "surgery" as the search word and found that the translation equivalent was"undergo". Then, more concordance lines were extracted with "undergo" in order to observe its collocations as the image below shows.

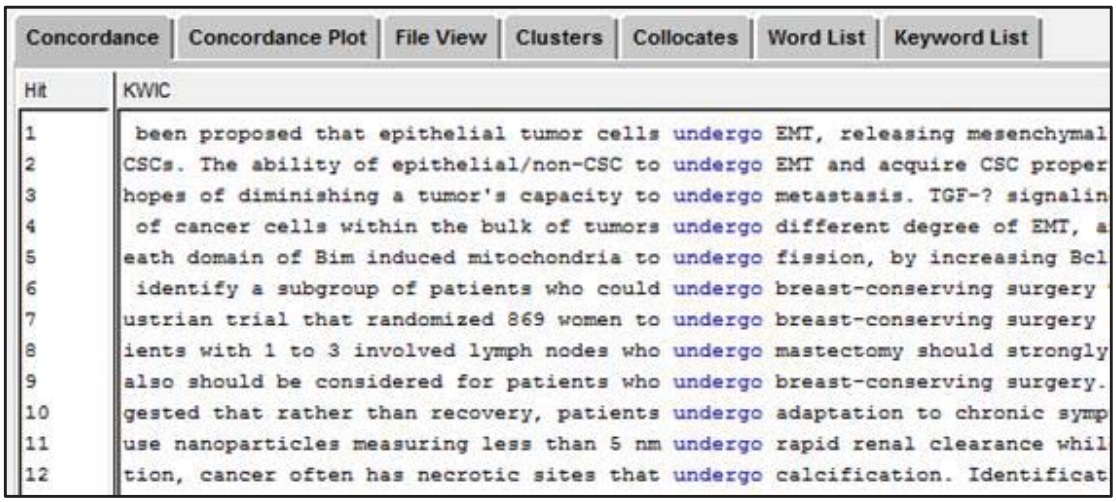


Image 18: Concordance for the search term *undergo*

6. Conclusion

Some translation students could think that second language acquisition makes them able to translate a technical or scientific text without any problem. But as we have seen at the beginning of our project, specialized language differs from global language. Therefore, they should be aware of that difference and acquire knowledge about the rules of LSP to write a specialized text.

As we have already pointed out the translators face with a set of problems when they translate a specialized text and in most cases they do not how to reach a solution despite the help of resources (e.g. dictionaries, glossaries, databases, and so on). So, this fact led us to build a specialized corpus.

It is true that many years ago, it was more difficult to compile a corpus due to the fact that it was a time-consuming process and the users were required to spend many hours collecting information in libraries. However, nowadays the translator can compile an electronic corpus rather fast. This is due to the great amount of information that is available online.

As we already know, comparable and parallel corpora are used in translation studies. But they are not so easy to compile. It is difficult to find translations of a source text to build a parallel corpus; whereas what sometimes make comparable corpus complicated to build is to meet the same characteristics for the two sub-corpora. In this project we compiled a monolingual corpus since our aim was to check its usefulness for medical translation.

The design criteria of a corpus always depend on the purpose for which you compile it. In our case, the corpus was compiled to extract linguistic information about the terms used in the IMRAD superstructure of medical research articles in order to indicate the benefits that the translators can obtain from it in case of translation problems. Our corpus consists of 45 full texts which comes to a total of 231.989 tokens. As was stated above, the Internet allows us to collect information easily. Nevertheless, not all information is reliable; consequently,

we have to be cautious choosing it. This fact was the main reason why we used the database the University of Valladolid's library to compile our corpus so that it will yield good results. We chose the journals that were listed at the top of the impact factor rank in the field of Oncology.

Regarding the texts included in our corpus, we chose to include the four typical parts of medical research paper: Introduction, Materials and Methods, Results and Discussion (IMRAD superstructure). Therefore, we did not choose chunk texts, because some important terminological and phraseological information could be excluded from our analysis. The texts deal with the same subject field (Oncology) and belong to the same text type (the research article) because we were interested in analyzing expert-to-expert texts belonging to a specific field.

It is a synchronous corpus, since the papers are from 2010 to 2014; therefore it can be considered as a representative sample of current specialized language. They are original texts, not translations. We believe that translations may be a problem for our analysis because these are not always expressed in a natural way.

In the selection of the program software, we chose AntConc mainly due to two reasons: it does not require any payment since it is free online, and it offers several tools that are necessary in corpus analysis.

The first step in our corpus analysis was the term extraction. Instead of using online term extractors, we made a decision to use our own corpus as a term extractor on the basis of our experiences by using these term extractors in our Applied Linguistics classes (Pizarro Sánchez 2013-14). This way we avoided the problems of *noise* and *silence.* Before extracting term candidates, the *stop list* was used to delete those words that we considered that were not necessary to extract. We extracted single term candidates and multiword term candidates by using the Wordlist and Cluster tools based on the approach by Pearson (1998:25). But we also used a concordance tool to extract those term candidates that are

sometimes more difficult to detect. Once term candidates were extracted, four specialized glossaries related with oncology were used to indicate the key terms. Although it was a little long process, we finally obtained the term list, which helped to solve our terminological problems.

The second step in our analysis was the collocate extraction. We have focused on lexical units (nouns) since we were interested in studying the use of noun terms in specific contexts. We have learnt in our Specialized Translation classes that there are not hardly any dictionaries and glossaries that include medical phraseological information in English (Mendéz Cendón 2012-13). This was the reason why we were interested in analyzing the collocations of nouns. As such, our analysis was focused on the patterns "verb + noun", "noun+ verb", "adjective+ noun" and "noun + noun". Apart from collocations, we also considered that it was important to extract another type of information, such as synonyms or acronyms since these have a high frequency of occurrence in specialized texts.

Once the KWIC concordance lines were generated, we noticed that the information retrieved from these lines revealed us what verbs, adjectives or nouns co-occur with the terms. Although the extraction and analysis of collocations may be a long process, the result is totally satisfactory. Therefore, the phraseological data extracted from our corpus can become "a type of documentation or source", which is, very helpful for us or for those who are dedicated to translate scientific research articles related to the topic of oncology.

A Spanish article was used in our concluding analysis because we needed to make a comparison between the usage of a specialized monolingual corpus with typical resources, such as dictionaries and glossaries in order to point out the advantages and disadvantages that can arise in the translation assignment. The same way that we look up a word in a dictionary, glossary or thesaurus, we can also look it up in a corpus. However, the results are different according to information provided. It is true that the previous resources provide you with translation equivalents and some examples of usage, in case of bilingual dictionaries, but their information is often limited. On the other hand, a corpus gives you

translation equivalents, synonyms, acronyms, phraseology, examples in several contexts and the definition of the term in some cases.
Moreover, we think that it was essential to explain possible ways of finding an equivalent in the corpus. For example, when our task was to find an equivalent for a particular verb we used lexical units that co-occur with the search term. In this regard, we consider that the translators should be creative and resourceful when they extract concordance lines, especially for those who are not familiar with corpora since translation equivalents are sometimes difficult to detect.

In addition to what has been mentioned previously, the following points illustrate the importance of building specialized monolingual corpora.

Firstly, translators specialized in scientific or technical fields usually have problems to find equivalents in dictionaries or glossaries because these do not contain information or they are outdated, whereas an up-to-date corpus can be used as term extractor, and the terminological problems will be solved easily.

Secondly, electronic corpora are very useful because you can detect and extract real and authentic examples of language, which allows you to write a translation that sounds natural, like an original text.

Thirdly, dictionaries and glossaries offer several possibilities of translation and sometimes the translator does not know what the appropriate equivalent is in the target language. On the other hand a corpus helps the specialized translator to select always the right translation equivalent. Then, there is the question of choosing the appropriate equivalent in your corpus in case that you decide to follow your intuition and you are not sure whether it is right or not.

Finally, the usage of a specialized monolingual corpus makes the translators to understand and acquire more knowledge about a particular subject field (in our case is Oncology).

As we have said, the task of the translators involves several challenges since they play the role of mediators between two languages (the source language and the target language) and cultures. Through this task, they may be faced with terminological and phraseological problems, which can affect the quality of their translation. Undoubtedly, the usage of corpora can ease their translation task because they are allowed to have access to a great amount of data. The information extracted from concordance lines helps them to solve their doubts about the meanings of words and, what is more important, to learn how these words combine with others within a sentence.

To conclude, we would like to highlight the usefulness of specialized monolingual corpora as a translation tool, by saying that they are a requisite for the translation activity. They play a significant role in the field of translation and therefore we consider that they should be integrated in the translation classes as part of the teaching curriculum.

7. References

Atkins, S. et al. (1992) “Corpus Design Criteria,” *Literary and Linguistic Computing,* 7: 1-16.

Baker, M. (1993) “Corpus Linguistics and Translation Studies: Implications and Applications”. In Baker, Mona/ Francis, Gill/ Tognini-Bonelli, Elena (eds.) *Text and Technology: In honor of John Sinclair.* Amsterdam/Philadelphia: John Benjamins, pp. 233-250.

Baker, M. (1995) “Corpora in translation studies: An Overview and Some Suggestions for Future Research,” *Target* 7, pp. 223-243.

Baker, M. (1996) “Corpus-based Translation Studies: The Challenge the Lie Ahead,” In Somers, Harold (eds.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 175 -186.

Bernardini, S. (2004) “Corpora in the Classroom: An Overview and Some Reflections on Future Developments.” In John Sinclair (ed.) *How to use corpora in language teaching*. Amsterdam: John Benjamins, pp. 16-36.

Biber, D. (1993) “Representativeness in Corpus Design”. *Literary and Linguistic Computing*, 8, pp. 243-257.

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

Bravo Gozalo, J.M. and P. Fernández Nistal (1998). “La lingüística del corpus, las nuevas tecnologías de la información y los Estudios de Traducción en la década de 1990”. In

Fernández Nistal, P. and J.M. Bravo Gozalo (eds.). *La traducción: Orientaciones Lingüísticas y Culturales.* Valladolid: Universidad de Valladolid, pp. 205-257.

De Santiago, Paula. "Introducción a la Traducción Automática." Class notes 2013-2014. Universidad de Valladolid.

Kübler, N. (2003) "Corpora and LSP translation". In Zanettin, Federico/ Bernardini, Silvia/ Stewart Dominic (eds.) *Corpora in Translator Education.* Manchester: St. Jerome Publishing, pp. 101-112.

Maia, B. (2005) "Terminology and Translation – Bringing Research and Professional Training together through Technology," *Meta: Translator's Journal,* 50. Èrudite. 22 Feb. 2014 <http://nelson.cen.umontreal.ca/revue/meta/2005/v50/n4/019921ar.pdf>

Mendéz Cendón, B. (2002) *Estrategias fraseológicas en el género discursivo de los artículos médicos en lengua inglesa.* Tesis Doctoral. Alicante: Biblioteca Virtual Miguel de Cervantes.

Mendéz Cendón, Beatriz. "Traducción de Lenguajes Especializados II (Inglés/Español: Ciencia y Técnica)." Class notes 2012-2013. Universidad de Valladolid.

Meyer, I. and K. Mackintosh (2002) "When terms move into our everyday lives: An overview of de-terminologization." *International Journal of Theoretical and Applied Issues in Specialized Communication,* 6. John Benjamins Publishing Company, pp. 111-138.

Olohan, M. (2004) *Introducing Corpora in Translation Studies.* London/New York: Routledge.

Pearson, J. (1998) *Terms in Context.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pizarro, Isabel. “Lengua Inglesa: Lingüística Aplicada III.” Class notes 2013-2014. Universidad de Valladolid.

Sinclair, J. (1987) *Looking Up*, London and Glasgow: Collins.

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

Tarp, S. (2007) “¿Qué requisitos debe cumplir un diccionario de Traducción del siglo XXI?” In Fuertes-Olivera, P.A. (ed.) *Problemas Lingüísticos en la Traducción Especializada.* Valladolid. Universidad de Valladolid, pp. 227-256.