COMPARA English Tutorial

Ana Frankenberg-Garcia

[last updated on 1 September 2008]

1. QUICK START TO THE CORPUS

Choosing your working language*

The Web interface to COMPARA is available in Portuguese and in English. Go to http://www.linguateca.pt/COMPARA/ and choose the language you feel more comfortable with. If your browser specifications bring you to the Portuguese interface and you wish to switch to English, click on This page in English on the top left-hand corner.

*This tutorial will refer to the English service alone. If you prefer to use the Portuguese interface, please follow the Aula Prática in Portuguese, available under Help.

Start using COMPARA

Click on Simple search. Write *mar* in the Portuguese to English search box and press **Search (from Portuguese to English)**.

The results you get are parallel concordances. The word *mar* can be seen in bold on the left-hand side of your screen, within the context of a larger stretch of text in Portuguese. The English equivalents are on the right. One side of the concordance is a translation of the other. To know which side is the source text and which side is the translation, check the letter&number code next to each concordance. The codes beginning with P indicate the source text is in Portuguese, and the ones beginning with E mean the source text is in English.

→ How many instances of *mar* are there?[1]

Scroll down your results to see different renderings of *mar* in English. As you can see, even though *mar* is very often equivalent to *sea*, you will also find words like *seas, seabed, maritime, ocean,* etc. on the English side of the concordance. There are also a few concordances in which there are no equivalents at all. Surprising? Not really. COMPARA is made up of human (not machine) translations, and human translators don't always translate word-for-word.

→ In what language is the source text of the first three concordances?[2]
→ In what language is the source text of the last three concordances?[3]
→ Are the texts in the concordances all the same length?[4]

A parallel concordance in COMPARA is always one full orthographic sentence of the source text and whatever matches it in the translation (remember that translators don't always preserve the sentence structure of the source text). Details on the criteria used for separating and aligning sentences in COMPARA are provided in Building COMPARA, available under Specific documentation.

Go back to the Simple search. You can do this by using your browser's back button or the Back to search form link on the top right. Change search direction by moving to the second, English to Portuguese tab on the search form, write *mar* in the corresponding search box and press **Search (from English to Portuguese)**.

---

[1] There are 221 occurrences of *mar* in the Portuguese part of COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[2] English (the letter&number code of the first three concordances begin with an E). This also means the Portuguese side of the concordances are translations.
[3] Portuguese (the letter&number code of the last three concordances begin with a P). This also means the English side of the concordances are translations.
[4] No, some concordances are very long, and some are very short.

→ How many occurrences of *mar* are there?[5]
→ Why are there so few?[6]
→ If you want to look up a Portuguese word, which search tab should you use?[7]
→ If you want to look up an English word, which search tab should you use?[8]

Go back to the Simple search and write *that* in the English to Portuguese search box and press **Search (from English to Portuguese)**.

→ How many occurrences of *that* are there in the corpus?[9]
→ How many concordances for *that* are shown?[10]

Searches that work

Query language

Go back to the Simple search and, using the Portuguese to English search tab, look up *MAR* (with capital letters) and then *Mar* (with a capital *M*).

→ How many occurrences of *MAR* are there?[11]
→ How different are the results for *Mar* and *mar*?[12]

COMPARA is case-sensitive by default. If you want your search to be case-insensitive, leave the case-insensitive option checked. This will retrieve *mar, Mar* and *MAR* all in one go.

Go back to the Simple search and click on Help to obtain more information on the query language used with COMPARA. To practise using it, try and look up the following in a single query:
→ *deus* and *Deus*[13]
→ *sit up*[14]

When your query involves a sequence of words, each separate word in the sequence must be surrounded by double-quotes: “like” “this”. The double-quotes are optional if your query is made of only one word.

→ *aberto* and *aberta*[15]
→ *book* and *books*[16]

---

[5] There are only 2 instances of *mar* in the English part of COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[6] This search box only looks for words on the English side of the corpus, and the English verb *mar* is much rarer than the Portuguese noun *mar*.
[7] The first tab (1. From Portuguese to English) should be used for queries which start in Portuguese.
[8] The second tab (2. From English to Portuguese) should be used for queries which start in English.
[9] There were 17954 hits for *that* in the English part of COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[10] Although there were 17954 hits, only 1000 concordances were shown. Words that are very common produce a very large number of hits, and to protect the rights of the copyholders of the texts in the corpus, only a limited, random selection of concordances can be shown.
[11] There were no occurrences of *MAR* in the Portuguese part of COMPARA 10.0.3. If you are using a later version of the corpus, this may have changed.
[12] Looking up *mar* retrieves common nouns, whereas *Mar* retrieves proper nouns (or *Mar* in sentence-initial position).
[13] Write **"deus"** in the Portuguese to English search box leaving the case-insensitive box checked or, alternatively, leave the case-insensitive box unchecked and write **"deus" %c**.
[14] Write **"sit" "up"** in the English to Portuguese search box.
[15] Write **"abert[oa]"** in the Portuguese to English search box.
[16] Write **"book(s)?"** in the English to Portuguese search box.

→ *perfeito, perfeita*, *perfeitos* and *perfeitas*[17]
→ *write, writes, wrote, writing* and *written*[18]
→ English words beginning with *un*[19]
→ English words ending in *ness*[20]
→ *has been* and *have been* with up to one word between the auxiliary and the past participle[21]
→ *emoção* without using the keys for cedilla and tilde.[22]

Most lexical searches in COMPARA can be done using just this. But if you want to learn more about the query language we use, please refer to the IMS Corpus Workbench Manual, available at:
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/

Realistic queries

Use the Simple search to look up *At last he had come to his senses* (remember you must put each word in between double-quotes).

→ How many occurrences of *At last he had come to his senses* did you find?[23]

You probably did not get any results for the above because it is a sentence that is not likely to be in any of the texts of the corpus. When using COMPARA (or any other corpus), full sentences or long sequences of words do not usually work. To work well with corpora, you must operate with realistic chunks of language that are likely to have been previously used in the texts that make up the corpus.

Now try searching for the following, and see how many occurrences of each there are:

→ "come" "to" "(my|your|his|her|our)" "senses"[24]
→ "to" "(my|your|his|her|our)" "senses"[25]
→ "(my|your|his|her|our)" "senses"[26]
→ "senses"[27]

When you don't find any hits for a long string of words, try working with a shorter query or replacing certain words in the string with a **".*"**, the wild card used to represent any word.

[17] Write **"perfeit[oa](s)?"** in the Portuguese to English search box.
[18] Write **"(write|writes|writing|written|wrote)"** in the English to Portuguese search box. Note that you must use vertical bars – and not slashes – to separate each inflection. In UK keyboards the vertical bar is on the top left, while in US keyboards it is below the backspace key.
[19] Write **"un.*"** in the English to Portuguese search box.
[20] Write **".*ness"** in the English to Portuguese search box.
[21] Write **"(has|have)" []* "been" within 3** in the English to Portuguese search box.
[22] Write **"emocao"** in the Portuguese to English search box and put a check in the diacritic-insensitive box. Alternatively, leave the diacritic-insensitive box unchecked and write **"emocao" %d** instead.
[23] There are no occurrences of *At last he had come to his senses* in COMPARA 10.0.3. If you are using a later version of the corpus, this may have changed.
[24] There are 2 occurrences of *"come" "to" "(my|your|his|her|our)" "senses"* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[25] There are 8 occurrences of *"to" "(my|your|his|her|our)" "senses"* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[26] There are 23 occurrences of *"(my|your|his|her|our)" "senses"* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[27] There are 52 occurrences of *senses* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.

Now use the Simple search to look up *greenhouse gas emissions* (remember to surround each word with double-quotes).

→ How many occurrences of this term did you find?[28]

COMPARA contains only fiction texts at the moment, and it is unlikely that you should find many technical terms in fiction. For this reason, COMPARA works best with words and expressions likely to be found in general language and the language of fiction.

→ Group the words below into words and terms that you might find in COMPARA, and words and terms that are unlikely to be found in COMPARA[29]:

What	if	concordances	mecânica dos fluidos
ele(s)	said	heart	heart valve replacement
quando	não	quotient	dia

Where are the results from?

You already know that for now there are only fiction texts in the corpus. But in what varieties of Portuguese and English are they? Who are the authors? Can just any translation be included in the corpus? What if the translation is not a very good one?

Next to each concordance there is a letter & number code. When pointing your cursor to the code, you can see who the author and translator of the concordance are and the titles of the texts in question. By clicking on the code, you can obtain the full reference of those texts.

Try any search on the Simple search and scroll down the results until you find a concordance line that interests you. Point the cursor to the letter & number code next to it to see the name of the author and of the translator and the titles of the texts in question. Next, click on the code to obtain its full reference of those texts in the Bibliographic references page.

You can also arrive at the Bibliographic references page by clicking directly on the link under Texts in COMPARA in the menu. Scrolling up and down Bibliographic references gives you information about all the texts in the corpus.

Use the information in Bibliographic references to answer the following questions:

→ How many texts by Julian Barnes are there in the corpus?[30]
→ Are the texts integral works or just extracts?[31]

Because of copyright restrictions, the texts in the corpus are generally 30% extracts of the entire work. These extracts are selected at random, from the beginning, middle or end of the book.

→ How many translations of David Lodge's *Therapy* are there in the corpus?[32]

---

[28] There were no occurrences of *greenhouse gas emissions* in COMPARA 10.0.3. If you are using a later version of the corpus, this may have changed.
[29] Words and expressions that you might find in COMPARA: *What, if, ele(s), said, heart, quando, não, dia*; words and expressions that are unlikely to be found in COMPARA: *concordances, mecânica dos fluidos, heart valve replacement, quotient.*
[30] There are 3 texts by Julian Barnes in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.
[31] Extracts.
[32] There are two translations of David Lodge's *Therapy* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.

→ What is the letter&number code for the text pair containing the Brazilian Portuguese translation of David Lodge's *Therapy*?[33]

Some source texts in COMPARA are aligned with more than one translation. Whenever this occurs, the first translation to be processed is identified by T1 at the end of the letter&number code, the second translation is identified by T2, and so on.

Click on the Quantitative summary of the Texts in COMPARA in the menu on the left. You will find here a summary of the information about the texts within COMPARA.

→ How many different source texts and translations are there in the corpus?[34]
→ How many English and Portuguese words are there in the corpus?[35]

At the time of writing this tutorial, COMPARA is the largest edited parallel corpus in the world.

→ What varieties of Portuguese and of English are there in the corpus?[36]

All varieties of Portuguese and of English are allowed in the corpus. Although there is no preference for any particular variety of Portuguese or English, some varieties are more represented than others.

→ When do the oldest and the most recent texts in the corpus date from?[37]

Texts published at any date may be included in COMPARA. Because of copyright restrictions, texts by living authors and translators or by authors and translators who died less than seventy years ago require special permission to be included in the corpus.

→ Can just any source text and any translation be included in the corpus?[38]

All source texts and translations in COMPARA must have been published. This means not just in multiple copies, but also copyright registered or recorded by a major indexing service. Although this does not in itself guarantee the quality of the source texts and translations, published texts are normally subjected to careful proofreading and editing, and they must merit publication in the first place. Also, only English and Portuguese source texts, and English directly translated from Portuguese or Portuguese directly translated from English are admitted in COMPARA. The interference of other languages is not allowed. Texts generated by machine-translation programs are also excluded.

[33] EBDL1T2.
[34] COMPARA 10.0.3 contains 72 source texts and 75 translations. The number of texts may have increased if you are using a later version of the corpus.
[35] COMPARA 10.0.3 contains 1436187 words in Portuguese and 1544294 words in English. These numbers may have increased if you are using a later version of the corpus.
[36] COMPARA 10.0.3 contains texts in Portuguese from Angola, Brazil, Mozambique and Portugal, and texts in English from South Africa, the United Kingdom and the United States. If you are using a later version of the corpus, there may be more varieties.
[37] The oldest text in COMPARA 10.0.3 is Mary Shelley's *Frankenstein*, published in 1818, and the most recent texts were published in 2002. If you are using a later version of the corpus, this may have changed.
[38] No.

Interesting queries

Having got an idea of how the Simple search works and of where the texts in COMPARA come from, the question now is what kind of queries are likely to be interesting. The answer to this is not simple, for there are countless questions to choose from, and different users will be interested in different answers. For example:

A language student might simply be interested in using COMPARA as an online, contextualized bilingual dictionary. Think of a word that you don't know how to say in Portuguese and try looking it up in COMPARA.

A language teacher might be interested in using COMPARA to teach about false friends, or verbs or prepositions that do not have a one-to-one equivalence in Portuguese and English, etc. Teachers can copy the results from COMPARA onto a word processor and then adapt them into a test or worksheet for their students.

A translator or lexicographer might be interested in the different ways a given word has been translated, see what the most frequent translations of the word are, or how different translations are used in different contexts, and so on.

A translation researcher might be interested in finding out if certain words and expressions are more frequent in translated language than in source texts, or if the translation sentences are longer than the source-text sentences, and so on. See Publications under Specific Documentation if you want to read about previous research that makes use of COMPARA.

Whatever your needs, at one point you might find that the Simple search is not enough. Which is when you should give the Advanced search a try.

2. THE ADVANCED SEARCH

Click on the link to the Advanced search. The form you have to fill in contains four different steps. To navigate along those steps you can click on their respective tabs or, alternatively, use the expanded view if you want to navigate without the tabs.

The expanded view is similar to an earlier version of the advanced search form. It is recommended for increased text size and improved accessibility or for repetitive queries.

Step 1 - Select language direction

When using the Advanced search, the first thing you must do is select the language direction of your search.

→ What language direction should you choose for looking up "*saudade(s)?"*?[39]
→ What language direction should you use for looking up *"eventually"*?[40]

COMPARA's default language direction is **From Portuguese to English**. Whenever you are entering words or expressions in English, you mustn't forget to change the language direction to **From English to Portuguese**.

[39] From Portuguese to English. Use it whenever your starting point is a word or expression in Portuguese.
[40] From English to Portuguese. Use it whenever your starting point is a word or expression in English.

Step 2 – Enter query

The box on the left is the one you use to write down your search expression, in the same way as in the Simple search. It can be used for words and expressions in either Portuguese or English, for you have already determined the language direction in step 1.

Alignment constraints

If you are conducting a search in English, the box for **alignment constraints** on the right allows you to look up words and expressions on the Portuguese part of the corpus at the same time. For example, if you want to know if there are any concordances with *eventually* on the English side and *eventualmente* on the Portuguese side, do as follows:

Select **From English to Portuguese** in step 1. Write *"eventually"* in the search box on the left and make the query case-insensitive. Write *"eventualmente"* in the box for alignment constraints on the right and check the case-insensitive box underneath. Submit your query.

→ How many concordances for *eventually* on the English side and *eventualmente* on the Portuguese side did you find?[41]
→ In what language is the source text of these concordances, and what can you conclude from that?[42]

Now if you want to see if there are any concordances with *eventually* on the English side but **without** *eventualmente* on the Portuguese side, go back to the Advanced search and do as follows:

Select **From English to Portuguese** in step 1. Write *"eventually"* in the search box on the left and put a tick in the case-insensitive option. Write *!"eventualmente"* in the search box on the right (with an exclamation mark before the search term) and make it case-insensitive. Submit your query.

→ How many concordances did you find and what can you conclude from that?[43]
→ Can you list some ways in which *eventually* has been rendered in Portuguese?[44]

When the search direction is **From English to Portuguese**, use the search box on the left for an expression in English and an alignment constraint in Portuguese. If you change the search direction in step 1 to **From Portuguese to English**, use the search box on the left to write a search expression in Portuguese and an alignment constraint in English. The use of alignment constraints is not compulsory. If you are not interested in looking up words and expressions in Portuguese and English at the same time, simply leave the box for alignment constraints blank.

Go back to Advanced search and clear the form. Click on tab 2.2 for further query options. This part of the form allows you to search for a number of additional features. Here is how:

---

[41] There were 3 concordances with the word *eventually* on the English side and the word *eventualmente* on the Portuguese side in COMPARA 10.0.3 (if you are using a later version of the corpus, this may have changed).
[42] The source text is in English (the letter&number code begins with an E), which means that *eventually* was translated into *eventualmente*, and not the other way round.
[43] There were 82 concordances with *eventually* on the English side and **without** *eventualmente* on the Portuguese side in COMPARA 10.0.3 (if you are using a later version of the corpus, this may have changed). It is possible to conclude that *eventually* is equivalent to something other than *eventualmente* in 96% of its occurrences in a corpus of almost three million words. *Eventualmente* therefore does not seem to be a very good translation of *eventually*.
[44] *acabar por, afinal de contas, por fim finalmente, até,* etc.

Translators' notes

Tick the box saying "translator's notes" and submit.

→ How many and in what language were the translators' notes retrieved?[45]
→ What do you have to do to look up translators' notes in the other language?[46]

Titles

Go back to Advanced search, clear the form and have a go at looking up titles on both the English and the Portuguese side of the corpus.

→ Are the titles retrieved the ones which make up the corpus (i.e., those listed in Bibliographic references)?[47]

The option "titles" does not give you the titles of the corpus texts themselves. Rather, it allows you to retrieve both real and fictional titles of books, newspapers, magazines, films, plays, television programmes, songs (etc.) *cited* in the corpus texts. To find out which texts make up the corpus, you can click in Texts in COMPARA or on the letter&number code next to the concordances.

Foreign words

Go back to Advanced search, clear the form and have a go at looking up foreign words and expressions on the Portuguese side of the corpus.

→ How many foreign words or expressions are there in the Portuguese part of the corpus? List four of them.[48]

Go back to Advanced search, change the language direction, and have a go at looking up foreign words and expressions on the English side of the corpus.

→ How many foreign words or expressions are there in the English part of the corpus? List four of them.[49]

→ Do the foreign words and expressions consist of common nouns alone?[50]

Proper nouns that are made of or include common nouns have also been considered **foreign**. For example, *Bouvard et Pecuchet* is considered foreign because of the French conjunction *et*. However, only the foreign words and expressions that have been highlighted by the author or the translator (usually in italics) can be retrieved automatically. There may be other words and expressions in the corpus which you consider to be in a language different from the main language of the text, but which are not automatically retrievable because the author or translator did not highlight them.

[45] There were 98 Portuguese translators' notes in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.
[46] Change the language direction in step 1.
[47] No.
[48] COMPARA 10.0.3 contained 2217 foreign words or expressions in the Portuguese part of the corpus. Here are four examples: Knees R'Us, Benson and Hedges, chignon, snob. If you are using a later version of the corpus, there may be more foreign words or expressions.
[49] COMPARA 10.0.3 contained 1181 foreign words or expressions in the English part of the corpus. Here are four examples: Genson, cojones, n'est ce pas, paraphema. If you are using a later version of the corpus, there may be more foreign words or expressions.
[50] No.

Within-sentence emphasis

Go back to Advanced search, clear the form and have a go at looking up emphasis in the Portuguese-English and then in the English-Portuguese direction.

→ Which language of the corpus uses stressed words more often?[51]

Named entities

Go back to Advanced search, clear the form and have a go at looking up named entities in both language directions.

Named entities are names used to identify brands, shops, hotels, companies, products, doctrines, etc. They can only be retrieved automatically if the author or translator of the corpus text has highlighted them.

Sentence division changes

COMPARA also allows you to have a look at the sentences that translators have split, joined, added to, deleted from and reordered in translation. Tick the appropriate boxes if you want to have a look at that.

Step 3 – Using only part of the corpus

COMPARA's Simple search uses by default all the texts in the corpus. But if you look at the Bibliographic references page and see what those texts are, you might not be interested in using them all. Step 3 in the Advanced search allows you to use only the texts you want.

3.1 Selecting texts according to language variety

Go to Advanced search, clear the form and make a case-insensitive search for "*it*" in the English to Portuguese direction.

→ How many occurrences of "*it*" and *"It"* were there?[52]

Now go back to Advanced search, do not clear the form, and tick the box saying South African English.

→ How many occurrences of *"it"* and *"It"* were there now?[53]

Do the same first for British, and then for American English.

→ How many occurrences did you find?[54]

[51] In COMPARA 10.0.3, there are more stressed words and expressions in English. This may have changed if you are using a later version of the corpus.
[52] There were 16487 hits for *it* and *It* in all the English texts in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[53] There were 901 hits for *it* and *It* in the South African English texts in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have increased.
[54] There were 9647 hits for *it* and *It* in the British English texts and 5939 hits in the American English texts in COMPARA 10.0.3. If you are using a later version of the corpus, these numbers may have increased.

The different results obtained for different varieties of English do not mean that people use word *it* more often in certain varieties of English than in others. First, you cannot make general claims about South African, British and American English using just the texts in COMPARA, which are only representative of a comparatively very small amount of fiction written in these language varieties. Second, the results are not proportional, because the South African part of the corpus is much smaller than the British and American English parts (see Quantitative summary).

Repeat the same searches and this time pay attention to the box in the results which says **Short description of the corpus used in the present search** to find out how many South African, British, and American English words were used in your queries.

→ Use this information to complete the table below[55]:

Variety of English	**occurrences of *it/It***	**Total number of words**	**% of *it/It***
American			
British			
South African			

→ Now repeat the search for *it* and *It*, restricting the corpus to English from South Africa and Portuguese from Angola. How many occurrences were there?[56]

Some language variety combinations, like South African English and Angolan Portuguese, are not available in COMPARA. If you are not sure which language variety combinations exist in the corpus, have a look at the Quantitative summary page.

3.2 Selecting texts according to dates of publication

You can also use only a subset of COMPARA by limiting your searches to texts published only before or after certain dates.

→ Clear the form and carry out a search for *computer* in the English-Portuguese direction for source texts published before 1975. How many are there?[57]
→ Go back and repeat the search for texts published after 1975. How many hits are there?[58]
→ Clear the form again and conduct a blank search for translations published before 1900. How many are there?[59]

---

[55] Table completed with figures taken from COMPARA 10.0.3. If you are using a later version of the corpus, these numbers may have changed:

Variety of English	**occurrences of *it/It***	**Total number of words**	**% of *it/It***
American	5939	673046	0.9
British	9647	1021729	0.9
South African	901	101962	0.9

[56] There were no hits for that query in South African English and Angolan Portuguese COMPARA 10.0.3, for there were no South-African English source texts translated into Angolan Portuguese and no Angolan Portuguese source texts translated into South African English.
[57] There was only one hit for *computer* in the English texts published before 1975 of COMPARA 10.0.3 – the word appears in a text by Nadine Gordimer published in 1974. If you are using a later version of the corpus, this may have changed.
[58] There are 41 hits for *computer* in the texts published after 1975 of COMPARA 10.0.3.
[59] All hits are from the same text, for there is only one translation published before 1900 in COMPARA 10.0.3: José de Alencar's *Iracema,* translated into English by Lady Isabel Burton, and published in 1886. If you are using a later version of the corpus, this may have changed.

In version 10.0.3 of COMPARA, the oldest text dates from 1818, and the most recent ones were published in 2002.

3.3 Keeping original texts and translations apart

→ Clear the Advanced search form, and search for *already* (case-insensitive) in the English-Portuguese direction. How many hits were there?[60]

→ Does the letter&number code of the concordances retrieved always begin with the same letter?[61]

→ Now go back to the Advanced search, do *not* clear the form, tick the box in step 3.3 saying that searches should go only from source texts to translations, and press search. How many hits were there?[62]

→ Does the letter&number code of the concordances retrieved always begin with the same letter?[63]

→ Go back to the Advanced search, do *not* clear the form, but this time leave the box saying that searches should go only from source texts to translations unchecked and put a tick on the other box, saying that the searches should go from translations back to source texts. Press search. How many hits were there?[64]

→ Does the letter&number code of the concordances retrieved always begin with the same letter?[65]

→ Is the word *already* more frequent in original or in translated English?[66]

Most translation researchers will agree that the language of translations is different from the language of texts that are not translations. Whenever you want to test claims about the differences between translational and non-translational language, this option is important. This option may also be important if you are using COMPARA for language teaching. There are times when you may want to shelter learners from the language of translations, and times when you might want to deliberately show students how certain words and language structures have been translated.

3.4 Selecting specific texts

Each source-text and translation pair in COMPARA is represented by a unique letter&number code. If you move your cursor over a code you can see which texts it stands for. The code is also linked to the Bibliographic references page, where you can obtain full references plus information on text length and language variety. In section 3.4 of the Advanced search, you can guide yourself by these codes to select whichever text pairs you wish to use in your searches.

---

[60] There were 916 hits for *already* in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.

[61] No. The first concordances shown begin with an E, meaning that the source texts are in English, and the translation sides of the concordances are in Portuguese. The last concordances shown begin with a P, meaning that the source texts are in Portuguese, and that the English parts of the concordances are English translations from Portuguese.

[62] There were 332 hits for *already* in the English source texts in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.

[63] Yes. All concordances begin with an E, meaning that only English source texts have been used in this search.

[64] There were 584 hits for *already* in the English translation texts in COMPARA 10.0.3. If you are using a later version of the corpus, this number may have changed.

[65] Yes. All concordances begin with a P, meaning that only English translated from the Portuguese has been used in this search.

[66] In COMPARA 10.0.3, *already* appears 332 times in 819660 words of original English, and 584 times in 724220 words of English translated from Portuguese (you can look up the total number of words in original and translated English in the Quantitative summary page, or in *Short description of the corpus used in the present search* which appears at the top of your results page.). *Already* is used more than twice as often in translated English!

→ In section 3.4 of the Advanced search click on the letter&number code EBDL1T1 to find out what this text pair is. Do the same for EBDL1T2. Which text pairs are these?[67]

Clear the Advanced search form, check the boxes for EBDL1T1 and EBDL1T2, and search for *suit* in the English to Portuguese direction.

→ What are the European and Brazilian Portuguese translations of *suit* given?[68]

→ Look at the numbers in brackets underneath each letter&number code. Do you have any idea why the numbers 555, 689 and 1339 are repeated?[69]

The numbers in brackets underneath each letter&number code serve to give a unique identity to every source text sentence in the corpus. The word *suit* appears in the 555th, the 689th and the 1339th sentences of the corpus extract of David Lodge's *Therapy*.

3.5 Searching within specific authors

In step 3.5, you can select specific authors and search only within the texts by these authors.

→ Use section 3.5 to look up *black* just in texts by Nadine Gordimer and then just in texts by Henry James. Which author uses the word more frequently?[70]

Step 4 – Choosing output

All the searches you've conducted so far, both in the Simple search and in the Advanced search, presented your results in the form of parallel concordances, which is COMPARA's default output. The Advanced search form allows you to get other types of results too.

Alignment properties

If you click on the box that says *show alignment properties*, you will be able to see whether the source text sentence upon which each concordance is based has been split, joined, deleted, added to or reordered in the translation.

Clear the form. In step 2.2, try out a few searches for sentences added to, deleted from, joined and split in translation while selecting **searches to go from source text to translations** in step 3.3 and, in step 4, you leave the concordance box checked and select **show alignment properties**. Press enter and look at the numbers underneath the numbers in brackets that identify each concordance line.

→ What do you think 1-0, 1-1/2, 1-2, and so on stand for?[71]

Hide translators' notes

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2.2, select **translator's notes**. Press enter to see English translators' notes. Go back to the Advanced search form, do *not* clear it, and select **hide translators' notes** in step 4. Press enter.

---

[67] David Lodge's *Therapy* translated into European (EBDL1T1) and Brazilian (EBDL1T2) Portuguese.
[68] In European Portuguese, *suit* has been rendered as *fato* and *é bom*; in Brazilian Portuguese, *suit* has been rendered as *agasalho*, *terno* and *não seria mau*.
[69] The numbers are repeated in this query because the source-text sentences are repeated.
[70] Nadine Gordimer.
[71] 1-0 identifies source text sentences that have not been translated. 1-1/2 identifies source text sentences that are equivalent to half sentences in the translation. 1-2 identifies source text sentences that have been split into two separate translation sentences, etc.

→ What do you see?[72]

Show POS

Go back to the Advanced search. Clear the form. In step 1, select the **Portuguese to English** direction. In step 2, write *gosto* and, in step 4, select **concordance** plus **show POS.** Press enter.

Instead of normal text, you'll see that there are slashes, letters and codes following each word. These letters and codes identify the POS (part-of-speech) category of each word.

→ What codes are used after the word *gosto*?[73]

All queries in the examples given in this tutorial are lexical queries which do not contain any grammar information. We have recently introduced grammatical annotation to COMPARA so that users can also use grammatical information in their queries. For example, grammatical annotation makes it possible to retrieve concordances for the word *gosto* used as a noun separately from concordances for *gosto* as a verb. The Portuguese grammar tags have been automatically inserted using the PALAVRAS parser for the Portuguese language and the English ones have been added using the CLAWS tagger. The output of both is currently being revised manually in order to improve accuracy.

Although the grammar annotation has not been fully revised yet, it is already possible to carry out queries containing grammatical information provided you remember there is still a small margin of error in the results. A separate manual to explain how to carry out these queries is underway. If you want to start learning about grammatical queries straight away, click on the help button next to box for entering queries. The last examples given involve grammar. For further information, see Grammar annotation in the menu for Specific documentation.

Distribution of forms

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *".*ly"* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution of forms**. Press enter.

→ What do you see?[74]

Distribution of part-of-speech

Go back to the Advanced search. Clear the form. In step 1, select the **Portuguese to English** direction. In step 2, write *"gosto"* in the search box and check the case-insensitive box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution of part-of-speech**. Press enter.

→ Is *gosto* more often a verb or a noun?[75]

---

[72] You should be able to see where translators' notes have been inserted, without having to see the notes themselves.
[73] You should see **gosto/N, gosto/V_fmc** and **gosto/V.** The first code identifies *gosto* as a noun; the next two codes identify *gosto* as types of verbs.
[74] All words ending in *ly* in the English part of the corpus. The most frequent one in COMPARA 10.0.3 was *only*, which appeared 2424 times. This number may have increased if you are using a later version of the corpus.
[75] In COMPARA 10.0.3, there are 152 hits for *gosto* as a noun, and 157 (109+48) hits for the word as a verb. This may have changed is you are using a later version of the corpus.

Distribution of lemmas

Go back to the Advanced search. Clear the form. In step 1, select the **Portuguese to English** direction. In step 2, write *caminho* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution of lemma**. Press enter.

→ Is *caminho* more an inflection of the verb *caminhar* or a noun meaning trail, path, etc.?[76]

Distribution of tense, person, number and gender

These distributions are only available for Portuguese.

If you search for [lema="trabalhar"] and select a **distribution of tense**, you will be able to see the different tenses of this verb represented in the corpus. Which tense is most frequent?[77]

If you search for [lema="sugerir"] and select a **distribution of person and number**, you will be able to see the different person and number inflections of this verb represented in the corpus. Which is the most frequent one?[78]

If you search for [lema="querido"] and select a **distribution of gender**, you will be able to see whether it is more frequent in the feminine or in the masculine form.[79]

Distribution of sources

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *"love"* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution of sources**. Press enter.

→ What text in the corpus mentions the word *love* most often?[80]

Distribution in original and translated text

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *"love"* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution of original and translated text**. Press enter.

→ Is the word *love* more frequent in original or in translated English?[81]

Distribution according to variety of Portuguese

Go back to the Advanced search. Clear the form. In step 1, select the **Portuguese to English** direction. In step 2, write *fato* in the search box. Skip step 3. In step 4, leave the concordance box checked and tick the box saying **distribution according to Portuguese variety**. Press enter.

---

[76] There are only 7 hits for *caminho* as an inflection of *caminhar* in COMPARA 10.0.3. All other hits pertain to the noun *caminho*. This may have changed is you are using a later version of the corpus.
[77] In COMPARA 10.0.3, the most frequent tense for the verb *trabalhar* is the infinitive.
[78] In COMPARA 10.0.3, *sugerir* appears most frequently in the third person singular.
[79] In COMPARA 10.0.3, *querido* is more often in the masculine form.
[80] In PBMA5, *love* is mentioned 38 times in 22203 words, i.e., 17 times in every 10 thousand words. Compare with EUEP1, where the word *love* only appears once in 24412 words, i.e., only 0.4 times in every 10 thousand words.
[81] In COMPARA 10.0.3, *love* is mentioned 451 times in 724220 words of English that has been translated from Portuguese (i.e., 6.2 times in every ten thousand words), and 296 times in 819660 words of original English texts (i.e., 3.6 times in every ten thousand words). There seems to be more love in English translated from Portuguese! If you are using a later version of the corpus, these numbers may have changed.

→ In what variety of Portuguese is *fato* more frequent? Read the concordances and try and explain why this may be so.[82]

Distribution according to variety of English

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *"colour"* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **distribution according to English variety**. Press enter.

→ In what variety of English is *colour* more frequent?[83]

Combined distribution of Portuguese and English search expressions

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *"actually"* in the search box and *"a(c)?tualmente"* in the alignment constraint box. Put a tick in the case-insensitive option underneath both boxes. Skip step 3. In step 4, leave the concordance box unchecked and tick the box saying **Combined distribution of Portuguese and English search expressions**. Press enter.

→ How many hits for *actually/Actually* and how many hits for *actualmente/atualmente/ Actualmente/Atualmente* were there? How many times did the two coincide in a single concordance?[84]

If you look at the **Corpus used in this search**, you can also conclude that *actually* figures 15 times in every 100 thousand words of English, while *actualmente* appears only twice in every 100 thousand words of Portuguese. In a corpus of three million words like COMPARA, the word *actually* is over seven times more frequent than *actualmente*!

Distribution per author

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *gun* in the search box. Skip step 3. In step 4, leave the concordance box unchecked and tick the **distribution per author box**. Press enter. Repeat the same steps for the word *killer*.

→ What author in COMPARA writes more about guns and killers?[85]

Distribution of colour

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *[sem="colour.*"]* in the search box. Select Kazuo Ishiguro in step 3.5. In step 4, leave the concordance box unchecked and tick the **distribution of colour** box. Press enter.

---

[82] The word *fato* is comparatively much more frequent in Brazilian Portuguese (in COMPARA 10.0.3, there are 68 hits in 557712 words, i.e., 12 hits in every 100 thousand words). The concordances show that the word has different meanings in Brazil and in Portugal and Mozambique, which can affect the frequency with which the word is used.

[83] In COMPARA 10.0.3, *colour* is comparatively more frequent in South-African English (15 hits in 100 thousand words), followed by British English (10 hits in 100 thousand words). There are only 0.6 hits for *colour* in every 100 thousand words in American English. If the query is repeated using *color* instead, there will be more hits in American English.

[84] In COMPARA 10.0.3, there were 232 concordances with *actually* and *Actually* on the English side, and 31 concordances with *actualmente*, *atualmente*, *Actualmente* or *Atualmente* on the Portuguese side. However, in only one concordance did the English and the Portuguese terms coincide, indicating that they are not good language equivalents. If you are using a later version of the corpus, these numbers may have changed.

[85] In COMPARA 10.0.3, Patrícia Mello uses the words *gun* and *killer* more frequently.

→ What colour does Kazuo Ishiguro use most?[86]

To look up colour in English, use *[sem="colour.*"]*. For Portuguese, use the following expression in the search box: *[sem="cor.*"]*

Semantic distribution

Go back to the Advanced search. Clear the form. In step 1, select the **English to Portuguese** direction. In step 2, write *[sem="colour.*"]* in the search box. Skip step 3. In step 4, leave the concordance box checked and tick the **semantic distribution** box. Press enter.

→ How many words related to colour are there in COMPARA?[87]
→ What kind of words are tagged with "colour"?[88]

For now, the only semantic category available in COMPARA is colour. If you are interested in using COMPARA to research another category of meaning, please contact us.

3. CONCLUSION

Now that you have finished this tutorial, try out your own searches and let us know if you have any questions or problems that have not been dealt with in our link to Help or Questions from Users. We also welcome any suggestions you might have. Use the feedback form available at *Send questions, comments and suggestions* .

At http://www.linguateca.pt/COMPARA/ComparaPublications.html, we maintain a page with a bibliography containing descriptions of COMPARA and studies that make use of COMPARA. Let us know if you have any publications or if you present a conference paper involving the use of COMPARA, so that we can add it to this page.

COMPARA is (partly) financed by the Fundação para a Ciência e Tecnologia (Portugal), with funds from POSI (POSI/PLP/43931/2001) and POSC (POSC 339/1.3/C/NAC) made available through FCCN.

[86] In COMPARA 10.0.3, Kazuo Ishiguro uses mostly *black*.
[87] In COMPARA 10.0.3, there are 3505 words related to colour in the English part of the corpus.
[88] Here are a few examples: *red-hot, orange, black-and-white, coloured, gingery-brown*, etc.