The corpus, its users and their needs: a user-oriented evaluation of COMPARA

Diana Santos
SINTEF ICT
Pb 124 Blindern
N-0314 Oslo, Norway

Ana Frankenberg-Garcia
Instituto Superior de Línguas e Administração (ISLA)
Lisbon, Portugal

Abstract

COMPARA is a bidirectional parallel corpus of English and Portuguese, currently with 3 million words. The corpus was launched in 2000 and at present it is possibly the largest edited parallel corpus publicly available on the Web, with roughly 6,000 corpus queries per month. This paper summarizes an analysis of six years of corpus use. We begin by looking at user studies for language resources, especially corpora, and then we provide a snapshot of COMPARA's users and their behaviour based on log analysis. Particular emphasis is given to the language interface preferred by users (Portuguese and English are possible), the choice between the Simple and Complex Search modes, the reasons underlying null-results and behaviour after truncated output. The data has pointed us to cases where COMPARA's Web interface can be improved, and provided insights about our users and the problems they face, although further studies that distinguish between different kinds of users remain necessary.

Key words

Parallel corpora; usability; evaluation; log analysis; interface design; error analysis; English; Portuguese.

1 Introduction

Any researcher involved in the creation of corpora will know that the time and trouble invested in the task is by no means negligible. Notwithstanding the huge amount of effort involved, it is unfortunate that most existing corpora today are only available to and understood by a small, restricted community of users. Talking about the potential advantages of corpora for language learning and research is one thing. Analysing the impact these resources have on (prospective) users is something else. While there is a considerable body of literature dedicated to the former, surprisingly little has been said about the latter.

In fact, there is a huge void regarding the evaluation of corpora in general. In addition to fundamental attributes such as overall quality, comprehensive documentation, up-to-date maintenance and long-term preservation, it is also important to consider the accessibility and usability of corpora. If there is to be a better match between corpora and their users, then it is necessary to ask who exactly current and prospective users are, how easily they can use the corpora in question, and how well their needs can be addressed by them. Rather than have users give up, adapt their needs or lower their expectations regarding what they can obtain from corpora, an attempt should be made to improve corpora and corpus software so as to better comply with user requirements. Indeed, it is believed that concern with usability should be one of the driving forces of software development in general, and this is precisely what lies behind the rationale for many decisions taken in relation to the COMPARA corpus, available at www.linguateca.pt/COMPARA/.

What follows is an analysis carried out to learn more about the users of the corpus and their general behaviour. As access to the corpus is online and requires no registration, the user profiles in this study are are based on log files. The analysis covers the period between the time COMPARA was first tried out in May 2000 and 31 August 2006.

The information obtained has made it possible to determine where corpus queries have been coming from; how often they are made; what exactly they consist of; and who the users of the corpus are likely to be. Particular attention has been paid to corpus queries that were unsuccessful so as to learn why they failed and how users reacted when that happened. User-oriented improvements that have taken place in COMPARA over its six years of existence are assessed, some remaining puzzles are described, and unexpected behaviour is discussed. While far from giving a full picture of user behaviour, the study reported here – which we believe to be the first of its kind – significantly increases knowledge of (parallel) corpus-browsing behaviour. To conclude, an account is given of different user classes, which we intend to study in the near future.

1.1 A brief presentation of COMPARA

COMPARA was developed under the scope of Linguateca, a resource centre for the computational processing of the Portuguese language. The corpus is an extensible bidirectional parallel corpus of English and Portuguese. In its current version 8.0, it contains around 3 million words. English from Britain, the United States and South Africa, and Portuguese from Portugal, Brazil, Mozambique and Angola are currently represented in the corpus in the work of 35 different authors and 45 different translators. Only published texts in English translated directly from Portuguese and Portuguese translated directly from English are admitted in the corpus. The corpus files are currently based on 74 different pairs of original and translated extracts[1] of fiction, randomly taken from the beginning, middle or end of books. Both contemporary and non-contemporary works are represented in the corpus, with the oldest original text currently dating back to 1837, and the most recent one having been published in 2000. Translation dates range from 1886 to 2002.

Like many other resources hosted by Linguateca, access to COMPARA on the Web is free and requires no registration. This access is made via the DISPARA interface, which is simultaneously available in English and Portuguese and offers users two different search facilities. The Simple Search enables users to retrieve parallel concordances from the entire corpus, in both the English to Portuguese and the Portuguese to English direction. The Complex Search allows users to do the same and, in addition to that, restrict searches to different types of sub-corpora, retrieve other types of results (apart from or excluding parallel concordances), and carry out more sophisticated queries: users can set alignment constraints, as well as look up translators' notes, titles, foreign words, emphasis, named entities, and sentences that have been added, deleted, joined, split and reordered in translation. The rationale behind having two different search facilities available in two different languages was to make COMPARA as widely accessible as possible. Target users include not only corpus and computational linguists, but also language learners, language teachers, university lecturers, students and translators anywhere in the world and with little or no prior experience of using corpora.

COMPARA was first tested online at the end of May 2000, with just two pairs of texts in the corpus. In November 2000 it was presented for the first time at the CULT 2K - Corpus Use and Learning to Translate - conference and, shortly afterwards, in January 2001, it was announced in the corpora list, with half a dozen pairs of texts (65 thousand words) and an embryonic search interface. Although copyright permission had been obtained for many more texts and the corpus could still be improved in many ways, it was felt that it was important to provide access to whatever was available as soon as it became available. At the time, there was no other publicly available parallel corpus for the English-Portuguese language pair, and it would be simpler to deal with any problems that arose if the corpus was still small (Frankenberg-Garcia and Santos, 2003).

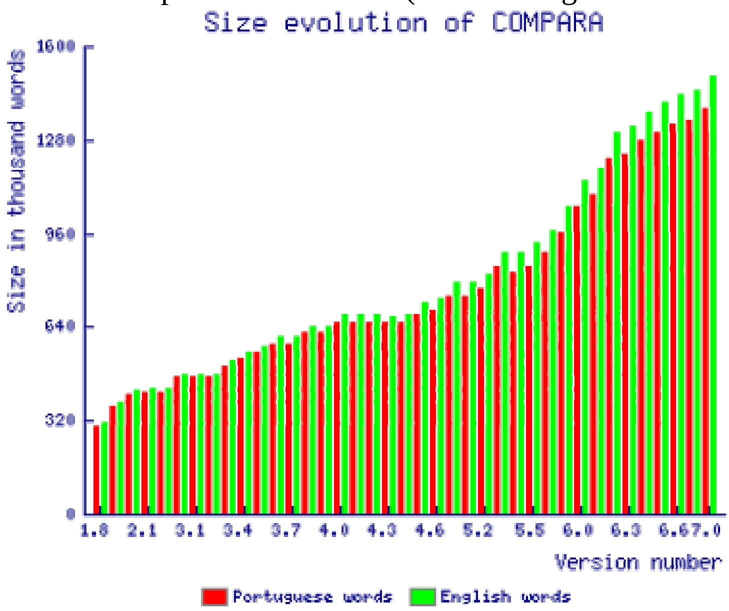


Figure 1: Size of COMPARA from version 1.8 (September 2002) to 8.0 (August 2006)

1.2 Usability in computational linguistics and corpus linguistics

Concern with the user has been a major trend in software engineering, such as in work on use cases (Jacobson, 1992) and within the tradition of human-computer interaction (see e.g. Nielsen, 1993; Dix et al. 1993; Helander et al., 1997). Disciplines like information retrieval (IR) and Web IR, including studies of Web search engines (Jansen et al., 2000) and library access (Jones et al., 2000) have also had a positive influence upon usability studies.

Within the field of computer-assisted language-learning (CALL), Noblitt and Bland (1991) analysed French learners using a CALL system and Frankenberg-Garcia (2005a) looked into the ways language learners chose to use electronic resources (corpora, termbanks and the Internet) and paper references in language production. Johns (1997) tested a concordancer for classroom use and Woolls (2000) reports on the user-driven design of a parallel concordancer. However, user-centred evaluation with respect to corpora remains scarce. Although the last two studies are specifically about corpora, they both deal primarily with the development of a *system* rather than with an integrated service to users (cf. Gaizauskas's (1998) distinction between evaluation of *systems* and *tasks*). Other studies about users and corpora have been conducted by Bernardini (2000), Kennedy and Miceli (2001) and Bianchi and Manca (2006). However, these studies focused on what corpus users needed to be taught rather than on how to make corpora more usable. In fact, the study of the way people perform corpus tasks with the view of improving the corpus service itself is practically unheard of.

In fact, while many corpus-minded researchers and educators are actively engaged in discussing the uses of corpora and teaching people how to use them, little attention has been paid to making corpora more user-friendly – which, as usability experts tell us, would then dispense with the need for teaching users. This last statement is, of course, a bit of an exaggeration, and both authors of the present paper are very keen on providing teaching and pedagogical material to support corpus use (for example, Frankenberg-Garcia (2004, 2006) and Santos (2006a, 2006b).

We deem it equally relevant, however, that teaching the whats and hows of corpora should be done with as little system-specific hindrances or misunderstandings as possible. Therefore, we must learn how people actually employ these systems in order to find out whether they entail unnecessary complications that can be simplified or even eliminated.

The present study has been inspired by the emerging discipline of Web usability studies (Masand and Spiliopoulou, 2000; Ivory and Hearst, 2001). Our primary objective here is to describe how COMPARA has been used according to records pertaining to a large set of user logs collected unobtrusively over a period of time. By studying how users perform corpus tasks, we wish to identify problem areas and then act on them, with the ultimate goal of making the corpus easier to use. We hope that this study and our decisions may inspire other corpus developers, and that the choices we made (or failed to make) may provide researchers in this area with grounds for comparison.

In a nutshell, the method employed in this study boils down to 1) observing users without disturbing them by analysing the “fingerprints” they leave when interacting with the corpus, and 2) trying to understand users' actions without directly communicating with them. Let us say from the start that – given our choice to make COMPARA as easy to use as possible – users do not authenticate, and we did not implement any cookies mechanism to keep track of sessions. So, it is not possible to single out individual users,[2] although there are work-arounds for some of these complications. Sullivan (1997) provides an enjoyable overview of the advantages and disadvantages of (server) log analysis, making the fine point that observation nicely complements experimentation. In this paper, we hope to show that the use of specifically-designed access logs can provide a wealth of information about corpus usability.

2 Usability of COMPARA: intentions and measurements

First, let us attempt to clarify what we mean by usability in a corpus context, both by defining the concept of usability and by explicating what is involved in making a corpus available to users. Starting with the latter, it is useful to distinguish between the following three different dimensions proposed by Santos (1998) (incidentally, one of the first papers ever on Web based access to corpora):

1. the bare corpus, i.e., the texts that form the corpus and their underlying selection and classification criteria (enriched with whatever information the texts are endowed with);
2. the corpus encoding system, i.e., the system that allows one to search the corpus and issue complex queries;
3. the interface between the above two, which is what the end user gets to see.

In COMPARA, the first of the above dimensions is referred to as COMPARA itself, or the COMPARA corpus, the second dimension is the IMS Corpus Workbench (Christ et al., 1999), and the third one is the DISPARA interface (Santos, 2002), behind which is all the software engineering environment required to create new versions of the corpus (with updates of both the corpus and its encoding system) and offer Web access to it.

Corpus usability, in turn, is taken here to mean the usability of the above three components as a whole.[3] Usability is defined by ISO (norm ISO9241-11) as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Now, the more concrete the group of users, the context of use and their goals, the easier it is to deal with the elusive concepts of effectiveness, efficiency and user satisfaction. This is not an easy task when what is at stake is a general public service on the Web like COMPARA.

In COMPARA – or, more precisely, in the COMPARA/DISPARA project – we tried to achieve a middle ground between creating a resource that would, on the one hand, correspond to the authors' expectations and wishes, and, on the one hand, meet the needs of a growing set of new users. So, while we tried to design COMPARA according to the state of the art in parallel corpus processing, we also made an undeniable effort to improve the interface and increase the functionalities it offered by observing (and listening to) the users underway. In our first paper about the corpus (presented at the CULT 2K Conference in November 2000, and later published as Frankenberg-Garcia and Santos, 2003), we even claimed that we intended further development of COMPARA to be user-driven.

With this said, let us admit that COMPARA is a real-world project, developed in a distributed fashion by people with other projects at hand, which means that it is unrealistic to guarantee that everything has worked out as smoothly. In particular, some problems which were easy to correct were detected much later than theoretically possible in an ideal world. Also, not all kinds of problems have unambiguous solutions. Often, things have to be tried out, to assess whether what seems to be an improvement is in fact helpful to the user out there. Some of this will be reported in the present paper, which is our first approach to come to grips with the usability of COMPARA as a whole.

We assume that our readers share with us a pre-theoretical idea of what a corpus can be used for and what kind of basic functionalities are expected, and also, that it makes sense that every function offered is logged in order for later study of both its popularity and eventual problems in its use.

We should like to make it clear, however, that we do not claim that unobstrusive studies are the best or the only way to come to grips with corpus usability. Traditional usability inspection methods are a natural complement to log analysis. Fortunately, the use of work-domain subjects instead of usability professionals has been recently argued for with the justification that "Typically, usability experts, software engineers or user interface (UI) designers do not have a thorough understanding of the context of use of a domain-specific work support system, but work-domain experts do" (Følstad, 2007), and since the developers of COMPARA are work-domain (corpora) experts as well, we can conduct our own inspection routines in a regular way associated to teaching or demonstration activities.

2.1 Log data and resources used

Let us start by describing how the "Web footprints" left by COMPARA's users can be analysed. As is customary in any Web application, we have two kinds of materials to help us understand users' behaviour (in addition, of course, to the invaluable feedback in the form of direct questions posed to our team, but which lie beyond the scope of the present article):

- standard Web server logs (in our case, Apache logs)
- service-specific logs, created for every query to COMPARA (the DISPARA system records various types of information related to a transaction, an example of which is given in Appendix 2).

Based on this, we have made use of two types of tools:

- general access statistics, which provide a general (and standard) quantitative view of the user mass[4];
- several specific programs that process the DISPARA logs and study specific aspects of interaction with COMPARA so as to test specific hypotheses and obtain fine-grained counts.

Most of the detailed analyses presented here refer to the period between the time COMPARA was first tried out in May 2000 and 30 September 2004, during which time there are records of 74,366 queries. However, in some more general analyses, we increased the time frame of the study so as to include more up-to-date data until 31 August 2006, amounting to 233,864 queries.

Although these queries include some that we – the corpus makers – made as users, it should be noted that practically no development work on the corpus is carried out on-line, and that the amount of test queries in the analysis that follows is negligible.[5]

2.2 Queries and sessions

In addition to individual queries (every search carried out in COMPARA), the concept of user session has been important from the start in transaction logs studies. Even in conceptually simpler applications like search engines, users in average perform more than one query per session. It is to be expected that this is even more likely to happen when interacting with a bilingual corpus like COMPARA.

When looking at the use of COMPARA beyond an individual request, we can introduce two different concepts of "user session":

1. session defined as a set of consecutive requests by a same user to COMPARA;
2. session defined as a navigation stretch around COMPARA's website as a whole, i.e. taking into account both navigation through help files, information files and corpus reference files.

The second of the above should give us some clues about the way COMPARA is used, and might allow us to, in a way similar to the study by Koch et al. (2005):

- measure the time employed to read documentation and/or help;
- discriminate between novice and advanced users by the way they entered and navigated in the site;
- assess how often users entered different parts of the site and how useful they seemed to be by inspecting movement to and from these particular pages.

Because of time constraints, however, it was not possible to perform the second kind of analysis. Although we have limited our analysis to the first type of user session, we hope to show that it produced plenty of material to consider. Simple grouping of requests by the same user allows us to study error recovery, related queries, and much more. As we will see in section 4, there is a large body of information that has to be seen in the context of what users are trying to do and when and in what order they do it.

3 Studying queries to COMPARA

This section contains information on the sum of all queries posed to the corpus until August 2006. It provides an overview of the use of COMPARA so far. We start by describing the geographical origin of queries, in Figure 2.

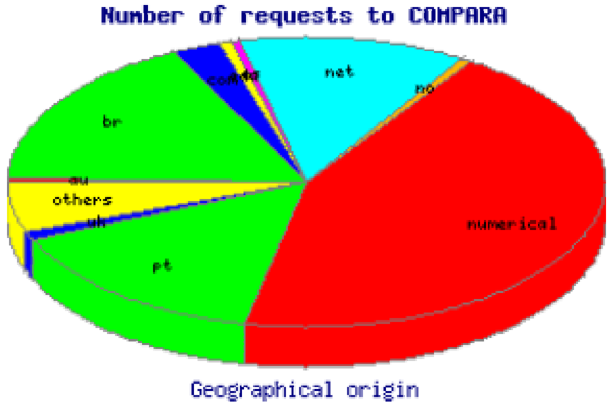


Figure 2. Where COMPARA users come from (May 2000-August 2006)

Because not all computers identify properly, in roughly half of the cases it was not possible to find out the computer's name and posit its geographical location. As can be seen, a large fraction of identifiable queries come from Brazil and Portugal. Figure 3 displays the distribution of queries over time. There is a steady increase in the number of queries, with interesting valleys during Brazilian and Portuguese holidays.

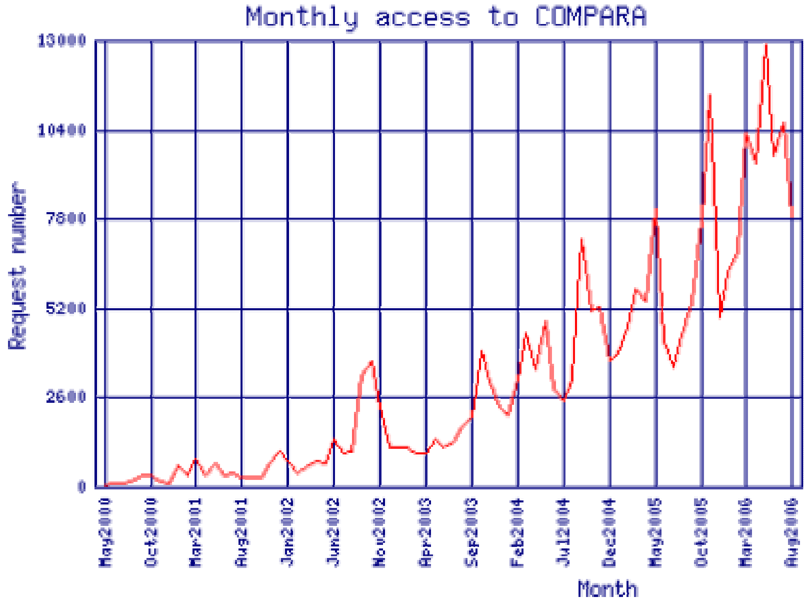


Figure 3. Number of COMPARA searches from 2000 to 2006

Although it is not possible to identify users any further, nothing prevents us from getting a broad but accurate picture as to which particular options they have employed. In the following sections we shall describe which part of the DISPARA interface users have chosen, what their queries were and the results they got.

3.1 Which part of the interface users choose

As already mentioned in section 1.1, every page in the COMPARA website is available in both Portuguese and English, so that people with very little Portuguese or very little English can still access them. This means that, irrespective of where users come from, it is possible for them to navigate through COMPARA's website in English or in Portuguese. Figure 4 shows which language users have chosen.

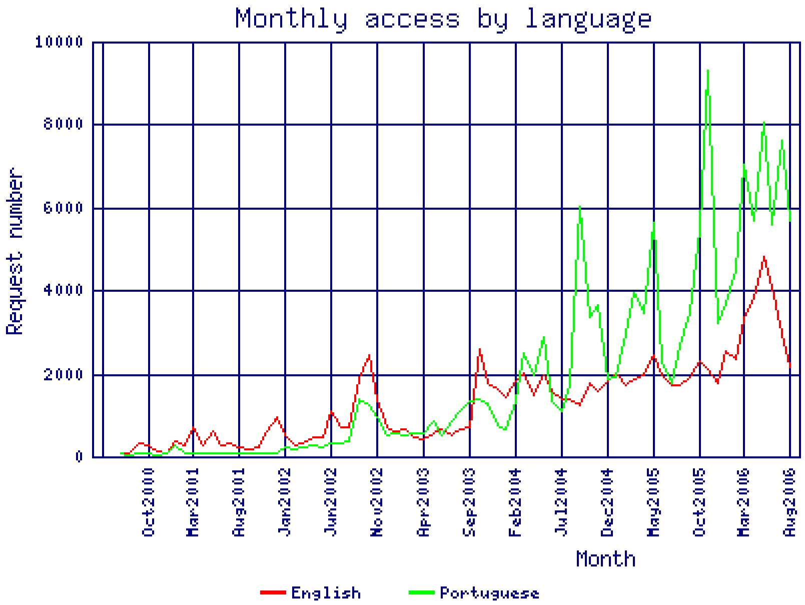


Figure 4: Interaction with COMPARA in English and in Portuguese

As can be seen in figure 4, until 2004 the preferred language was English, which comes as a surprise inasmuch as most identifiable users of the corpus come from Brazil and Portugal. We expected that the vast majority of users to be native Portuguese speakers. If this was true, why would they use English instead of their mother tongue? This rather unexpected situation prompted us to look into the language part in further detail (see section 4.4 below). In any case, the fact that both language services were actually *used* suggests that the two were not redundant: giving users the possibility of choosing between them seems to have been a valid design issue.

Apart from choosing the language they wish to use when interacting with COMPARA, users can also select between two different search interfaces. Very early in the project we decided to provide users with the option to choose between the Simple Search, which was made as simple as possible, not to discourage people with reduced computer skills, and the Complex Search, which provides a rich set of alternatives and offers a lot of querying power to a power user (see screenshots in appendix 3). The first thing users are requested to do when searching COMPARA is to choose which kind of search interface they want to use. Advanced users are expected to have bookmarked the Complex Search page and go directly to it.

Figure 5 depicts how often the two kinds of interface have been employed since the date they were provided. The comparative popularity of the Simple Search seems to indicate that the idea of providing users with an easy, no-frills interface has been well worth its while, and even advanced users, who often resort to the Complex Search, may use the shorter and more direct form offered by the Simple Search whenever the query they have can be dealt with in this limited mode.

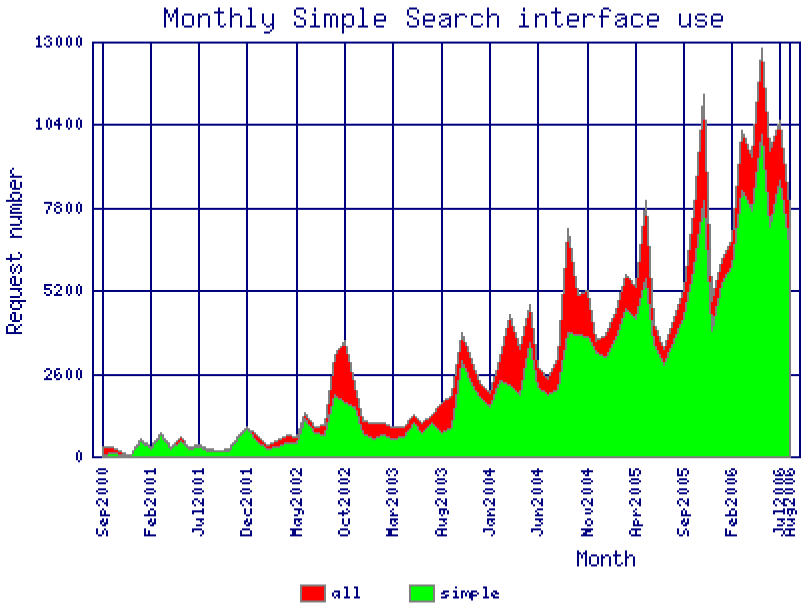


Figure 5: Interaction with COMPARA in Simple or Complex mode

3.2 What COMPARA users ask for

Users who select COMPARA's Simple Search are only able to retrieve concordances from the corpus. The user has no other choices apart from searching from Portuguese to English or from English to Portuguese. This choice is also available in the Complex Search mode, and the results pertaining to the distribution of these choices in the two interfaces are displayed in table 1.

Table 1: Search Direction in Simple Search and Complex Search (Jan 2003 - Sep 2004)

Search Direction	Complex Search	%	Simple Search	%
Portuguese to English	13,466	55.9%	16,209	89.4%
English to Portuguese	10,640	44.1%	1,929	10.6%

While Complex Search users interrogated the corpus in both language directions in a balanced way, Simple Search users have been overwhelmingly more interested in observing the translation of Portuguese into English. It is worth noting that the Portuguese to English direction is both the first button available in the Simple Search mode and the default search direction in the Complex Search. To search from English to Portuguese in the Simple Search, users would have to fill in the second box that appears on their screen, and users of the Complex Search would have to press one extra button in the form. The extent to which filling in the first box in the form (in the Simple Search) or not pressing the extra button to change the language direction (in the Complex Search) was intentional will be discussed later, in section 3.3.1.

While in the Simple Search users are only able to retrieve Portuguese-English or English-Portuguese parallel concordances, in the Complex Search six different kinds of results are currently available in one fell swoop (although by default the system only displays concordances):

1. Concordances
2. Distribution of forms
3. Distribution of part-of-speech
4. Distribution of lemma
5. Distribution of sources
6. Distribution in original and translated text

7. Distribution according to Portuguese (and/or English) variety

8. Combined distribution of Portuguese and English search expressions.

In addition to these different types of output, in the Complex Search users can restrict their searches to specific sub-corpora, which they can select text by text or by restricting the corpus in terms of publication date, language variety, and original vs. translated text. Other options available in the Complex Search that users might wish to retrieve are translators' notes, foreign words, emphasis, titles, named entities and different types of alignment (sentences preserved, added to, deleted from, split, joined and/or reordered in translation). Tables 2 to 5 summarize what users asked for and what they got in the 56,260 queries carried out in the Complex Search mode. When looking at the results, it is important to note that the different types of options available in the Complex Search were introduced at different moments in time, which makes it difficult to present general usage data comparable over the project's lifetime as a whole. Since COMPARA has evolved and changed in the course of its development, the information available through the logs has also changed. A short history of COMPARA is provided in Appendix 1 to allow for statistically minded readers to compensate for the apparent mismatches. The dates each new feature was introduced are provided in the tables. Percentages are given both relative to total interaction with COMPARA and relative to the periods where options were available.

Table 2: Type of output requested by users in the Complex Search (56,260 requests)

Type of output	**Number of queries**	**% in general**	**% in the period**	**Available since**
Concordances	54,893	97.57	97.57	Sep 2000
Distribution of forms	1,857	3.30	3.30	Sep 2000
Combined distribution in EN and PT	1,759	3.13	3.13	Sep 2000
Distribution of sources	1,322	2.35	2.35	Sep 2000
Distribution in original vs. translated text	939	1.67	2.12	Oct 2003
Distribution by English language variety	355	0.63	1.15	Oct 2004
Distribution by Portuguese language variety	254	0.45	0.82	Oct 2004
Distribution by part-of-speech[6]	236	0.42	1.87	Jan 2006
Distribution by lemma	58	0.10	0.87	May2006

Table 3: Sub-corpus selection in the Complex Search (56,260 requests)

Corpus selection	**Number of queries**	**% in general**	**% in the period**	**Available since**
Only originals	8,732	15.52	15.52	Sep 2000
Only translations	1,564	2.94	2.94	Sep 2000
Specific variet(y/ies) of Portuguese	11,516	20.47	20.47	Sep 2000
Specific variet(y/ies) of English	7,946	14.12	14.12	Sep 2000
Specific texts in the corpus	4,894	8.70	8.70	Sep 2000
Texts published in specific dates	816	1.45	1.48	June 2001

Table 4: Language varieties chosen in the Complex Search (56,260 requests)

Language variety	**Number of queries**	**%**	**% in the period**	**First text became available**
Brazil	10,011	17.80	17.80	from start
United States	6,288	11.18	11.18	from start
United Kingdom	4,371	7.76	7.76	from start

Portugal	1,883	3.34	3.34	from start
South Africa	1,743	3.10	3.16	8 January 2001
Angola	428	0.76	0.94	31 Aug 2003
Mozambique	351	0.62	0.69	14 Oct 2002

Table 5: Requests for other features available in the Complex Search (56,260 requests)

Features	Number	% in general	% in the period	Available since
Alignment	3,217	5.70	5.70	Sep 2000
Foreign words	2,072	3.68	3.68	Sep 2000
Translators' Notes	1,778	3.16	3.16	Sep 2000
Emphasis	922	1.64	1.64	Sep 2000
Titles	840	1.49	1.49	Sep 2000
Named entities	408	0.72	0.88	Aug 2003

These tables prompt some discussion: First of all, the disproportionate amount of interest in singling out Brazilian Portuguese as compared to all other varieties offered is noteworthy. Further work has to be done to see whether these selections correlate with Brazilian users (or users located in Brazil) or, rather, with specifically Brazilian linguistic or cultural phenomena.

Another interesting piece of information, now more relevant to corpus use in general, is that users of COMPARA are clearly more conversant with the option of restricting the corpus to their specific goals (Tables 3 to 5) than with asking the system to produce other kinds of output, apart from concordances (Table 2). This may reflect a genuine interest of users for real examples and not so much quantitative data, but it may also indicate that users may (wrongly) use several selections in a row to get at what distribution would give them with one request. In any case, it seems that concordances are cognitively easier to grasp than distributions, and this may require some specific action to lead users to make further use of the latter. Even if we discount the cases where an additional kind of output was required together with a concordance, the option of asking for concordances alone was taken in 52,195 requests, making up 92.8% of all interactions.

3.3 What COMPARA users get

After examining the kind of query users have carried out, it is just as important to investigate what COMPARA has provided them with.

Table 6 displays the number of hits returned for each concordance output requested (excluding zero results). It should be noted that, for copyright reasons, COMPARA implements a threshold on the maximum number of results returned, which depends on the current size of the corpus and on the size of the subcorpora selected by the user. Again, this dynamic threshold has changed with time, and as the corpus grows it is expected that higher numbers of hits will be returned to users. The number of searches that reached the threshold and were therefore truncated by the system is presented as well (until September 2004 the upper limit was 1,000 hits per query).

Table 6: Number of hits per query returned

Number of hits returned	Frequency	Truncated output
1-9	53,053	
10-99	46,245	164
100-999	18,744	1,601
>1000	4,126	4,126

Total (excluding zero hits)	122,168	6,264

A critical issue in any search service is the queries that result in an empty result set. Zero occurrences are known to be discouraging, even though they do not necessarily reflect usability problems. In COMPARA, as shown in figure 6, queries that produced zero hits amount to roughly half the total number of queries, and this figure has been fairly consistent throughout the existence of the corpus, although we seem to discern some improvement of late. More details about them are given in section 3.3.1 below.

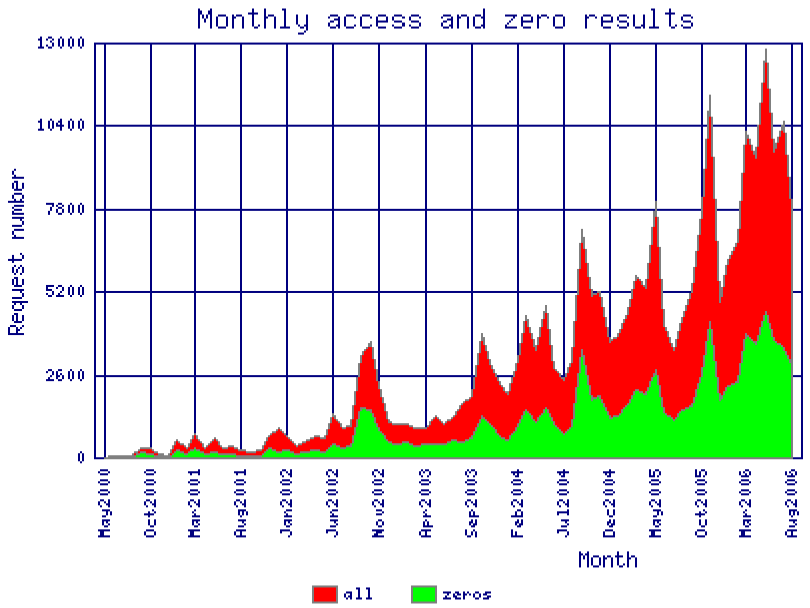


Figure 6: Zero results consistently amount to half the queries to COMPARA

3.3.1 Classification of zero results

In an attempt to assess the possible causes for such a large number of empty queries, a random selection of around one thousand queries returning zero hits (from the logs until September 2004) was analysed in terms of the following six categories:

I. Other language: This category was used to describe queries where search expressions in one language were used to look things up in the other language of the corpus. In other words, it was used to describe queries where users looked up words and expressions in Portuguese in the English part of COMPARA, and queries where users looked up words and expressions in English in the Portuguese part of the corpus. Even if the search expression existed in the corpus, users would be unlikely to get any results because they were looking it up in the other language of the corpus.[7]

II. Wrong syntax: COMPARA is encoded in the IMS-CWB system, so all queries must comply with its syntax. The category for “wrong syntax” was used to describe queries which were not in accordance with the IMS-CWB requirements. Badly-formed syntax excludes from the outset all chances of obtaining successful results.

III. Wrong syntax and other language: This category was used to classify queries that produced no results because of both the language and the syntax used. It is therefore a combination of categories 1 and 2.

IV. Nonsense and misspellings: This category describes queries without syntax and “other language” problems, but with nonsensical strings like *fjhkltr* and misspellings like **accidently, studpid* and *atravez.*[8]

V. Empty search: This category describes cases where users hit the search button without entering any query, and cases when they set up an alignment constraint without a query.[9]

VI. Well-formed but not found: This category describes queries where the part of the corpus being interrogated and the language used to interrogate it matched, the spelling and syntax were correct, and there were no nonsensical or empty searches, but still no results were returned. This category describes therefore well-formed queries that presented none of the procedure problems outlined in categories I to V.

The distribution of queries according to these categories is summarized in table 7.

Table 7: Overall analysis of randomly selected 988 "no results" queries to COMPARA

Type of problem	***f***	**%**
I - Other language	295	29.9
II - Wrong syntax	205	20.7
III - Wrong syntax and other language	64	6.5
IV- Nonsense and misspellings	76	7.7
V- Empty search	17	1.7
Sub-total	**657**	**66.5**
VI – Well-formed but not found	331	33.5
Total	**988**	**100**

Almost two-thirds of the queries returning no results could be traced back to problems accounted by categories 1 to 5, while around one third of the queries produced no results despite appropriate use of the corpus. The three most frequent categories – "other language", "wrong syntax" and "well-formed but not found" - are analysed in greater detail below.

3.3.1.1 Zero results because of wrong language direction

The 359 queries which involved searching for Portuguese expressions in the English part of the corpus, and English ones in the Portuguese part of the corpus (categories I and III) represented the single most frequent problem behind why users got no results.

As the procedure for selecting which language of the corpus to use is different in the two search forms available in COMPARA, it was important to find out whether one search form was producing better results than the other. COMPARA's Simple Search contains two separate boxes, one for entering searches within the Portuguese side of the corpus, and one for typing in searches within the English side. Entering the query and selecting the language is therefore a one-step procedure. In COMPARA's Complex Search, there is one box for entering the query, and a separate button for selecting in which language of the corpus the search is to be carried out.

As it turned out, both the one-step procedure of the Simple Search and the two-step route of the Complex Search proved to be problematic: of the 359 queries pertaining to categories I and III, 219 were carried out in the Simple Search and 138 were conducted using the Complex Search.[10]

On the surface, these results suggest that the one-step procedure of the Simple Search was the one that caused most problems. However, if one remembers that the Simple Search has been used more than twice as often than the Complex Search (see figure 3), then the two-step procedure of the Complex Search seems to be proportionally more problematic. One reason for this could be that, unlike the Simple Search, in the Complex Search there is a set default language: queries are by default automatically conducted in the Portuguese part of the corpus. And indeed, as shown in table 8, most “other language” problems occurring in the Complex Search can be traced back to the use of English queries in the default, Portuguese part of the corpus.

Table 8: Analysis of “other language” problems in the Complex Search

Type of problem	***f***	**%**
English search expressions in Portuguese corpus	109	79.0
Portuguese search expressions in English corpus	28	20.3
Other language in alignment constraint	1	0.7
Total	**138**	**100**

3.3.1.2 Wrong syntax

The second most frequent usability problem had to do with the query syntax, with categories II and III accounting for 269 queries of the sample. A closer look at this category revealed different types of problems, which are summarized in table 9.

Table 9: Analysis of query syntax problems

Type of problem	***f***	**%**
Quotation marks	140	52.0
Regular expressions	65	24.2
Quotations & expressions	3	1.1
Case sensitive	39	14.5
No diacritics	22	8.2
Total	**269**	**100**

The most frequent syntax problem had to do with quotation marks. Most quotation-mark problems occurred when users failed to understand that the IMS-CWB syntax requires the marks to be used around each separate element of a search string. Sometimes, however, users understood that, but failed to open or close the quotations marks, inserted too many marks, or left empty spaces inadvertently within them. Badly-formed regular expressions were the second most frequent syntax problem, but occurred only half as often. Many of them can be traced back to users attempting to apply the syntax of other systems to an IMS-CWB encoded corpus. For example, looking up *"dis*"* instead of *"dis.*"*, or *is+regarded* instead of *"is" "regarded"*.

Another recurring syntax problem occurred when users failed to use the '%c' command to make the query case insensitive. For example, users wrote queries in block capitals to look up words that they did not necessarily seem to want in block capitals, wrote the first word of search strings with a capital letter, even if they did not appear to be looking for sentence-initial position, and searched for proper names with small letters. Although the need to use the %c command in these cases was explained in the help file, few users bothered to look it up.

Not using diacritics while not using the '%d' command to make queries insensitive to diacritics was the next most frequent problem. Naturally, it only affected searches containing Portuguese words with accents, tildes and cedillas: for example, *"refem"*, *"nao"* and *"comecar"*. Again, the explanation in the help file regarding this command was by and large overlooked.

3.3.1.3 Well-formed but not found

The 331 queries that returned no results despite exhibiting none of the mechanical setbacks accounted for by categories I to V were then examined in greater detail, and it was possible to identify three major problems:

a. The search term was not found in the sub-corpus used (but exists in the corpus).
b. The search term was not found because it was unrealistic.
c. The search term was plausible but was not found at the time the query was carried out.

The first problem was observed in 50 of the 331 well-formed queries that returned no results (15.1%). Although the log files did not contain sufficient information to see exactly what sub-corpora had been used for each search, it was possible to find the search term sought when repeating the query using the entire corpus (version 6.0).

The problem of unrealistic queries was detected in 77 of the 331 well-formed queries that returned no results (23.3%). Two types of queries were classified as unrealistic. First, those that contained obviously technical terms (such as *greenhouse gas emissions*, *peer review mechanism*, *livres de nitrofuranos* and *hortifrutiganjeiro*). COMPARA contains only fiction texts, and it was deemed unrealistic to expect to find technical terms in it. The second type of unrealistic queries were those that returned no results or had just one single hit in a large monolingual corpus. Portuguese expressions not found in COMPARA 6.0 were tried out in CETEMPúblico and NILC/São Carlos, two large monolingual corpora of European and Brazilian Portuguese, with 180 million words and 32 million words respectively. English search terms not found in this same version of COMPARA were tried out in the BNC (100 million words) and in the Bank of English demo (56 million words that include 10 million words of American English). As version 6.0 of COMPARA contained just over two million words (one million Portuguese and one million English), it was considered unrealistic to expect it to contain expressions not found or found only once in much larger monolingual corpora. Thus the queries classified as unrealistic included unlikely sequences such as *mad honey* and *buzz kill,* and queries resulting from search strategies that are employed in search engines but fail to match textual corpora, such as omitting very frequent grammatical words (e.g., *assistir parto* instead of *assistir* ***ao*** *parto,* or *chefe redacção* instead of *chefe* ***de*** *redacção*).

The third problem observed, i.e., the search term was plausible but was not found at the time the query was carried out, affected 204 of the 313 well-formed queries that returned no results (61.6%). Despite COMPARA being already a sizable corpus insofar as parallel corpora are concerned, it must be remembered that its first public version contained only 65 thousand words. As the sample returning no results analysed dates back from the very beginning of the corpus, some of the queries returning no results in the early stages of the corpus may not necessarily return no results today. To check whether this may have indeed occurred, the 204 plausible and well-formed search expressions in the sample were tried out again in COMPARA's much larger version 6.0, and 80 of them (39%) no longer returned no results.

It should in any case be emphasized that zero results are not always discouraging nor do they always indicate a too little textual base. Sometimes, they may be significant and interesting, such as when one is checking for false friends and gets no hits, or when the results pointing out that "standard" translations, in the sense of Gellerstam (1986),[11] were not used. One should therefore be especially careful not to equate negative results in category VI with problems.

3.3.2 Classification of truncated results

We also analysed a random selection of cases (before September 2004) where the output was so large that, to protect the rights of copyright holders, COMPARA had to truncate the concordance presented to the user. Table 10 displays a first categorization of plausible causes for queries having such huge output.

Table 10: Why truncated results

Possible explanation	**f**	**%**
Empty search	25	12.5
Wrong direction in simple search	5	2.5
Very general non-lexical item	22	11.0
Very frequent lexical item	148	74.0
Total	**200**	**100**

Undoubtedly, the most common reason for truncation is the search for very frequent lexical items, such as auxiliaries, pronouns or prepositions. However, other factors concur to yield too many results. 15% of the cases could be due to user mistakes: selecting the wrong language direction when using the Simple Search interface, or issuing a null query. From this cursory perusal, however, it seems that most cases occur because users do focus on broad subjects, either inadvertently or because they are genuinely interested in them. Some examples of “very general non-lexical items” are suffixes, prefixes, adverbs ending in *-ly* and verbs with clitics.

4 COMPARA user sessions

Until this point, we have been treating every request to COMPARA as independent, which is obviously a gross simplification. We will here try to link queries under the concept of user session and see what else can be concluded.

4.1 Using logs to identify sessions and users

As mentioned in section 2.2, we did not look at the general navigation in all pages in COMPARA, but concentrated on user sessions as reflected by service-specific COMPARA logs, which are only activated by the user actually querying the corpus.

Corpus lookup is not an application like general Web search in two particular details: on the one hand, we expect users to take some time looking at the results and go on with their work; on the other hand, it should be possible that a lot of experimentation with the system takes place before users start working (solving their own goals). This means that we expect a larger number of queries per session than is usual in general purpose search engines.

Let us first tackle the question of user identification and session identification: We did not employ any complicated algorithms to define an individual session, such as those described by He et al. (2002). Rather, we used a simple heuristic: the same computer address on the same day was taken to reflect the same user.[12] A “session” within COMPARA is therefore defined as a number of questions posed by the same user (read: computer id) in the same date (read: day). We have then implemented two corrections to this simple definition: (a) merging two sessions from the same user that cross midnight (and whose joined duration does not extend six hours), and (b) marking as specially suspect sessions those which have such a temporal density of queries and such a sizable number of queries that they are bound to reflect a classroom environment.

Another considerably more complicated question emerges when attempting to identify different sessions by the *same* user in order to measure properties such as user faithfulness or user progress. Since users are purposefully not required to register (in order to make the resource maximally usable and unobtrusive), the only identification available is the IP address of the computer used to get access to COMPARA. This makes it difficult to know whether the same user is involved in different (temporally distant) transactions. In a nutshell, non-intrusive techniques are not good at measuring user faithfulness. Still, we present some data in tables 11 and 12, to be read with special care. Note that this data is gathered only from the COMPARA logs, i.e., from people who did ask questions to COMPARA, not from people who visited the COMPARA site but did not query the corpus.

Table 11: "Users" and sessions

Number of sessions	**Number of "users"**
At least 1	23,198
At least 2	2,305
At least 3	1,055
At least 6	392
At least 10	208
Average: 1.4 sessions per "user"	

Table 12: "User" faithfulness: common users in the periods considered: upper half, intersection of the years; lower half, period spanning from the Y year until the X year

Same users	2001	2002	2003	2004	2005	2006
2000	20	14	5	4	5	3
2001	-	127	38	119	97	27
2002	127	-	100	85	60	36
2003	13	100	-	105	76	38
2004	7	32	105	-	191	103
2005	3	18	36	191	-	292
2006	1	8	16	62	292	-

Intra-session behaviour, i.e., what a user does inside one session, yields considerably more reliable data. All sessions are analysed independently. Table 13 provides information on the number of sessions considered and the number of queries per session (single-query vs. multiple-query). Information on sessions ending in null results and on sessions returning positive hits is also provided. These overall results are sub-divided into "null results included" (sessions which include at least one query returning no results) and "null results alone" (sessions in which all queries yielded null results).

Table 13: Description of logs in term of user sessions

Kind of session	**Total**	**null results included**	**null results alone**
Number of sessions	33,026	24,542	9,535
Single-query sessions	9,983	4,372	4,372
Multiple-query sessions	23,043	20,170	5,163
Suspiciously active sessions[13]	110	108	0
Sessions ending in null-results	14,127	14,127	9,535
Sessions ending in positive hits	18,899	10,415	0

Tables 14 to 16 present sessions in terms of the requests they contain. While table 14 presents the session population at a glance, table 15 details sessions in terms of number of queries and table 16 and 17 classify them according to duration[14] and query density. However, care must be taken when interpreting sessions in terms of how long they last and how many queries are posed. As He et al. (2005: 358) note in an analysis of user behaviour in the context of interactive question answering, "performing more query iterations does not necessarily lead to higher accuracy, nor does it necessarily take more time".

Table 14: Description of sessions

Statistic	Value
Average number of queries per session	6.98
Median number of queries per session	3
Average number of null results per session	2.53
Average number of null results per multiple session	3.63

Table 15: Session size in terms of number of queries

Number of queries	Number of sessions
2	5,891
3	3,905
4	2,542
5-10	5,962
11-20	2,554
21-50	1,643
51-100	399
more than 100	147

Most sessions are short, both in number of queries and in duration. There are, however, cases with a large number of queries (even discounting the suspiciously large ones), as well as a sizeable number of sessions that went on for longer than one hour. It is not possible to know whether the user is analysing the results before issuing further queries, or spending time in between reading documentation or even doing searches in other corpora or on the Web.

We would be interested in knowing whether similar data could be obtained for other corpora on the Web, but being the first (to our knowledge) to publish this kind of data, we cannot benefit from comparable studies.

Table 16: (Multiple) session size in terms of duration

Time	Number of sessions
less than 1 minute	5,524
1-5 minutes	6,278
5-10 minutes	1,965
10-15 minutes	1,063
15-30 minutes	1,729
30-60 minutes	1,660
1-2 hours	1,469
2-3 hours	934
3-4 hours	353
more than 4 hours	2,068

Table 17: (Multiple) session size in terms of number of queries per hour (for the 17,152 sessions with more than two requests)

Queries per hour	**Number of sessions**
less than 1	684
1-5 queries/hour	2,229
5-10 queries/hour	1,541
10-20 queries/hour	1,721
20-30 queries/hour	1,178
30-40 queries/hour	936
40-50 queries/hour	827
50-60 queries/hour	710
60-70 queries/hour	645
70-80 queries/hour	582
80-90 queries/hour	487
more than 90 queries/hour	5,612

Having clarified what our working definition of session is and given a corresponding quantitative description, below are three analyses of user behaviour that are based on the session concept. Many further interesting questions could be pursued, but these will have to be left to later studies.

4.2 Impact of adding a help message in case of zero results

What led us to attempt to study the reasons behind null results (as presented in section 3.3.1 above) was the expectation that, if we could recover what users meant, we might find out (a posteriori) how the problems that they encountered could be avoided or solved.

If one cannot prevent users from making mistakes (which would obviously be the ideal solution, if feasible), nor can we automatically reconstruct what they wanted, one can at least yield fairly informative repair messages. Whether such informative repair messages are effective is another issue, to which we turn now.

Having noted that users consistently made syntax errors when attempting to recover a sequence of words (even though information on the appropriate syntax is available in the help file), an error message specially tailored to deal with this situation was implemented in December 2002 (version 2.2). We have investigated here whether this message actually made any difference by counting the percentage of successful recoveries after its implementation. By successful recoveries we mean: a user session with a zero result that ended with a non-zero result.

Table 18 gives the number of multiple sessions with zero results, and how many got a positive follow-up, i.e., sessions where zero results were followed by a more “comforting” answer, both before and after December 2002. Not all cases of 0 followed by non-zero are recoveries, in the sense that the user may have changed query in the middle of a session. In addition, the only null results that could be fixed were the ones due to faulty input (as seen in table 7, amounting to 66.5%).

Table 18: Recovery from zero results in multiple sessions

Period	Number of sessions	Sessions with 0 results	%	Sessions with recovery	%
2000-2002	3,528	2,397	67.9	1,451	60.5
2003-2006	20,077	15,690	78.1	10,399	66.3

Although a 5.8% improvement is not impressive, we expect that the error message has in fact helped users. However, even if habitual users learned from an error message pertaining to previous sessions, it is not guaranteed that the distribution of causes of zero results is constant over time. For this, an error evolution analysis must still be undertaken.

4.3 Reaction to too many occurrences

Too many occurrences may indicate that some query refinement is in order. According to table 6, 6.1% of the results were truncated. What do people do in that case? Do they simply repeat the query, are they happy, or do they try to get it broken down into more specific subcases?

Before September 2004, there were 988 sessions that yielded a truncated output, 432 of which received no follow-up. This means that in 43.7% of the cases, the users were apparently satisfied with a large number of answers and did not continue the session. To have a better idea of what happened in the remaining 556 sessions where users continued to query the corpus after obtaining truncated results,[15] a random selection of 200 queries producing a truncated output that were then followed by another query were analysed. The results are displayed in table 19:[16]

Table 19: Action after too many results

Kind of action	***f***	**%**
Query changed to something different	101	60.1
Query refined	19	11.3
Query refined but still truncated	8	4.76
Failed attempt to refine query (error)	8	4.76
Query repeated with different selection	8	4.76
Query repeated with request for different output options	6	3.57
Query repeated without changes	15	8.92
Query broadened	3	1.78
Total	**168**	**100**

The results in table 19 lead us to assume that the majority of users are content with a truncated answer, even though “query changed” may in a few cases be an artificial way of getting answers to the same questions via a different query. In 25.6% of the cases, the user tried to avoid truncation by either refining the selection or refining the search string. The cases covered by “Query repeated with request for different output options” are of two kinds: either the user changed interface (from Simple Search to Complex Search) to see whether more results would be provided (obviously in vain), or the user asked for distribution details: even though only a limited number of concordances could be seen, these users could at least find out which parts of the corpus contained them. This information is extremely relevant and shows that some users are very familiar with the capabilities provided by COMPARA.

On the other hand, when users issue exactly the same query twice, two possible explanations can be devised: they are checking whether the random output is repeated or is again random, or they got confused and simply tried again.

Finally, to generalize a search (instead of refining it) may be interpreted as a mistake, but it can also correspond to the wish to have an idea of a larger mass (COMPARA gives the exact number of matches, although it truncates the concordance).

4.4 The language option: chance, laziness or bug?

Since the DISPARA interface is strictly parallel, i.e., all HTML pages, error messages and result pages exist in the two languages, we also looked at the use of the English and Portuguese service as a parameter, after noticing the surprisingly high use of the English mode of interaction, mentioned in section 3.1 above.

Giving it some thought, we realized that, in the beginning of COMPARA's history, English might have been preferred because the default URL www.linguateca.pt/COMPARA/ gave the English welcome page. Therefore, to use Portuguese, one more click was necessary, and people might not even notice the option to switch languages.

In order to test whether this explanation was right, we changed the default welcome page to Portuguese in April 2004 to check whether English had been "chosen" by default, or whether this reflected genuine user preferences.[17]

If users preferred the English service because they did not know a service in Portuguese existed, one would expect the majority of interactions with COMPARA to turn to Portuguese. Looking at Figure 4 above, however, we note some difference after the change, but the amount of interaction in English did not significantly decrease. This is rather surprising, and may have several different concurrent explanations:

- long-time users were used to the "English" COMPARA and people in general don't like changes;
- teachers of English ask their students to use the English interface;
- most casual users of COMPARA arrive there by chance and don't speak Portuguese;
- most users are directed to COMPARA via a search engine (and if you look for COMPARA in Google, for example, what you get first is the English page[18]);
- people misguidedly turn to the English page to run queries in the English-Portuguese direction (instead of selecting the appropriate button in the form).

Only the last hypothesis can be easily tested by looking at the logs, it is also the one more closely related to COMPARA's design: if people turned to the English interaction language *in the interface* when they meant language direction *in the corpus*, the number of zero results in the interaction language English would be much higher (because it would include the cases where people were looking for English items in the Portuguese corpus). However, the number of zero results when the interaction language is English compared to Portuguese, measured until September 2004, does not lend weight to this hypothesis, since they are comparable: 16,520 vs. 15,626 zero results.

Also, it sounds improbable that a user changes language in the middle of a session (remember that a session is defined by a number of actual requests to COMPARA, not as a navigation sequence). Still, we looked at changes of the interface language within a session with more than one query, and found that the number of such changes is very small indeed (see table 20). However, they correlate highly with changes of interface mode as well, which seems to indicate that such users may be in an exploratory and/or teaching mode or that these numbers correspond to groups of students in a classroom with the same IP address.

Table 20: Change in language of interface per session

Interaction language	Number of sessions
Portuguese service alone	12,577
English service alone	9,587
English service changed to Portuguese	367
Portuguese service changed to English	210

More than one change	302

4.5 Do users progress in their interaction with COMPARA?

One possible, albeit limited, way of assessing user progress is by analysing intra-session moves from Simple Search to Complex Search, overviewed in Table 21. Our hypothesis was that users starting in the Simple Search and becoming more confident might go on to try the Complex Search mode, while the other kind of transition would reflect the user's giving up of trying to make sense of the complexity offered in Complex Search.

However, and given our discussion of Figure 5 above, things may be more complicated here as well. In any case, the number of shifts is unexpectedly high (occurring in 3,957 sessions), which prompts for a more detailed study of this kind of user behaviour.

Table 21: Change of search mode in 23,043 plural sessions

Kind of search interface	**Number of sessions**
Sessions with Simple Search only	15,745
Sessions with Complex Search only	3,285
Move from Simple to Complex Search	1878
Move from Complex to Simple Search	637
More than one change	1,442

5 Results and further work

Summing up, in this first investigation of user behaviour and query patterns based on log observation, we were able to detect some general usability problems as well as gather material for further research and development of COMPARA. We also got a much broader picture of what people out there are doing with COMPARA than any questionnaire would give us.

In fact, we identified (a) some possible failures in understanding the interface design, (b) some typical mistakes, and in some cases (c) unexpected behaviour that requires further study. We were also able to (d) assess the popularity of different options, which may guide us in further design.

In some cases, we have already taken action since this study was conducted (for example, by significantly improving the documentation, and by simplifying the way users can make queries case-insensitive or do without the use of diacritics), as the readers may confirm by actually trying out the system. A detailed study of the impact of those and other changes will have to be left for the future.

We have just scratched the surface of all there is to do in a serious user analysis of a computational service. Just by doing what was reported in the present paper, however, we considerably increased our knowledge of users and the uses the corpus is put to, and were able to come up, in some cases, with a set of suggestions for improvement, as well as plenty of material for further interesting research questions and to guide the future development of COMPARA. We hope, in addition, that our work can be inspiring for other corpus developers so that both tools and methodology can be reused and improved. Also, by presenting some quantitative data, we offer researchers who want to perform comparable analyses of their own services a basis for comparison.

We conclude this paper with a brief discussion of what is still not known and of what we wish to learn with respect to the use of COMPARA.

5.1 Higher-level research questions and the user class issue

The long-term objective of a user study is to answer considerably more relevant, higher-level questions for a corpus service, such as:

- Do users capitalize on corpus growth?
- Does their performance with COMPARA increase with use?

Neither of these questions is easy to answer, though. Even if we had managed to identify users across sessions unambiguously, which, as already mentioned several times above, we did not, these issues would remain very difficult to assess. In order to explain why, we have to invoke the notion of user class.

We hypothesise (and hope) that the COMPARA service on the Web has several kinds of relevant users:[19]

- translators, who use COMPARA as inspiration or oracle in their daily work;
- language students of either language, who use COMPARA to help them to learn the (other) language;
- contrastive researchers, who use COMPARA to test or discover differences and similarities between the two languages;
- NLP researchers, who use COMPARA to assess relevant generalizations to be used in their systems and to get materials to evaluate their systems;
- researchers in translation studies, who use COMPARA to study translation;
- language teachers of either language, who use materials extracted from COMPARA to prepare exercises and tests.

An attempt of semi-automatic discovery of “who is who” is well beyond our current capabilities, but we need an indication of which class a session belongs to in order to start addressing the higher-level questions outlined above.

Having made this clear, let us go back to the question of capitalizing on corpus growth. Given that the number of words in COMPARA has seen a steady increase over the years (as demonstrated by figure 1), it is possible that people who wanted to validate their studies with more data - an important methodological issue raised in Santos and Oksefjell (1999) - might issue the “same” query after corpus increase. On the one hand, this behaviour would be expected of a researcher with long-term goals and a particularly keen interest in one specific phenomenon, or in the case of scholars replicating the studies of others in COMPARA.[20] On the other hand, however, a translator or a student using the corpus like a bilingual dictionary would probably not be interested in repeating a query over time. To complicate matters further, a tutorial for COMPARA has been made available since August 2004 (Frankenberg-Garcia, 2004), and it is to be expected that the examples it contains will be repeated more often than by chance. Reliable statistical studies based on logs after September 2004 will have to remove the tutorial examples from the data set or at least consider them very carefully.

The second higher-level question raised is even more general and central to the usability quest: assessing whether, over time, users make fewer errors and/or bolder, more complex queries. However, it would be misleading to carry out a straightforward temporal analysis of this issue for the number of COMPARA users is constantly growing and new information and functionalities are continuously being added to the corpus. A further complication is that not all kinds of users are supposed to “progress” to other kinds of tasks. For example, an experienced translator who keeps interrogating COMPARA in Simple Search with steady satisfaction should not be considered as a failure. In contrast, a researcher doing serious contrastive analysis would probably need to learn to take advantage of the more powerful functionalities of the corpus in order to do a better job.

One of the future goals of mining COMPARA logs is the empirical validation (or discovery) of different user classes, who may require different functionalities and a different system behaviour. Although we are far from being able to develop reliable criteria to identify them automatically, we will start to develop heuristics for assigning a class membership for particular sessions, and possibly cross this information with site-navigation patterns so that we can engage in a finer study of the uses and problems of different kinds of users of COMPARA.

Acknowledgements

This work, as well as the development and maintenance of COMPARA, was (partially) supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI. We are grateful to the Linguateca team for help with the project development, more specifically to Luís Sarmento, Paulo Rocha, Luís Costa and Luís Cabral for help with graphical packages and the maintenance of the COMPARA site, to Rosário Silva for help with wording of the Portuguese interface and pertinent remarks about the site and some of its behaviour options, and to Nuno Seco for discussion on log processing. Finally, we thank Asbjørn Følstad for relevant usability references and advice.

References

[All URLs last checked 19 September 2006]

Bernardini, Silvia. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (eds) *Rethinking language pedagogy from a corpus perspective*. Frankfurt am Main: Peter Lang, 225-234.

Bianchi, Francesca & Elena Manca. (2006). Discovering language through corpora: Needed abilities and student difficulties in corpus analysis. Paper presented at the *Seventh International Conference on Teaching and Language Corpora* (TaLC 7), Bibliotheque National de France, Paris, 1-4 July 2006.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther Koenig (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. March 8, 1999 (CQP V2.2). Institute for Natural Language Processing, University of Stuttgart. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/

Dix, Alan, Janet Finlay, Gregory Abowd & Russell Beale. (1993). *Human-Computer Interaction*. Hillsdale, NJ: Prentice Hall.

Følstad, Asbjørn. (2007). Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work Support Systems. *International Journal of Human-Computer Interaction* **22** (1).

Frankenberg-Garcia, Ana. (2004). COMPARA English Tutorial. 28 September 2004. http://www.linguateca.pt/COMPARA/Tutorial.pdf.

Frankenberg-Garcia, Ana. (2005a). A peek into what today’s language learners as researchers actually do. *International Journal of Lexicography*, **18** (3), 335-355.

Frankenberg-Garcia, Ana. (2005b). A corpus-based study of loan words in original and translated texts. *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747-9398.

Frankenberg-Garcia, Ana. (2006). Raising teachers' awareness to corpora. Keynote lecture presented at the *Seventh International Conference on Teaching and Language Corpora* (TaLC 7), Bibliotheque National de France, Paris, 1-4 July 2006.

Frankenberg-Garcia, Ana & Diana Santos. (2003) Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, Silvia Bernardini & Dominic Stewart (Eds.), *Corpora in Translation Education* (pp. 71-87), Manchester: St. Jerome Publishing.

Gaizauskas, Robert. (1998). Evaluation in language and speech technology. *Computer Speech and Language*, **12** (4), 249-62.

Gellerstam, Martin. (1986). Translationese in Swedish novels translated from English. In Lars Wollin & Hans Lindquist (Eds.), *Translation studies in Scandinavia* (pp. 88-95), Lund: CWK Gleerup.

He, Daqing, Ayse Göker & David J. Harper. (2002). Combining evidence for automatic Web session identification. *Information Processing and Management,* **38**, 727-747.

He, Daqing, Jianqiang Wang, Jun Luo & Douglas W. Oard. (2005). Summarization Design for Interactive Cross-Language Question Answering. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck & Bernardo Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004)* (Bath, UK, 15-17 September 2004) (pp. 348-362). Lecture Notes in Computer Science 3491. Berlin, Heidelberg: Springer.

Helander, Martin, Thomas Landauer & Prasad Prabhu (Eds). (1997). *Handbook of Human-Computer Interaction*. Amsterdam: North-Holland.

Ivory, Melody Y. & Marti A. Hearst. (2001). The State of the Art in Automating Usability Evaluation of User Interfaces. *ACM Computing Surveys,* **33** (4), 470-516.

Jacobson, Ivar (1992). *Object-Oriented Software Engineering: A Use-Case Driven Approach*. Reading, MA: Addison-Wesley.

Jansen, Bernard J., Amanda Spink & Tefko Saracevic. (2000). Real life, Real users and real needs: a study and analysis of user queries on the Web. *Information Processing and Management,* **36**, 207-227.

Johns, Tim. (1997). Contexts: the Background, Development and Trialling of a Concordance-based CALL program. In Wichmann, Anne, Steven Fligelstone, Tony McEnery & Gerry Knowles (Eds.), *Teaching and language corpora* (pp. 101-15). London & New York: Longman.

Jones, Steve, Sally Jo Cunningham, Rodger McNab & Stefan Boddie. (2000) A transaction log analysis of a digital library. *International Journal on Digital Libraries,* **3** (2), 152-169.

Kennedy, Claire & Tiziana Miceli. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology,* **5** (3), 77-90.

Koch, Traugott, Koraljka Golub & Anders Ardö. (2005) Users browsing behaviour in a DDC-based web service: A log analysis. To appear in *Cataloging & Classification Quarterly*. Manuscript at: http://www.lub.lu.se/tk/publ/CCQ05-manuscript.doc.

Masand, Brij & Myra Spiliopoulou (Eds.). (2000) *Web Usage Analysis and User Profiling: International WEBKDD'99 Workshop (San Diego, CA, USA August 15, 1999), Revised Papers* [Lecture Notes in Artificial Intelligence 1836]. Berlin/Heidelberg: Springer.

Nielsen, Jakob. (1993). *Usability Engineering*. Boston, MA: Academic Press.

Noblitt, James S. & Susan K. Bland. (1991). Tracking the Learner in Computer-Aided Language Learning. In B. Freed (Ed.), *Foreign Language Acquisition Research in the Classroom*, pp. 120-131. Lexington, MA: D.C. Heath.

Santos, Diana. (1998). Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. In Antonio Rubio, Natividad Gallardo, Rosa Castro & Antonio Tejada (Eds.), *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), Vol. 1 (pp. 475-481). ELRA.

Santos, Diana. (2002). DISPARA, a system for distributing parallel corpora on the Web. In Elisabete Ranchhod & Nuno J. Mamede (Eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)* (pp. 209-218). LNAI 2389, Berlin, etc.: Springer.

Santos, Diana. (2006a). Breves explorações num mar de língua. *Ilha do Desterro,* **50**.

Santos, Diana. (2006b). Corpora. Primeira Escola de Verão da Linguateca, Universidade do Porto, 13 July 2006, http://www.linguateca.pt/escolaverao2006/Corpora/CorporaEscolaVerao.pdf.

Santos, Diana & Caroline Gasperin. (2002). Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In Manuel González Rodríguez & Carmen Paz Suárez Araujo (Eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002) (pp. 597-604). ELRA.

Santos, Diana & Susana Inácio. (2006). Annotating COMPARA, a grammar-aware parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1216-1221. ELRA.

Santos, Diana & Signe Oksefjell. (1999). Using a Parallel Corpus to Validate Independent Claims. *Languages in contrast,* **2** (1), 117-132.

Santos, Diana & Paulo Rocha. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001) (pp. 442-449), ACL.

Santos, Diana & Luís Sarmento. (2003). O projecto AC/DC: acesso a corpora / disponibilização de corpora. In Amália Mendes & Tiago Freitas (Eds.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002)* (pp. 705-717). Lisbon: APL.

Sullivan, Terry. (1997). Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files. *Proceedings of the 3rd Conference on Human Factors and the Web* (Boulder, CO, June 1997), http://www.pantos.org/ts/papers/rrr.html.

Woolls, David. (2000). From purity to pragmatism; user-driven development of a multilingual parallel concordancer. In Simon Botley, Tony McEnery & Andrew Wilson (Eds.), *Multilingual corpora in Teaching and Research* (pp. 117-133). Amsterdam and Atlanta: Rodopi.

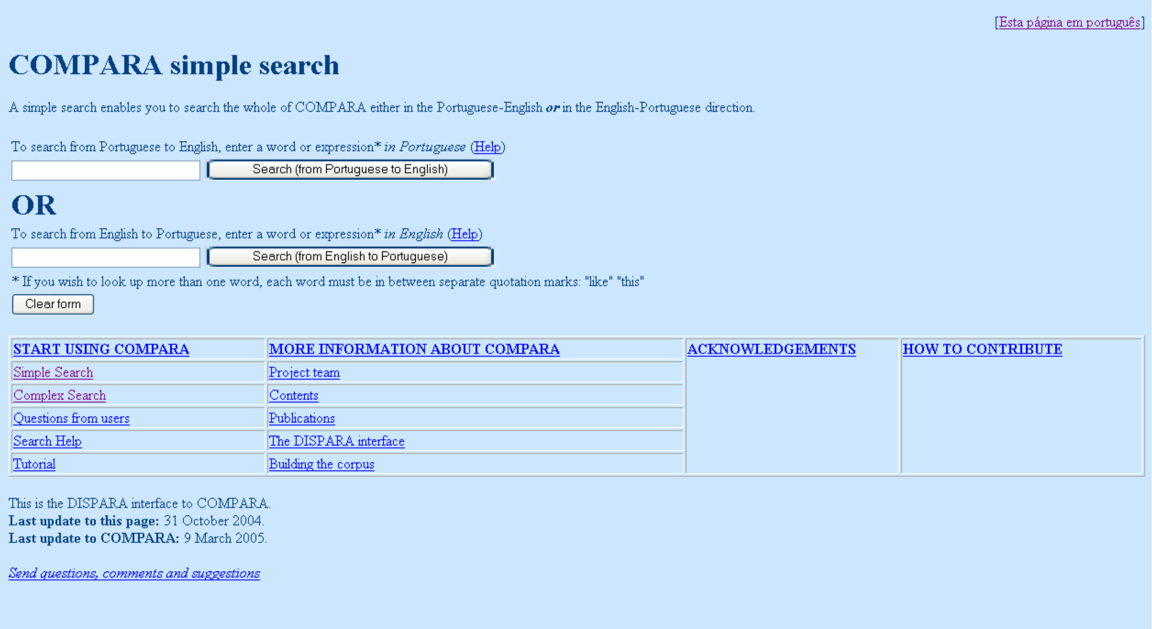
Appendix 1. Short history of COMPARA

Date	Event
May 2000	First version of COMPARA
September 2000	Introduction of Simple and Complex Search modes
September 2002	Start of version control
November 2002	Change of COMPARA's URL
December 2002	Insertion of a detailed message in case of zero results
April 2004	Change of default page for COMPARA
June 2005	First annotated version made publicly available

Appendix 2. Example of COMPARA specific logs

```
++++++++++
Thu Sep 30 22:59:39 CEST 2004
 82.154.17.197
lingua  eng
simples sim
concordancia   on
palavra_port
palavra_ing    "why" %c
accao   Search (from English to Portuguese)
----------
Resultados     772
++++++++++
Thu Sep 30 23:02:36 CEST 2004
 82.154.17.197
lingua  eng
complexa       sim
accao   Submit query
corpus  COMPARA_ING
palavra_port   "resum.*"
palavra_ing
quandoori      depois
dataori
quandotrad     depois
datatrad
concordancia   on
----------
Resultados     17
++++++++++
Thu Sep 30 23:30:42 CEST 2004
 200.216.152.32
lingua  port
simples sim
concordancia   on
palavra_port
palavra_ing    "thankful" "for"
accao   Pesquisar (de inglês para português)
----------
Resultados     3
++++++++++
```

Appendix 3: Screenshots of Simple Search and Complex Search pages



[Esta página em português]

COMPARA simple search

A simple search enables you to search the whole of COMPARA either in the Portuguese-English ***or*** in the English-Portuguese direction.

To search from Portuguese to English, enter a word or expression* *in Portuguese* (Help)

Search (from Portuguese to English)

OR

To search from English to Portuguese, enter a word or expression* *in English* (Help)

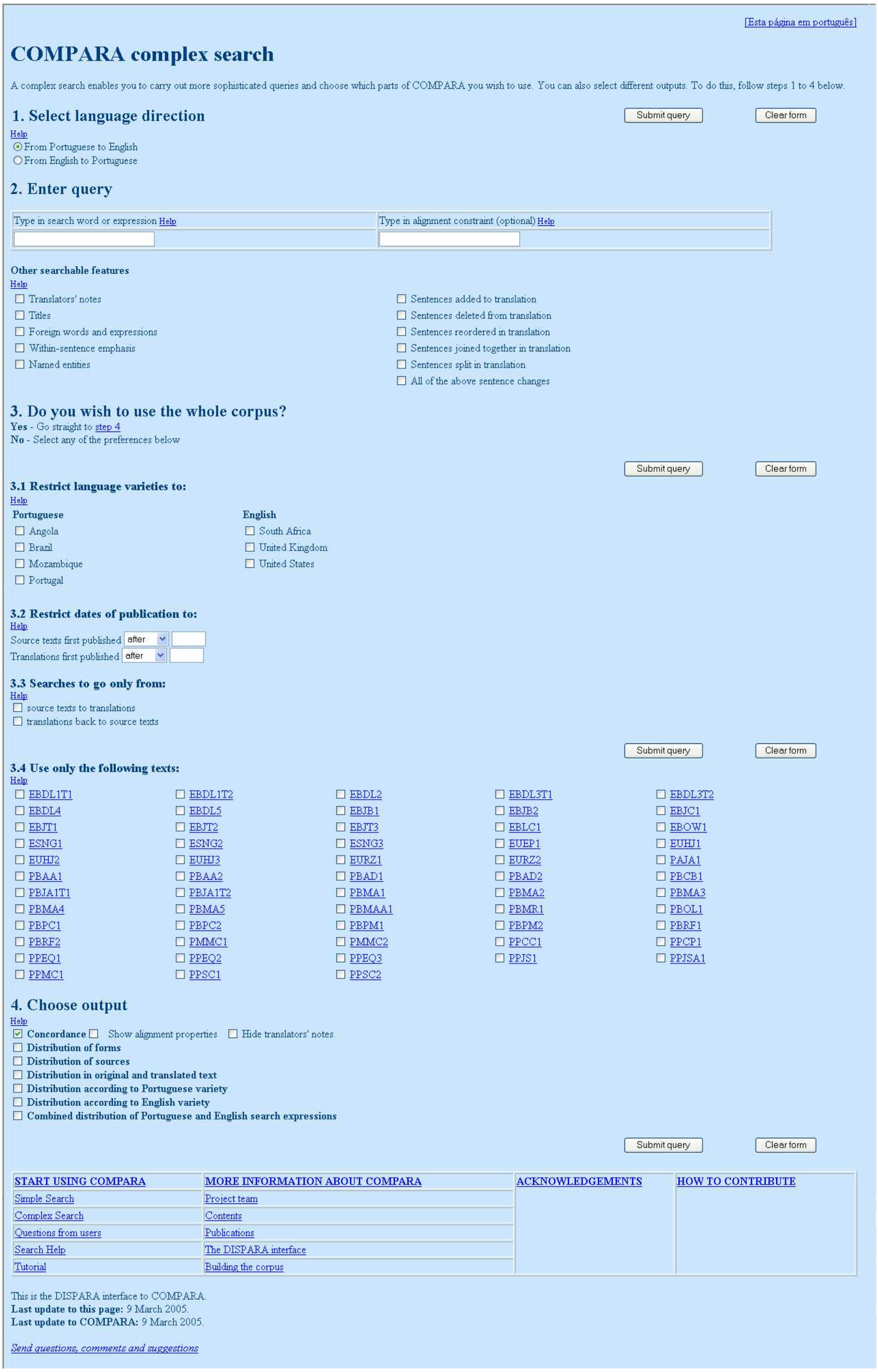
Search (from English to Portuguese)

* If you wish to look up more than one word, each word must be in between separate quotation marks: "like" "this"

START USING COMPARA	MORE INFORMATION ABOUT COMPARA	ACKNOWLEDGEMENTS	HOW TO CONTRIBUTE
Simple Search	Project team		
Complex Search	Contents		
Questions from users	Publications		
Search Help	The DISPARA interface		
Tutorial	Building the corpus		

This is the DISPARA interface to COMPARA.
Last update to this page: 31 October 2004.
Last update to COMPARA: 9 March 2005.

Send questions, comments and suggestions



[Esta página em português]

COMPARA complex search

A complex search enables you to carry out more sophisticated queries and choose which parts of COMPARA you wish to use. You can also select different outputs. To do this, follow steps 1 to 4 below.

1. Select language direction

Help

● From Portuguese to English
○ From English to Portuguese

2. Enter query

Type in search word or expression Help	Type in alignment constraint (optional) Help

Other searchable features

Help

□ Translators' notes	□ Sentences added to translation
□ Titles	□ Sentences deleted from translation
□ Foreign words and expressions	□ Sentences reordered in translation
□ Within-sentence emphasis	□ Sentences joined together in translation
□ Named entities	□ Sentences split in translation
	□ All of the above sentence changes

3. Do you wish to use the whole corpus?

Yes - Go straight to step 4
No - Select any of the preferences below

3.1 Restrict language varieties to:

Help

Portuguese	**English**
□ Angola	□ South Africa
□ Brazil	□ United Kingdom
□ Mozambique	□ United States
□ Portugal	

3.2 Restrict dates of publication to:

Help

Source texts first published after
Translations first published after

3.3 Searches to go only from:

Help

□ source texts to translations
□ translations back to source texts

3.4 Use only the following texts:

Help

□ EBDL1T1	□ EBDL1T2	□ EBDL2	□ EBDL3T1	□ EBDL3T2
□ EBDL4	□ EBDL5	□ EBJB1	□ EBJB2	□ EBJC1
□ EBJT1	□ EBJT2	□ EBJT3	□ EBLC1	□ EBOW1
□ ESNG1	□ ESNG2	□ ESNG3	□ EUEP1	□ EUHJ1
□ EUHJ2	□ EUHJ3	□ EURZ1	□ EURZ2	□ PAJA1
□ PBAA1	□ PBAA2	□ PBAD1	□ PBAD2	□ PBCB1
□ PBJA1T1	□ PBJA1T2	□ PBMA1	□ PBMA2	□ PBMA3
□ PBMA4	□ PBMA5	□ PBMAA1	□ PBMR1	□ PBOL1
□ PBPC1	□ PBPC2	□ PBPM1	□ PBPM2	□ PBRF1
□ PBRF2	□ PMMC1	□ PMMC2	□ PPCC1	□ PPCP1
□ PPEQ1	□ PPEQ2	□ PPEQ3	□ PPJS1	□ PPJSA1
□ PPMC1	□ PPSC1	□ PPSC2		

4. Choose output

Help

☑ **Concordance** □ Show alignment properties □ Hide translators' notes
□ **Distribution of forms**
□ **Distribution of sources**
□ **Distribution in original and translated text**
□ **Distribution according to Portuguese variety**
□ **Distribution according to English variety**
□ **Combined distribution of Portuguese and English search expressions**

START USING COMPARA	**MORE INFORMATION ABOUT COMPARA**	**ACKNOWLEDGEMENTS**	**HOW TO CONTRIBUTE**
Simple Search	Project team		
Complex Search	Contents		
Questions from users	Publications		
Search Help	The DISPARA interface		
Tutorial	Building the corpus		

This is the DISPARA interface to COMPARA.
Last update to this page: 9 March 2005.
Last update to COMPARA: 9 March 2005.

Send questions, comments and suggestions

[1] For copyright reasons, extracts are generally 30% of a text, and are consequently not all the same size.

[2] In fact, what users leave as "identification" is the IP address of the computer they are using, and there is no univocal correspondence between different users and different IPs: The same computer may be used by many people, and the same users may have a different IP number when connecting at different times, especially if they are connected to the Web through an ISP or their institution is protected by a firewall.

[3] It is possible, of course, to evaluate these different parts separately, especially when corpus texts are distributed as raw text, without a corpus encoding system or dedicated interface. In Santos and Rocha (2001) and Santos and Gasperin (2002) some preliminary evaluation of bare corpora is presented, but without special focus on usability.

[4] These tools were developed in the context of the AC/DC project (Santos and Sarmento, 2003) and are currently in use to display visits to the Linguateca site.

[5] Possibly one query or so has been issued per version, to test whether COMPARA was working after installation of a new version, or to revise the automatic computation of alignment type in some complex cases, see Santos (2002), but they are not statistically relevant. Conversely, the first author in the scope of her research has also often posed non-trivial queries to the development version of COMPARA for convenience reasons, but these queries have not been taken into account in the numbers discussed in the present paper.

[6] This option and the following one, which presuppose grammatical annotation of the corpus, are so far only available for Portuguese; see Santos and Inácio (2006) for more information.

[7]It must be noted that, although most queries conducted in the "other language" signify that the user has made a mistake, there is one important exception. Namely, it is quite legitimate for a user to search for a Portuguese word in the English part of the corpus (or vice-versa) if the user is interested in finding out if certain words have been left untranslated. A search for the English word *sitcom* in the Portuguese part of the corpus, for example, returns 25 hits in COMPARA 6.0. However, a more probable way to find out if any words have been left untranslated would be to use the option for looking up words classified as foreign (an option available in the Complex Search interface since September 2000), see Frankenberg-Garcia (2005b).

[8] Words without diacritics (e.g. *nao, avancar, ninguem*) were not considered as misspellings, for the IMS-CWB syntax allows users for this possibility. Whether the user was expecting a behaviour like the one of major search engines, or was genuinely interested in the question of whether the corpus contained such spelling errors, we have no way to tell except if we inspected their subsequent query behaviour, which we did not.

[9] In fact, in our original design it was not clear what was the semantics of an empty query if other options were selected, as "Translation notes" or "Sentence split by translation", and this has had different implementation strategies: either alignment units or words have been returned at different times. So leaving the word search empty is not necessarily an error from the user point of view. But if absolutely nothing had been selected by the user, the system would return an error message and zero results.

[10] The two search forms have been available since the corpus was first announced. A couple of queries carried out in a test phase prior to the existence of the Simple and Complex Search could not be classified according to either one of these forms.

[11] Gellerstam (1986) points that uses of Swedish *ländа* to express English *arrive* are marks of translationese instead of good translation practice.

[12] This may bring problems if a modem user chooses to disconnect from the Internet in the middle of a study. Another interpretation could be that the user actually had more than one session with COMPARA in a single day.

[13] Arbitrarily defined as sessions with more than 30 queries and more than 90 queries per hour. Not that these sessions were not removed from the data presented anywhere else in the paper.

[14] We compute duration as the temporal distance between the first and last request in a given session. A user must spend some time appreciating the last results, presumably the best ones he got, but we cannot estimate this time.

[15] Note that we did not consider the 275 cases where truncated output was due to an empty query plus subcorpus selection, which leads COMPARA to produce a random selection of translation pairs, since in that case the next request cannot be meaningfully interpreted. We expect that some of these requests were due to mistakes, while others were issued to have an idea of the subcorpus at stake.

[16] Due to a problem in the logs with concurrent searches, it was not always possible to identify the search which led to the truncated result. Therefore 32 cases had to be excluded from the detailed analysis.

[17] Later on, and as a consequence of the Linguateca portal growing larger, bringing the need for an increase in the homogeneization of the services offered, www.linguateca.pt/COMPARA/ started in May

2004 to point to an even more Portuguese-inspired page, which also lists, on the left handside, resources other than COMPARA that are available.

[18] If you don't select the search language, that is, if you don't look for pages in Portuguese.

[19] By non-relevant users we mean (1) people who come to COMPARA by chance and try to use it as a search engine, and (2) colleagues interested in looking at COMPARA to mimic its functionalities for e.g. other language pairs. Even though the latter are most welcome, they are not ultimately interested in *using* the corpus, so their interaction patterns cannot be considered relevant.

[20] It must be noted that COMPARA is still young, so there are not yet many published studies based on COMPARA that could be validated during the time the logs refer to.