

Corpus analysis for indexing: when corpus-based terminology makes a difference

*Débora Oliveira*¹

*Luís Sarmiento*¹

*Belinda Maia*¹

*Diana Santos*²

¹Linguatca node at FLUP
University of Porto, Portugal

²Linguatca
SINTEF ICT, Norway

dmoliveira@letras.up.pt

las@letras.up.pt

bmaia@mail.telepac.pt

Diana.Santos@sintef.no

Abstract

This paper describes preliminary work in corpus-based indexing of a sizeable specialized Web portal, with (comparable, but not parallel) information in Portuguese and English. The interdisciplinary work involved illustrates the urgent need to create greater cooperation between information retrieval and corpus-based terminology.

The aim of the work described is twofold: to provide a case for terminology-based search engine deployment while improving the availability of the information in a specific Website, and to suggest further measures in order to successfully marry IR and corpus-based terminology.

After presenting the Corpógrafo, a fairly mature Web-based environment for terminology work which includes information extraction capabilities such as named entity recognition and semi-automatic harvesting of definitions and semantic relations, we describe the practical task we wanted to deal with, namely, improving Linguatca's search system over its website by adding properly chosen index terms.

We begin by a general introduction to the portal, with a first analysis of the current state and limitations of the search engine, Busca, together with a study of its users, based on the logs. Then we explain the assumptions underlying our attempt to improve the efficiency and usefulness of dedicated search engines with a terminology-inspired methodology.

Finally, we describe the empirical work, providing a quantitative description of the terminological units detected from a qualitative and quantitative point of view, and suggesting how the work can be employed to give Busca more intelligent search capabilities.

We finish the paper by emphasizing the need to build bridges between IR and terminology in the training phase of terminologists.

1. Introduction

Interdisciplinary cooperation is not for the faint-hearted, and it requires a great deal of mutual understanding, tolerance and mental gymnastics to make it work properly. ‘Corpus Linguistics’ is no stranger to this problem, as an analysis of this conference programme will show. The authors of this paper combine a variety of talents in computer science, computational linguistics, terminology and translation, and much has been learnt from the experience of trying to coordinate objectives.

Theoretically, it would seem to make good sense to bring information retrieval and corpus-based terminology work together for the mutual benefit of both areas. However, this means bringing together computer experts in sophisticated word-crunching with a discipline that traditionally aims at linguistic precision, while simultaneously trying to reconcile this aim with the dynamic nature of concepts and the terms that represent them. If we add to this mixture of interests those of the ‘clients’ searching a particular website for information, we create an ambitious framework in which to conduct research.

The Corpógrafo, the Web-based environment for corpus-based terminology research developed in Linguateca, provides the potential for creating special domain corpora, semi-automatic term extraction and accelerated preparation of terminology databases. It is the product of brainstorming between computer science, natural language processing, and terminology and translation pedagogy. It can, and does, clearly accelerate the terminologist’s work considerably and its potential in this area is already being proved. It also aims to go further and provide semi-automatic extraction of definitions and semantic relations although, of course, much depends on the type of texts available in the domain for this to be successful.

However, even expert terminologists do not always find it easy to define semantic relations and the resulting ontologies, even when working ‘manually’ with experts in the time-honoured way. Information science, on the other hand, struggles to circumvent this complex process by providing increasingly ingenious ways of extracting and processing information from either raw or tagged text.

Information retrieval and terminology both suffer from a phenomenon that is common to other areas: a lack of synchronization between a) the experts’ view of what is needed; and b) what the ‘general public’ actually needs, or thinks it needs. In the world of technical communication, for example, this has given rise to the profession of ‘technical writer’, a person who is trained to write texts that explain technical matters in a way the ‘general public’ can understand.

The analysis of Linguateca’s search engine, Busca, and the logs of its usage, clearly shows that the ‘general public’ includes a wide variety of people, and the requests made of Busca show that there is no easy way to solve everyone’s problems. Apart from anything else, the Linguateca site contains a wide variety of material – language resources, tools, catalogues of useful information, a large bibliography and plenty of full texts – and devising a common search function for all this is

probably impossible. As we shall see, using corpus based terminology will improve one type of search tool, but not all.

2. Corpógrafo: a robust Web-based terminology environment

Corpógrafo (Sarmiento et al., 2004) is a web-based environment that allows users to perform a wide variety of linguistic and knowledge engineering studies using their own personal corpora. Corpógrafo provides a broad set of tools that range from simple concordance to semi-automated ontology building from text. Corpógrafo does not require any programming skills and helps the users in all steps of these studies, starting from the compilation and pre-processing a personal corpus to the export of results in XML files. Each user has his/her own personal area on Corpógrafo web server and may compile a process his/her own corpora without the need for installing any software (for a more detailed description of Corpógrafo please refer to the corresponding poster being presented in the same conference).

One of the most useful features of Corpógrafo is a rich set of tools for extracting terminology, definitions and semantic relations among terminological units. These tools speed up significantly the task of compiling specific domain terminology, and glossary (or ontology) building. Corpógrafo is able to propose to the user lists of terminological units, definitions and possible relations among such terminological units, which the user may then validate and store in personal databases. This is in fact one of the most relevant points about the Corpógrafo: it helps the user in course of a specific task, and tries to reduce the amount of effort and time needed for accomplishing that task. Therefore, Corpógrafo relies on a set of robust semi-automatic methods for terminological extraction (Sarmiento, 2005) and definition extraction and semantic relations identification, instead of aiming at fully-automated methods that are too difficult to develop with present technology and resources. On the other hand, these semi-automated methods have enough precision to present the user with good options that should not take too much time to validate.

Corpógrafo has been under intense development for more than two years and has reached a fairly mature state of development. We have recently re-implemented most of the basic infrastructure of Corpógrafo to use a standard database system, which is now responsible for executing most of the pattern matching procedures. We are also in the process of adding a named-entity recognizer for Portuguese to the set of available tools, as well as additional search tools over the entire Portuguese Web.

Corpógrafo has been extensively used by more than a hundred undergraduate and graduate students in Portugal and Brazil. However, we are now interested in applying Corpógrafo to some real world needs in order to push its development even further. The present study was seen as a good opportunity for this because it could impose new practical needs that would not have been predicted otherwise and it did indeed impose those needs, especially some regarding terminology analysis. Most of the statistics presented here were produced automatically by Corpógrafo using new extensions that were added because of specific needs of the study. Additionally, it provided an opportunity for general debugging of the system as well as for improving the user interface.

3. The problem to be solved

Linguatca is a distributed language resource centre for Portuguese which aims at contributing to the quality of Portuguese language processing. As part of its activities, Linguatca has maintained an increasingly large website at <http://www.linguatca.pt> since mid 1998. The main goal is to provide a comprehensive starting point for anyone interested in studying or developing research about Portuguese. In the site several on-line resources (corpora, tools, publications, etc) produced by Linguatca, as well as other similar resources produced or maintained by other researchers, are made available to the general public.

In the beginning, there were only a few resources listed in the Linguatca site, as can be appreciated in Oksefjell & Santos (1998), and the website's search engine, Busca, started as a small functionality whose purpose was to incrementally make use of the structure already present in the pages. General care was given to "Search for particular people" (Person Search), and a very simple keyword search (Free-text Search), weighted simply by the occurrence of the word in a title or not, was implemented as a 3-month student assignment by Tom Funcke.

When the site began to grow and a system for maintaining the catalogue was developed by Paulo Rocha (see Santos, 2000 for an overview), some further development took place in the search engine, in order to achieve higher integration with the rest of the site. A cache was added, processing of rtf, ps and pdf files was included, and the whole system was rewritten to use CQP (Christ et al., 1999) as the underlying search engine (thus using the same technology for searching the site and searching the corpora in the AC/DC project). In this step, weighting was dropped, but speed and robustness were considerably increased. The Person Search then became independent of the Free-text Search.

Two years later, another kind of search was implemented: the search over our publication database (Publication Search), felt to be necessary, given the growing size of publications listed in our site. Although the same label, Busca, was used, it names in this case a traditional database search and is not, in any way, related to the rest of Busca. In other words, although from a user's point of view Busca offers three kinds of search, from an implementation point of view they are completely separate. In the present paper we will be solely concerned with Busca as a Web search engine over specialized pages in a particular domain. No further references will be made either to the Person Search or to the Publication Search.

It should be made clear, however, that the ongoing work reported in the present paper gave origin to a fourth kind of search, which we may call "Search by terminological units and/or named entities", that still coexists with the present "Free-text Search", where all words are treated as keywords, no matter what their frequency, distribution or grammatical category. Figure 1 shows the options a user gets when selecting Busca.

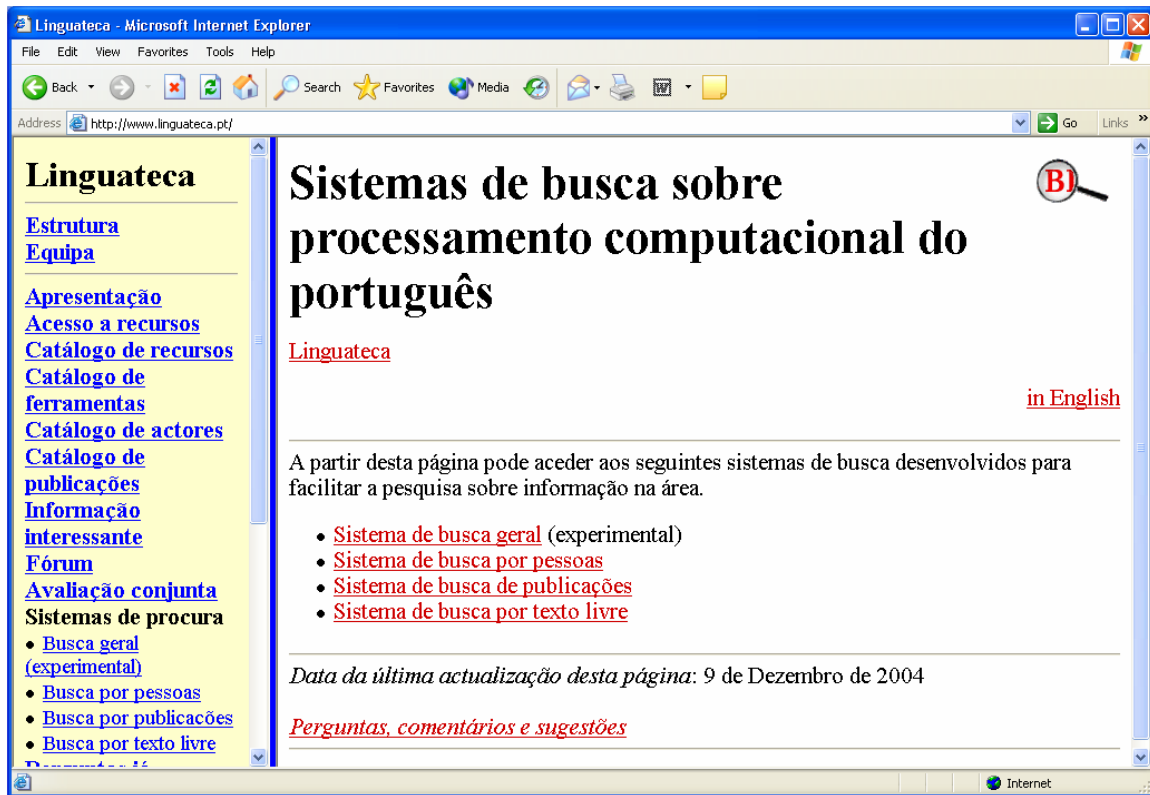


Figure 1- The search possibilities currently offered by Busca

Busca has not been successful in helping users find the appropriate resources for several reasons, some technical, others inherent to Linguateca's website and still others inherent to any information retrieval system's current architecture. There are many problems with word-indexed search systems, which are not specific to any particular search engine, but which have motivated considerable research in IR in general and in the improvement of Web search systems, in particular. We will describe some of the problems here with the help of Busca.

First of all, our website is not a uniform document collection. Currently the site holds around 1300 web documents and points to approximately 2500 external links, which makes it difficult for the site maintainers to keep simple navigation mechanisms. The site stores very different types of resources, including many on-line tools, each with its own specific web interface and on-line documentation. Each of these resources would in principle require different indexing techniques and relevance ranking.

Then, the users of our site range from highly specialized researchers to undergraduate students or even occasional users, a heterogeneous target audience which greatly complicates the development of a consensual information structure that may suit every type of user.

Now, and this is the reason why we started working on this problem in the first place, Busca did not make use of any knowledge about the domain to improve performance. Our hypothesis was that such knowledge could be used for indexation and ranking purposes, as well as to help users reformulate their search query when needed.

For example, Busca was not sensitive to domain specific terminological units (TU) and therefore was not able to differentiate between a relevant keyword of the domain at stake and a completely irrelevant general language word. In fact, Busca did not even perform any type of term weighting such as the conventional tf-idf weighting to improve result ranking. A slightly ill-formed query containing a general language word (e.g.: “system”) will probably put Busca off the right trail. Such performance becomes frustrating for the user after a few tries.

Furthermore, Busca is not able to make use of the knowledge provided by the way a query is formulated, i.e. Busca does not make use of pragmatics. In other words, Busca does not know what a search query reveals about what a user expects to find. For example, if a user searches for “corpora” using Busca, he is probably looking for on-line access to corpora rather than publications about corpora. However, for a search word such as “corpora”, which is very frequent over the entire site, Busca would retrieve many unrelated documents with a large number of occurrences of this word, which would hardly suit the users’ needs.

An interesting hypothesis that we consequently studied was whether Busca could be provided with simple heuristics that helped distinguish between different search strings and consequently different types of results.

4. Studying the logs

We started our work by conducting a simple examination of Busca logs between January 2003 and April 2005. Following Jansen et al. (2000) we analyzed the logs based on query frequency and length (number of words). Several of the queries logged had been made by our own research team so, in order to prevent biased data being included in the analysis, we excluded all possible internal requests from the study, which left us with 1527 queries.

The top 20 most frequent search strings are listed in table 1, and a quick look will show a very irregular usage pattern: while part of these queries are clearly related to the site’s content, such as “corpus”, “verbos”, “linguagem natural”, many other are either not related at all with the content of the site, or seem to be incompletely or incorrectly formulated. For example, the top three search-strings seem to be incomplete terminological units, suggesting that the simple knowledge of terminologically relevant units could be helpful in identifying these frequent cases. Table 2 shows the top 20 search string, excluding single-worded ones.

As far as repetition is concerned (Figure 2), these 1527 queries result in 1098 different search strings. It is possible to observe that the vast majority of the queries (76%) were tried only once and that the difference between queries occurring once and twice is particularly high (approximately 60%). Still, a significant number of queries (about 10%) were repeated 3 or more times.

We also studied the length of queries according to the number of words. Not surprisingly, the most common search strings are single-worded (54%), followed by two-word search strings (22%), such as for example “linguagem natural” (5), “floresta sintáctica” (3), or “dicionário técnico” (3). The resulting distributional plot (Figure 3) shows a reasonably soft descending profile with a significant number of three or more words search strings (ex: “processamento de texto” (3) or “textos com conectores” (2), “verbos de movimento”(2), “redação coerencia e coesão” (4), “análise sintáctica

dos adjetivos” (2)). It should be noticed that some of the multi-word search strings include expressions such as “o que é...” (“what is..”), “quais os tipos de...” (“what types of...”), “porquê...” (“why...”), “gostaria de traduzir...” (“I would like to translate...”), which may reveal some inexperience from the users or simply that they are expecting to find a question-answering system.

search string	#
Variações	10
Adjunto	9
Cabeça	8
Verbos	7
Corpus	5
corpus da folha de são Paulo	5
linguagem natural	5
Peniche	5
registros do que é Conjunções coordenadas	5
Sexo	5
Tesouro	5
Tradução	5
Trail	5
About	4
Adjetivos	4
Admir	4
Árvore	4
Autor	4
Concordância	4
Consultoria	4

Table 1 - Top 20 most frequent search strings

search string	#
corpus da folha de são paulo	5
linguagem natural	5
Registros do que é Conjunções coordenadas	5
creme de legumes	4
ele é nada mais nada menos que um idiota	4
há momentos	4
língua portuguesa 7% AA série	4
o cortiço	4
redação coerência e coesão	4
signo linguístico	4
Vanguarda europeia	4
verbos irregulares	3
adjunto adnominal	3
setem publico um milhao de palavras	3
comparable corpora	3

concordancia verbal	3
dicionário técnico	3
emprego do artigo	3
ensino%2C portugues%2C lingua estrangeira	3
floresta sintactica	3

Table 2 - Top 20 most frequent search strings, excluding single-word search strings

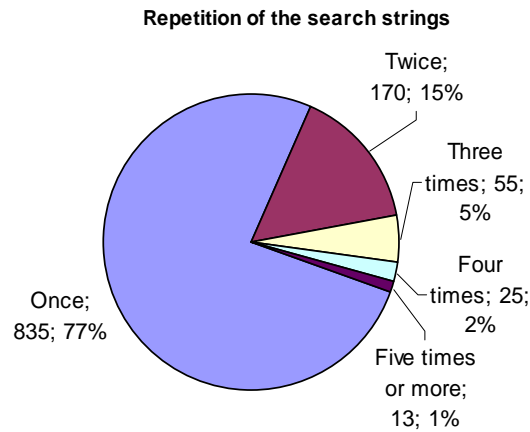


Figure 2 – Search string repetition in Busca’s logs

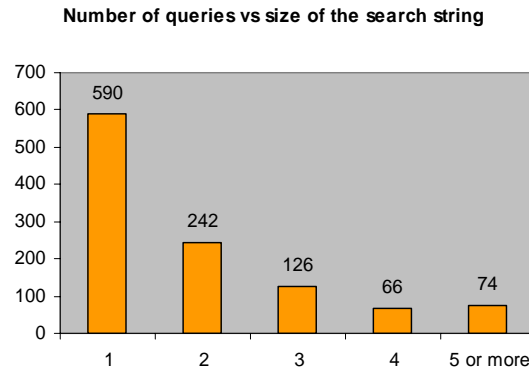


Figure 3 – Distribution of search strings according to size (in number of words)

Another interesting fact that it was possible to extract from the logs is that almost all search strings (94.8%) are singular. The top 10 plural forms are listed in table 3, and they tend to be mainly (complete or incomplete) terminological units related with specific issues of grammar or linguistics, such “verbos” (verbs), “adjetivos” (“adjectives”), “fonemas” (“phonems”), etc.

search string	#
Variações	10
Verbos	7
registros doque é Conjuções coordenadas	5
Adjetivos	4
bolsas	3
Corpora	3

Fonemas	3
Nomes	3
Proverbios	3
verbos irregulares	3

Table 3 - Top 10 most frequent plural form search strings

4.1 Complementing results with Google search strings

Since the Busca logs amounted to only 1527 queries, we decided that it would be useful to obtain more information from our Apache web server logs. We were particularly interested in knowing the keywords used to get to our web site because they might also be the words a user would type in Busca. From the Apache web logs we were able to extract 106,375 requests coming from Google. The most frequent search strings found are displayed in Table 4. As expected, the search string that brings most users to Linguateca's site is "Linguateca". The following search strings are mostly related with general language resources that people usually look for on the web, such as for example "dicionario ingles portugues on line"(ex: "on-line English to Portuguese dictionary").

In order to cope with all the possible variations and word orderings in the search strings, we divided them into single words and compiled a new list containing the number of occurrences of each individual. The results are given in table 5. Again, one of the most frequent single words in the search strings coming from Google is "dicionario" and "dicionário" (note the enormous number of people who do not use accents because they know that searchers like Google are not accent sensitive), which suggests that most of the people that come to Linguateca through major search engines are looking for general language resources.

Search string	# queries
linguateca	832
dicionario ingles portugues on line	812
literatura infantil	625
livrarias	602
portugues para estrangeiros	582
priberam	463
compara	457
avalon	451
editoras	431
power translator	431
livrarias portugal	424
dicionario portugues ingles on line	392
dicionario portugues aurelio	391
português para estrangeiros	384
dinalivro	381
dicionario portugues	360
curriculum vitae	349
dicionario portugues ingles	334
dicionario portugues on line	315
Enciclopedias	310

Table 4 - Top 20 most frequent search strings

Word in search string	# occurrences
de	36151
portugues	18102
dicionario	14228
dicionário	11725
ingles	10920
download	8757
português	8419
on	8270
line	7966
para	7941
em	6746
da	5612
inglês	5349
do	5063
e	5054
online	4953
portuguesa	4230
língua	3350
tradução	3034
Termos	2895

Table 5 - Top 20 most frequent single words in the search strings

4.2. A closer look at search strings

After this more quantitative analysis, we proceeded with the analysis focusing on the semantics of the search strings from the Busca logs (i.e. not the Google search strings). From simple inspection, and excluding queries that obviously have no connection with the content of the website, we were able to identify six types of search queries:

1. **Queries about informatics in general.** E.g.: “CAD”, “Pascal”, “Java”, “Autocad 2000”, etc. The user seems not to be very well-informed about the content of the website and is looking for technical resources that are probably not covered by Linguateca’s website (it is possible that such a user has reached our site after a search in an external engine).
2. **Queries about topics concerning mainly the Portuguese language (literature, grammar, use).** E.g.: “figuras de estilo”, “verbos”, “Tipos de Sujeito Indeterminado e Oração sem Sujeito”, “verbo inacusativo”, “expressões idiomáticas”, “falsos amigos”, “provérbios”, “exemplos de orações condicionais”, etc. In this case, the user is searching for what seems to be the answer to general doubts about Portuguese. We may assume the user would be happy to find definitions, explanations or examples about the topic and so documents with such information would be preferred. This category also includes queries that reveal that the user is looking for very specific items such as idiomatic expressions (“dar com os burros n’água”), or is searching for material about certain topics: “propostas de testes de português”, “textos com uso de próclise”.
3. **Queries about specific fields or knowledge domains.** E.g.: “extracção de informação”, “terminologia”, “semântica lexical”, “Portuguese language history”, “tradução automática”, etc. The search string indicates a topic but no precise question in mind, which suggests that the user may be looking for general information or pointers to documentation about these topics.

4. **Queries about general tools or resources.** E.g.: “corpora”, “dicionário”, “conjugador de verbos”, etc. These are examples of queries that strongly suggest that the user is trying to find resources, either for download or for on-line use.
5. **Queries about specific tools or resources.** E.g.: “Cetempúblico”, “Cetenfolha” (two corpora distributed by Linguateca), “COMPARA”, “Corpógrafo”, which indicates that the user is trying to find a particular resource that he knows about.
6. **Queries that seem to be intended for our on-line concordance tools rather than for the search engine.** E.g.: “sem nada”, “abonad.+”, “ansioso para”, “porém (ocorrências)”. These queries are interesting because they show that the user knows what he wants to do, i.e., use our on-line concordancer, but he cannot find it.

All six cases suggest that the user has a different goal in mind, and also different levels of knowledge about the content of the site and the areas that it covers. They also show that users are relatively familiar with terminological units, especially noun phrases, and use them in search expressions naturally (even if the TUs are inappropriate in respect to the content of our website). Sometimes, however, as we have seen before, users type incomplete, ill-defined or misspelled terminological units.

It should also be mentioned that user requests may be those concerning subjects that are not easy to find otherwise (one should observe that “access to resources” is above the option “search site” in Linguateca’s main menu displayed in Figure 1).

4.3. Preliminary consideration for the improvement Busca

From these observations and from the point of view of a terminologist it became clear that something could be done to improve Busca without the need for an extremely sophisticated technological apparatus.

First of all each document in the site should be indexed using only the TUs it contains. This can be performed quite easily if the complete list of TUs is known (although choosing the best indexing TUs requires much more information and linguistic analysis, see e.g Peñas et al (2001). This would also be very useful in detecting misspelled search strings. Secondly, knowing all possible variants and synonyms of a given TU would help in dealing with several different wordings that usually mean the same. For example, there are small differences in the spellings between the European and Brazilian variants of many of the relevant terminological units (TUs) (ex: “adjectivo” and “adjetivo”), which should be dealt with robustly by Busca. Finally, for some more problematic search strings (ambiguous, incomplete) there should a set of TUs that would be suggested to allow the user to reformulate the query. This would significantly increase the usability of the system.

While the third option might be something the site maintainers should think hard about (in fact, it depends on what kind of suggestions we are thinking of), the first two are typically the result and aim of terminological work: which TUs best represent the concepts in an area, and how should synonyms and related terms be identified, and this is why we employed Corpógrafo. As a side effect of this study, the Corpógrafo underwent many changes motivated by the need to obtain specific data.

5. Empirical work

5.1 The corpus

We started by selecting a subcorpus made from 178 files in different formats (78 html, 57 pdfs, 16 rtf, 16 Ms-Word and 11 plain-text) and composed of documents in Portuguese stored in our main web server or pointed from our catalogue, including articles and other technical publications included in our publication catalogue. (This was a hopefully balanced subset of the Busca search domain).

Since most of our users are native speakers of Portuguese (for the current statistics, see <http://acdc.linguateca.pt/estatisticas.html>), at this stage we only considered documents in Portuguese (all varieties) and left the English section of the site to a later stage.

The corpus total number of tokens in this corpus is 969462, therefore, approximately 1M. Corpógrafo allowed us to extract and manually validate 1209 terminological units, ranging from single word TUs to 5 or more word TUs (see figure 4 below).

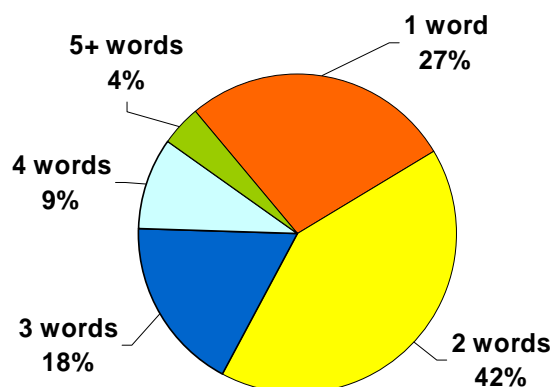


Figure 4 - Distribution of TU according to the size (in number of words)

Using Corpógrafo we were able to obtain some statistics about each TU, namely values regarding their frequency (in number of occurrences per million words - opm), and their distribution in the collection, i.e. the number of documents in which they occur. Tables 7, 8 and 9 show the top 10 most frequent TUs (up to three words) while table 10, 11, and 12 show the most widely distributed TUs (up to three words) in the collection.

TU	# - opm	# docs
português	2288 - 2.360	140 - 78.7%
texto	2189 - 2.258	126 - 70.8%
sistema	2095 - 2.161	116 - 65.2%
palavra	4027 - 4.154	114 - 64.0%
corpus	1842 - 1.900	102 - 57.3%
linguagem	866 - 893	96 - 53.9%
termo	1461 - 1.507	93 - 52.2%
recurso	613 - 632	92 - 51.7%
processo	917 - 946	90 - 50.6%

ferramenta 474 - 489 82 - 46.1%

Table 6 Top 20 most frequent single-word TUs

TU	# - opm	# docs
língua portuguesa	339 - 350	92 - 51.7%
modelo conceptual	323 - 333	4 - 2.2%
linguagem natural	256 - 264	56 - 31.5%
modelo coclear	132 - 136	1 - 0.6%
língua natural	127 - 131	22 - 12.4%
classe gramatical	118 - 121	26 - 14.6%
sintagma nominal	115 - 119	27 - 15.2%
rede neural	108 - 111	4 - 2.2%
tradução automática	103 - 106	33 - 18.5%
categoria gramatical	97 - 100	30 - 16.9%

Table 7 - Top 20 most frequent two-word TUs

TU	# - opm	# docs
base de dados	436 - 450	48 - 27.0%
banco de filtros	146 - 151	1 - 0.6%
aquisição de vocabulário	130 - 134	2 - 1.1%
português do brasil	94 - 97	26 - 14.6%
impedância da partição	85 - 88	1 - 0.6%
banco de dados	80 - 83	16 - 9.0%
sistema de interrogações	76 - 78	3 - 1.7%
unidade de alinhamento	64 - 66	4 - 2.2%
taxa de disparos	63 - 65	1 - 0.6%
sinal de fala	58 - 60	3 - 1.7%

Table 8 - Top 10 most frequent three-word TUs

TU	# - opm	# docs
palavra	4027 - 4154	114 - 64.0%
modelo	2470 - 2549	49 - 27.5%
português	2288 - 2360	140 - 78.7%
texto	2189 - 2258	126 - 70.8%
sistema	2095 - 2161	116 - 65.2%
corpus	1842 - 1900	102 - 57.3%
termo	1461 - 1507	93 - 52.2%
verbo	1446 - 1492	73 - 41.0%
tradução	1114 - 1149	80 - 44.9%
dicionário	1023 - 1055	73 - 41.0%

Table 9 - Top 10 most distributed single-word TUs

TU	# - opm	# docs
língua portuguesa	339 - 350	92 - 51.7%

linguagem natural	256 – 264	56 – 31.5%
tradução automática	103 – 106	33 – 18.5%
língua inglesa	82 – 85	31 – 17.4%
categoria gramatical	97 – 100	30 – 16.9%
sintagma nominal	115 – 119	27 – 15.2%
classe gramatical	118 – 122	26 – 14.6%
português europeu	33 34	24 – 13.5%
corpus paralelo	52 – 54	23 – 12.9%
língua natural	127 – 131	22 – 12.4%

Table 10 – Top 10 most distributed two-word TUs

TU	# - opm	# docs
base de dados	436 – 450	48 - 27.0%
português do brasil	94 – 97	26 - 14.6%
engenharia da linguagem	17 – 18	19 - 10.7%
sinal de pontuação	39 – 40	18 - 10.1%
banco de dados	80 – 83	16 - 9.0%
tipo de texto	30 – 31	14 - 7.9%
processamento do português	14 – 15	13 - 7.3%
catálogo de recursos	13 – 13	12 - 6.7%
lista de palavras	34 – 35	12 - 6.7%
português de portugal	40 – 41	12 - 6.7%

Table 11 - Top 10 most distributed three-word TUs

As expected, there are different winners according to frequency or distribution over the corpus. Figure 5 provides a graphical display of this data for all TUs. Each point in the chart represents a TU plotted in logarithmic scale. The horizontal axis measures the base 10 logarithm of the distribution, while the vertical axis measures the logarithm of the frequency (in opm).

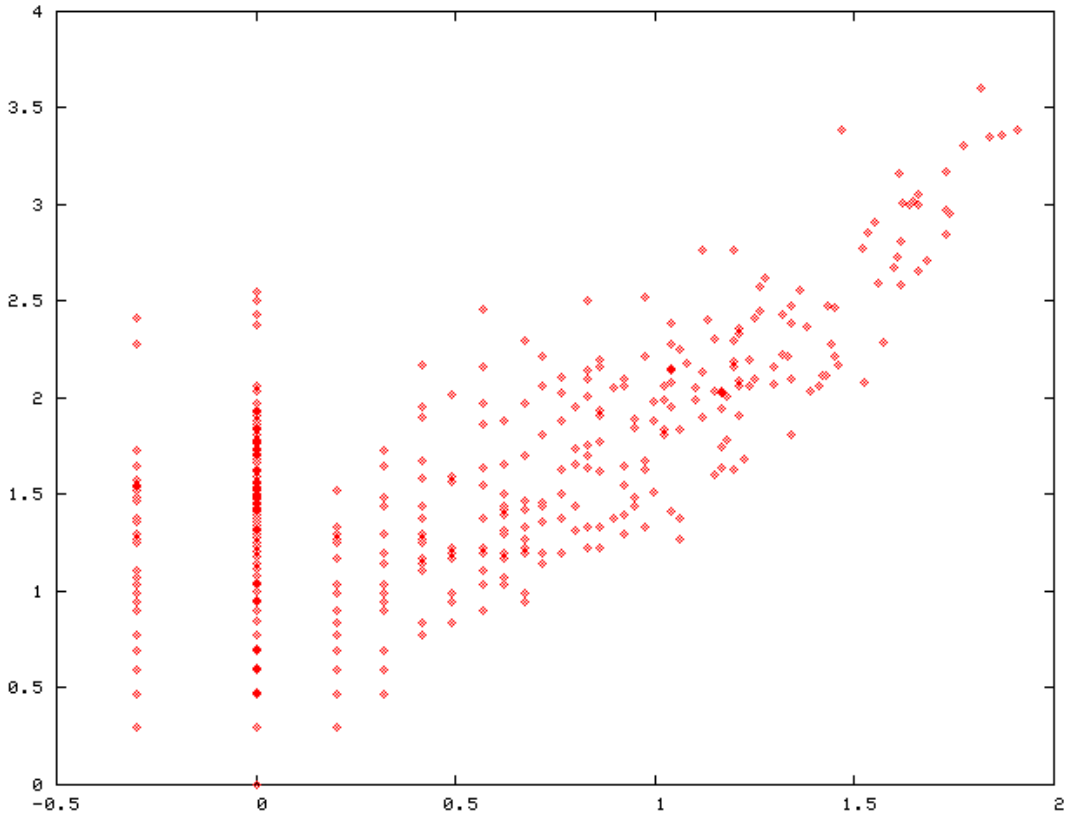


Figure 5: Frequency and Distribution of the 1209 TUs extracted. The axis are set to logarithmic scale.

This chart may be divided into 3 main regions:

1. Region 1 (upper-left sector), corresponding to frequent but not widely distributed TUs. E.g.: “modelo coclear”, “taxa de disparos”. These are usually compound words.
2. Region 2 (upper-right sector), which includes frequent and widely distributed TUs. E. g.: “análise”, “corpus”, “modelo”, “linguística”, etc. These are usually very generic TUs, and most of the times single words (they nevertheless have multiple possible modifiers).
3. Region 3 (lower-left sector) where less frequent and less distributed TUs may be found. E.g.: “verbo intransitivo”, “relação semântica”, “vibração macromecânica”.

The lower-right corner would correspond to less frequent TUs that nevertheless occur in several documents. This sector is empty because it is very difficult to find valid TUs that are simultaneously widely distributed and not too frequent. Region 1 is particularly interesting because it contains TUs that occur very frequently but in a small number of documents, which suggests that they would be good indexation terms: their ability to discriminate documents is very high. TUs from Region 3 could also be good indexation terms, especially because they occur in just a small number of documents. Region 2 is populated by TUs usually omnipresent in NLP and computational linguistics documents.

After the terminological extraction we also compiled several other items that might help the user during the search, namely:

1. Synonyms Portuguese (53 pair) - E.g.: “adjetivo: adjetivo”, “bibliografia: documento: publicação”;
2. Translation equivalents between Portuguese-English (107 pairs)- E.g.: “dicionário: dictionary”;
3. Synonyms English (23 pair)- E.g.: “parsing system: parser”;
4. Acronyms in Portuguese and English (81)- E.g.: “RI: Recuperação de Informação”.

Unfortunately the Corpógrafo did not help much at this specific stage: these lists were compiled manually, in much the same way as any translator would do in the course of his/her daily routine. This task confirmed the need for including in Corpógrafo appropriate semi-automatic methods for speeding up bilingual terminology alignment and for acronym matching.

5.2 The terminological units

We then looked in more detail into the TUs found, attempting to classify them according to several linguistic criteria:

Morphology and Syntax. There are multiple possible morphological combinations in the extracted TUs (see Table 12). Single word TUs are mostly either the names of available tools and resources, such as “COMPARA” and “Smorph”, or common nouns such as “alinhador”, “anotador” and “corpus”, which are general words in the specific knowledge domain. Multiword TUs are usually a combination of different morphological categories. The most common combinations in our corpus are *Noun + Preposition + Noun* and *Noun + Adjective*.

POS	occur.	%	Examples
CN + ADJ	504	41,6	vagueza grammatical, sumarização automática
CN	226	18,7	dicionário, gramática
CN + PRP + CN	178	14,7	sistema de tradução, sinal de fala
PN	52	4,3	COMPARA, Corpógrafo
CN + PRP + CN + ADJ	37	3,1	reconhecimento de dígitos isolados, resolução da ambigüidade lexical
CN + PN	35	2,9	dicionário Aurélio, sistema Edite
CN + PRP + CN + PRP + CN	28	2,3	arquitectura do sistema de interrogações, processo de aquisição de vocabulário
CN + ADJ + PRP + CN	20	1,7	Legendagem automática de notícias, reconhecimento óptico de caracteres
CN + PRP + PN	19	1,6	modelo de Kanis-Deboer, teorema de Bayes, rede de Elman
Acronym/abbreviation	14	1,2	bd, cce, IA, lil
CN + ADJ + PRP + CN + ADJ	9	0,7	processamento automático da linguagem natural, criação semi-automática de recursos lexicais
CN + ADJ + PRP + PN	3	0,2	modelo auditivo de Seneff, modelo coclear de Goldstein
Other POS structures	84	7	

Table 12 –The distribution of existing POS structures (ADJ – adjective; CN – common name; PN – Proper Name; PRP - Preposition)

An interesting case is the *Noun + Adjective* form. In TUs with such structure the head noun is often a general TU (e.g.: “análise”, “corpus”) specialized by the addition of specific adjectives (such as “morfológica”, “anotado”). This confirms Caraballo’s (1999) remark that “one possible indicator of specificity is how often the term is modified. It seems reasonable to suppose that very specific nouns are rarely modified, while very general nouns would usually be modified”. For example, for the two nouns “corpus” and “homonímia”, respectively occurring 463 and 21 times in our corpus, there was only one modifier (categorical) for the latter, as opposed to the situation for “corpus” displayed in table 13 below. We will use the specificity concept to suggest improvements for Busca.

	Modifier	# Occurrences
Corpus	Comparado	1
	Etiquetado	1
	Multilíngue	1
	Jornalístico	2
	Anotado	6
	Eletrônico	17
	Paralelo	21

Table 13: Modifiers of the TU “corpus”.

Semantic Classification. After this study of morpho-syntactic features, we also analysed the identified TUs according to some possible semantic categories, developed for the specific purpose of dealing with the pragmatics of the search in the Linguateca site (i.e., we do not claim that this is general enough for all purposes).

1. **Language resources.** E.g.: “corpora”, “CETEMPúblico”, “dicionário”, “Wordnet”, “COMPARA” etc.
2. **Tools and systems.** E.g.: “anotador”, “analizador morfológico”, “Corpógrafo”, etc.
3. **Actions and processes.** E.g.: “aquisição de vocabulário”, “extração de terminologia”, “anotação de corpora”.
4. **Specific theories and models.** E.g.: “modelo auditivo de Seneff”, “algoritmo de Earley”, etc.
5. **Linguistic concepts and phenomena.** E.g.: “polissemia”, “ambiguidade lexical”, “verbo inacusativo”, “advérbio de tempo”, “adjectivo”, etc.
6. **Disciplines or knowledge fields.** E.g.: “lexicografia”, “engenharia da linguagem”, “inteligência artificial”, “semântica lexical”, etc.

The proposed classification is based on our knowledge about the content of the site and about what users may be searching for. Notice that what it does is very similar to the classes we proposed after analysing the Busca logs and it tries to be a compromise between our view of our own site and possible user needs. The specific identification of named entities (in our case, mainly NLP resources) to improve (Web) IR is obviously not a new idea (see Pasca (2004) for recent work).

6. Suggestions for improvement of the search

After doing the work reported above, we are able to offer the following suggestions for search improvement in the Linguateca site, i.e. both improvement of Busca's search capabilities and user satisfaction.

Since most of the Busca queries are single words (see figure 3), special care has to be taken for dealing with these cases. For example, if a user inputs a general word alone in the search box, i.e. a TU that may be modified in several ways, Busca could present the corresponding list of possible modifiers and furnish a set of specialization options to the user, instead of providing a large set of results. On the other hand, if the word is matched against a known name of a resource or tool (e.g. "COMPARA") then a very good option would be to presenting the user with a direct link to that resource/tool homepage, in addition to a list of results. It is important to say that both these solutions can easily be implemented in a fully-automated way.

Another very simple but extremely useful change would consist in improving the current query processing mechanism to be able to deal more robustly with the different varieties of Portuguese. Although Portuguese varieties present minor variation in the morphology (in this case we are dealing mostly with European and Brazilian varieties), this variation apparently tends to occur frequently in many TUs relevant to this domain. Similar query processing improvements are needed in order to deal with the loose usage of accents ("dicionário" vs "dicionario"), which users are quite used to because of the behaviour of general search engines.

Query processing could be further improved, by using synonym lists, acronym lists and translation equivalents, which can be used to expand the user queries. This, however, could lead to problematic issues, because these transformations are not always unambiguous. Nevertheless, some improvements could be tried, especially by using synonyms lists.

Other substantial improvements could be achieved by making more sophisticated changes in Busca, which nevertheless could be reasonably implemented. For example, one could cluster the results according to the modifiers, as is done by some search engines (see <http://clusty.com/> or <http://beta.exalead.com/search>).

Additionally, after performing a semantic classification of the kinds of TUs that are covered by the site (according to the classification described above), we can build different and more discriminated behaviour of the search system, in several ways. An interesting solution could be to try to identify the user's kind of need from the semantic classification of the TU s/he inputs and allow for "intelligent" suggestions of other similar resources. For examples, if a user types "corpora" which is identified as a resource, the system could present pointers to several resources available (not necessarily just corpora), or point directly to our catalogue of resources. The same could be applied for tools.

Further improvement could be achieved by combining the knowledge about the semantic classification of the keywords with some pragmatic rules of thumb, such as: if one is interested in a particular technology/tool/resource, one may be interested in systems that apply or implement such a technology or function, one may be able to provide to a user who inputs "morphology" in the search box a choice between "scientific discipline", "applications that deal with morphology" (such as morphological analysers, stemmers, morphological generators, POS taggers), "specific systems that perform any of these tasks" (such as Palavroso, PALMORF, etc.) or even "evaluation" (given

that there is substantial material on how to evaluate morphological analysers at the Linguateca site). Such rules could be coded manually and applied after having the correct semantic classification of each TU (probably more refined version of the proposed classification scheme), or have an automatic text categorization system similar to the one described by Aires et al. (this volume).

Although we have not yet explored (and tested) the Corpógrafo tools for finding semantic relations, there are other interesting options to explore if some semantic relations between TUs are known. For example, one may be able to add as indexing terms those TUs not explicitly present in the documents themselves, but already known to be intrinsically related to TUs in those documents, like relating “algoritmo de Earley” to “análise sintáctica”, or “CLEF” to IR. One kind of semantic relation that is especially relevant is the hypernymy / hyponymy relation, since a user does not know at which level of generality the texts were indexed. If the search terms are too general, the search system should help the user to specialize; if the search terms are too specific, but providing the system has a richer terminology available, it can suggest a generalization. Knowing these relations is therefore important, but we are still in the process of compiling that information using the Corpógrafo.

We hope to present a terminology referring to the areas covered by the Linguateca site on which Busca can rely to produce this adaptive behaviour in the near future.

7. Concluding remarks

We have not yet been able to index the site with the result of our work, and therefore cannot but hope that subsequent user testing will show that this actually helped users to find relevant information in our site.

Still, we were able to produce a list of indexing TUs that may be useful for browsing, and which should provides a bird’s eye of our site for newcomers. We are still considering ways to visualize this as a site map, though.

On the practical side, we were able to test Corpógrafo’s capacity for information retrieval on a real world website. The exercise provided for a set of improved functionalities, as well as proving its robustness. We hope that this paper can be an inspiration for using Corpógrafo to do similar things to those attempted here, namely knowledge organization of a particular site using terminological units and semantic relations among them.

During the work, it was observed that bridging two areas of expertise such as IR and terminology, with different cultures and values, brings severe communication problems that can only be avoided or solved by continuous effort. All the authors had to learn how to communicate with the other field, and we found out how difficult it is in practice to agree on terminological and semantic grounds, long after the high level goals had been consensually agreed upon. Perhaps the most significant problem, however, is that IR and terminology, as areas of expertise, are each going through a phase of intense development.

Computer science offers new ways of realizing the age-old dream of classifying and organizing knowledge. While there still are several different (and conflicting) paradigms within the domain to deal with information and classification, such as IR, knowledge engineering and the new semantic Web effort, they are still finding out how to cope with the challenges posed by natural human language, not to mention those posed by the human beings who use it.

'Terminology', insofar as it was even considered an independent discipline, was traditionally considered a sub-area of lexicography, in which special domain experts and linguists supposedly co-existed to produce systematically organized language in which the experts could or should discuss their domains. Terminology work also included the semantic organization of the terms / concepts and various types of thesaurus work, or knowledge organization. The interest of the linguists was chiefly for activities like translation, or to serve the needs of language politics, as in the EU, Canada and Catalonia, and less often as a subject for research. Only recently has it developed an academic culture of its own.

Therefore, while NLP inside IR has not yet defined itself as mainstream in computer science, 'Terminology' is, at present, having to redefine and reorganize its activities because of computer science. Both areas, by definition, have to cope with all the problems of being interdisciplinary, in itself a serious academic problem, and both areas need each other if real progress is to be made. The traditional chasm between the mentality of the sciences – for IR – and that of the humanities – for terminology - desperately needs bridging. IR needs linguistic expertise, terminology needs computer power. Together they may even fuse into a new discipline – but we are not going to hazard a guess as to what this discipline will call itself.

Acknowledgements

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

References

Aires, Rachel, Diana Santos and Sandra Aluísio (2005) "Yes, user!": compiling a corpus according to what the user wants, this volume.

Caraballo, Sharon A. and Eugene Charniak. (1999) Determining the specificity of nouns from text, in *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 63-70.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther Koenig (1999) The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. Institute for Natural Language Processing, University of Stuttgart, March 8, 1999 (CQP V2.2), <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/> (accessed June 15th, 2005)

Jansen, Bernard J., Amanda Spink and 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.

Maia, Belinda, Luís Sarmiento, Diana Santos, Luis Miguel Cabral & Ana Sofia Pinto (2005) CORPÓGRAFO – an online suite of tools for the construction and analysis of corpora, semi-automatic extraction of terminology and the construction of conceptual databases. Poster to be presented at Corpus Linguistics 2005.

Maia, Belinda (2005) Terminology and Translation – bringing research and professional training together through technology in *Proceedings of the META Symposium - For a Proactive Translatology* (Université de Montréal, Québec, Canadá, 7-9 April 2005), to appear

Oksefjell, Signe and Diana Santos (1998) Breve panorâmica dos recursos de português mencionados na Web, in Vera Lúcia Strube de Lima (ed.), *III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98)* (Porto Alegre, RS, 3-4 de Novembro de 1998), 38-47.

Pasca, Marius (2004) Acquisition of Categorized named Entities for Web Search, in David Grossman, Luis Gravano, ChengXiang Zhai, Otthein Herzog, David A. Evans (eds.), *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management*, 137-145.

Peñas, Anselmo, Felisa Verdejo and Julio Gonzalo (2001) Corpus-based terminology extraction applied to information access, in *Proceedings of Corpus Linguistics*, Lancaster University, UK, 2001. <http://nlp.uned.es/~anselmo/articulos/cl2001.pdf> (accessed June 15th, 2005)

Santos, Diana (2000) O projecto Processamento Computacional do Português: Balanço e perspectivas, in Maria das Graças Volpe Nunes (ed.), *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)* (Atibaia, SP, 19-22 de Novembro de 2000) (São Paulo: ICMC/USP), 105-113.

Sarmiento, Luís, Belinda Maia and Diana Santos (2004) The Corpógrafo - a Web-based environment for corpora research, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation* (Lisboa, Portugal, 25 May 2004), 449-452.

Sarmiento, Luís (2005) A Simple and Robust Algorithm for Extracting Terminology, in *Proceedings of the META Symposium - For a Proactive Translatology* (Université de Montréal, Québec, Canadá, 7-9 April 2005), to appear.