INTERNATIONAL WORKSHOP

NATURAL LANGUAGE PROCESSING METHODS AND CORPORA IN TRANSLATION, LEXICOGRAPHY AND LANGUAGE LEARNING

held in conjunction with the International Conference
RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by

Iustina Ilisei, Viktor Pekar and Silvia Bernardini

Borovets, Bulgaria

17 September 2009

International Workshop

NATURAL LANGUAGE PROCESSING METHODS AND CORPORA IN TRANSLATION, LEXICOGRAPHY AND LANGUAGE LEARNING

PROCEEDINGS

Borovets, Bulgaria
17 September 2009

ISBN 978-954-452-010-6

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Programme Committee

Marco Baroni (University of Trento)
Jill Burstein (Educational Testing Service)
Michael Carl (Copenhagen Business School)
Gloria Corpas Pastor (University of Malaga)
Le An Ha (University of Wolverhampton)
Patrick Hanks (Masaryk University)
Federico Gaspari (University of Bologna)
Adam Kilgarriff (Lexical Computing)
Marie-Claude L'Homme (Université de Montréal)
Ruslan Mitkov (University of Wolverhampton)
Roberto Navigli (University of Rome 'La Sapienza')
Miriam Seghiri (University of Malaga)
Pete Whitelock (Oxford University Press)
Richard Xiao (Edge Hill University)
Federico Zanettin (University of Perugia)

Organising Committee

Iustina Ilisei (University of Wolverhampton, United Kingdom)
Viktor Pekar (Oxford University Press, United Kingdom)
Silvia Bernardini (University of Bologna, Italy)

TABLE OF CONTENTS

Search techniques in corpora for the training of translators

Verónica Pastor
TecnoLeTTra
Universitat Jaume I
Av. de Vicent Sos Baynat, s/n. 12071
Castellón de la Plana, Spain
vpastor@trad.uji.es

Amparo Alcina
TecnoLeTTra
Universitat Jaume I
Av. de Vicent Sos Baynat, s/n. 12071
Castellón de la Plana, Spain
alcina@trad.uji.es

Abstract

In recent years, translators have increasingly turned to corpora as a resource in their terminology searches. Consequently university translation courses should include training for future translators in the effective use of corpora, thus enabling them to find the terminology they need for their translations more quickly and efficiently.

This paper provides a classification of search techniques in electronic corpora which may serve as a useful guide to the efficient use of electronic corpora both in the training of future translators, and for professional translators.

Keywords

Electronic corpora for translators, search techniques, corpus queries, translation resources, translation training.

1. Introduction

Terminology is a key factor in translators' work. The development of specialized fields has grown hand in hand with advancements in science and technology. These market demands explain why translators are calling for resources to satisfy their terminological needs quickly and effectively [1].

Dictionary creation cannot keep pace with developments in specialized fields. Many studies show dictionaries to be deficient in the lack of information they include, speed of content update, and the limited ways of accessing contents. For this reason, translators are increasingly turning to other resources, such as the Internet and corpora, to search for the terminology they need.

In this paper we analyze the search techniques offered by a range of electronic corpora. Our search technique classification is aimed to provide translation teachers with a reference to help them teach students how to use corpora efficiently. This classification may also be of interest to professional translators who want to further their knowledge of electronic corpora techniques in order to improve their query results.

2. The need for corpora in translation

Market demands require translators to work against tight deadlines and with rapidly evolving vocabulary. According to Varantola [22], fifty per cent of the time spent on a translation is taken up with consulting reference resources. Many studies have revealed that dictionaries do not satisfy all translators' terminological queries [5, 9, 16]. Gallardo and Irazazábal [10] suggest that the terminology translators need, apart from equivalents in different languages, should also include contexts and information about the concept that allow translators to decide how and where to use a term.

In this vein, Zanettin [23] states that the use of corpora in translation training was commonplace even before the development of electronic corpora. Snell-Hornby [21] and Shäffner [18], for instance, argue that by studying similar texts in the source and target languages translators may identify prototypical features that are useful for the target text production.

Since the development of electronic corpora, the need for these tools has become more evident, especially as a terminology resource for translators. Several authors state that translators need new terminological resources, such as corpora [3, 4, 11, 15], which complement dictionary and database use [8, 12, 19] and satisfy specific terminological problems quickly and reliably.

Some studies have demonstrated that translation quality improves when translators use corpora in their terminology searches. Zanettin [23] conducted an experiment with translation students from the School for Translators and Interpreters at the University of Bologna. He shows that comparable corpora[1] help translation students to compare the use of similar discourse units in two languages and facilitate the selection of equivalents adapted to the translation context. Bowker [3] carried out a study with translation students from the School of Applied Language and Intercultural Studies at Dublin City University. She found that corpus-aided translations are of higher quality than translations carried out only with the aid of dictionaries. In a subsequent study, Bowker [4] suggests various ways a target language corpus can be used as a terminological resource for translators.

Despite the usefulness of corpora, the need to use a range of resources to access terminology is a daily problem facing translators. According to Alcina [1], if translators

[1] Zanettin [23] defines a comparable corpus as a collection of independent original texts in two or more languages with similar content, domain and communicative function.

have to undertake terminological tasks, whether searching in a corpus or on the Internet, time is wasted and their translation efficacy is poorer. As Varantola [22] states, the success of a query depends on the intelligent use of search tools.

Translation students should receive quality training at university level in the use of new electronic resources in order to respond to the demands of companies and institutions [2]. Any training in electronic resources should also include electronic corpora search techniques[2]. If this were the case, translators would spend less time and effort acquiring the competences to query corpora efficiently once they have embarked on their professional career. If translators know how to use search techniques in electronic corpora, they will be able to satisfy their terminological needs more quickly and efficiently and the quality of their translations will improve.

3. Corpora examined in the analysis

This study analyzed the search functions of various stable online corpora interfaces. Because we wanted to analyze corpora that are easily accessible to translators, we selected those that are available online. All the corpora analyzed incorporate interfaces that allow different types of queries.

It is worth noting that many of these corpora are not specifically designed for translators. In addition, each corpus explains its own query options, but few studies provide a comprehensive and systematized classification of all the search techniques that can be used in a corpus.

Our classification will provide an overview of all the search techniques that have been incorporated in electronic corpora to date. We will use this classification in future research to discover which of these search techniques are useful for translators, in order to create electronic corpora adapted to translators needs, as well as to teach translators the range of search techniques used in electronic corpora.

In this section we briefly describe the corpora analyzed, focusing on the particular features of each corpus. Specific examples of queries in the corpora are included in our search technique classification.

The **Corpus de referencia del español** (CREA) and the **Corpus diacrónico del español** (CORDE) are two monolingual online corpora developed by the Real Academia Española. CREA contains modern Spanish texts from 1975 to 2004. CORDE includes Spanish texts written up to 1975. Both corpora allow the use of distance criteria between words. Corpus filters such as field, author and work, date, register, and geographic area can be applied. Statistical data on search results, concordances and clusters are also available.

At Brigham Young University (BYU), Professor Davies created an online interface for a set of monolingual corpora: the **Corpus del español** (Spanish from 1200s-1900s), **Corpus of Contemporary American English** (US English from 1990-2008), **BYU-British National Corpus** (British English from 1980-1993), **TIME Corpus** (US English from 1923-present), **BYU-OED Oxford English Dictionary** (Old English-1990) and **Corpus do Português** (Portuguese from 1300s-1900s).

This interface allows the user to search one or more word forms, lemmas or parts of speech. Part of speech restrictions can be applied. Searches can also be limited by genre or over time. It compares the frequency of words, phrases or grammatical constructions by genre or over time. The user can search for collocates of a word or compare collocates of two words. Another particular feature is the semantically-oriented search, which enables the user to search for synonyms[3] of a word. Finally, customized lists of words or phrases may be created for use in a query.

The **British National Corpus** (BNC) is a monolingual corpus of modern spoken and written British English. The online interface allows the user to search for a word or phrase. More complex queries can be carried out using the SARA/XAIRA search tool (www.oucs.ox.ac.uk/rts/xaira) or directly from the online search box using the BNC Corpus Query Language, or the online CQP edition of the BNC[4].

The **Hellenic National Corpus** (HNC) is an online monolingual corpus containing 47 million words developed by the Institute of Language and Speech Processing. It covers written Modern Greek since 1990. One feature of this corpus is that it allows the user to define the distance between three words, lemmas or parts of speech within the same query[5].

BwanaNet is an online corpus search tool developed to query a collection of specialized corpora from the Institut Universitari de Lingüística Aplicada (IULA) at the Universitat Pompeu Fabra. This collection of corpora includes original and parallel texts in Catalan, Spanish and English, from the fields of Computing, Environment, Law, Medicine, Genome, Economy, and other specialized areas.

This interface generates lists of word forms, lemmas or parts of speech. Users can search for concordances of one or more words, lemmas or parts of speech. Part of speech restrictions have two features: 1) the option to delimit, in a grammatical construction, the number of subsequent occurrences of the same category (between 0 and 9), and 2)

[2] Alcina [2] presents a didactic proposal divided into four levels of specialization in Computerized Terminology. In this proposal she includes training to query online corpora or other formats, as well as the use of corpora search tools.

[3] For more information on semantically-oriented searches, see Davies [7]. We include examples of this type of search in our search technique classification.

[4] Available at http://bncweb.lancs.ac.uk after registration.

[5] Most corpora allow distance to be defined between two elements only.

the search for a word form or lemma that excludes a particular part of speech. The user can also limit the search to a section of the corpus (titles, lists, tables, text). Other queries can be carried out using the Corpus Query Processor language[6].

COMPARA is an online bidirectional aligned corpus of English and Portuguese. To query this corpus, the user needs to be familiar with the CQP language. The interface allows the user to limit the search to linguistic variants of Portuguese or English, date, author, etc. In addition, concordance formats can be modified, for instance by displaying alignment properties or part-of-speech tags.

4. Classification of search techniques in corpora

Search techniques are options that a user can apply to a resource to obtain a result. We distinguish three elements in a search technique: a query probe, a query resource and a query outcome. The *query probe* is the word or phrase introduced by the user in the interface of a resource. The *query resource* is the resource or part of the resource in which the word or phrase is searched. The *query outcome* is the result obtained in a query when a probe is searched in a resource.

In this paper we present a classification of search techniques in electronic corpora that focuses on the query probe, the query resource and the query outcome. An example of a corpus search technique could be to use an exact word as a probe, e.g., we look for the word *play* in an English monolingual corpus (resource) to obtain a list of concordances—the outcome—of the word *play*, which includes expressions such as *play the piano, play football* or *play the role of.*

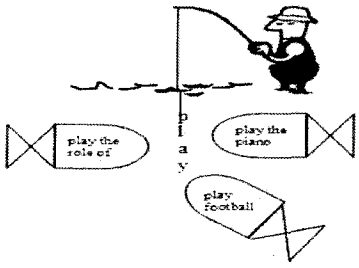


Figure 1. Representation of a search technique in an electronic corpus

Below, we explain in more detail the search techniques that can be used in an electronic corpus, and provide examples of how these search techniques are applied in the corpora analyzed.

[6] The Corpus Query Processor (CQP) manual is available at http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.

4.1 Query probe

The query probe is an expression the user tries to find by introducing it in a corpus interface. We categorize query probes as follows: lexical expressions, grammatical expressions, numbers, hybrid expressions, and non-continuous combinations of expressions. Filters may be applied to restrict probes.

4.1.1 Lexical expressions

Lexical expressions can take a single form or a lemma, or a sequence of forms or lemmas.

A **lemma** is the base form of a word, i.e., it is a word without inflectional morphemes. The lemma of a noun is its form with no gender or number morphemes. The lemma of a verb is the infinitive.

A lemma is a useful way of retrieving all the forms in a corpus that are tagged with that lemma. For example, if we introduce the lemma *do* in the BYU-British National Corpus, the corpus retrieves all the forms of this verb that appear in the corpus: *do, did, does, done, doing, etc.*

A **form** can be exact or partial. An **exact form** is a complete word. It can be useful for finding a particular form of a word in the corpus. For instance, we can search the plural word *houses* in any of the corpora analyzed.

A **partial form** is an incomplete word. The omitted part of the word is replaced by a wildcard. The most frequent wildcards are the asterisk (*), which replaces one or more characters, and the question mark (?), which replaces only one character. Partial forms can be useful if we want to search all the words that start, end or contain a specific sequence of characters. For example, if we introduce the partial form *hous** in the COMPARA corpus, the following complete forms are retrieved: *house, housewife, housekeeper, house-doctor, houses, housing, household*, etc.

Lexical expressions can also be **sequences of two or more forms or lemmas**, which may be exact or partial. An **exact sequence** is a phrase or combination of forms or lemmas that appear in the corpus in the same order as those searched for. Exact sequences can be introduced to see the context in which a particular expression is used. In the following example we introduce the exact sequence of the forms *raining cats and dogs* in the BNC. Two contexts are retrieved: 1) "It was raining cats and dogs and the teachers were running in and out helping us get our stuff in and just couldn't do enough for us." 2) "What must you be careful of when it's raining cats and dogs?"

A **partial sequence** is a combination of forms or lemmas in which one or more forms or lemmas are replaced by a wildcard. It can be used to search for an expression when we only know some of the words contained in it. In this example we introduce a partial sequence in the BYU Corpus del español: the verb *llover* as a lemma, followed by the preposition *a*, and then a wildcard, [llover] a *. Our search results include Spanish

expressions referring to 'raining heavily', such as *llovía a cántaros, llueve a torrentes, lloviendo a mares, lloviendo a raudales, lloviendo a chuzos, llovía a baldes,* etc.

Frequent sequences representing concept relationships, also called linguistic patterns[7], can also be introduced. Some of these patterns can be used to retrieve, for instance, defining contexts in a corpus (*is a, known as, is defined, is called,* etc.).

4.1.2 Grammatical expressions

Grammatical expressions are constructions made up of parts of speech. They may contain a single part of speech or a sequence of parts of speech. Grammatical expressions can be useful to find words or sequences of words by introducing their parts of speech in the corpus.

This search technique is a feature of BwanaNet, BNC, BYU corpora, COMPARA and HNC. For example, if we introduce the grammatical expression "adjective+noun" in the BwanaNet English Law corpus, the following expressions are retrieved: *commercial legislation, fiscal protection, Social Fund,* etc.

4.1.3 Numbers

Numbers can be exact or partial. If we introduce an **exact number** in the corpus, it is retrieved in the same form as it was introduced. A **partial number** is a number combined with a wildcard. In this case, the corpus retrieves all the numbers containing the sequence of numbers introduced with the wildcard. A number search can be useful to find words that appear in the same context as a significant number. For instance, if we introduce the number 640 in the BwanaNet Spanish Computing corpus, the word *píxel* appears, because the corpus retrieves the typical computing measurement *640×480 píxels*; the term *memoria RAM* is also found, because another specific measurement retrieved by the corpus is *640 Kb de memoria RAM.*

4.1.4 Hybrid expressions

Hybrid expressions combine lexical expressions, grammatical expressions and numbers. They can be useful to find expressions in which we know the form or the lemma of some of the words and the part of speech of other words. For example, we introduce in the BwanaNet English Law corpus a hybrid expression made up of a grammatical expression followed by a lexical expression: "adjective+*law*". The following expressions are retrieved *organic law, civil law, common law, Federal law, budgetary law*, etc.

4.1.5 Non-continuous combination of expressions

This search technique consists of introducing an element in a corpus and establishing the distance in which a second element must also appear. The first and the second element can be any of the query probes explained above: a lexical expression, a grammatical expression, a number or a hybrid expression.

In the following example we combine two lexical expressions in the BNC within a distance of 5 positions. The first expression is the form *Cytomegalovirus* and the second, an exact sequence of forms, the linguistic pattern *is a*. As a result we obtain some defining contexts of the word *Cytomegalovirus*.

Table 1. Results of the search for the form *Cytomegalovirus* within 5 positions of distance from the linguistic pattern *is a*

Cytomegalovirus (CMV) **is a** virus with many similarities to the herpes virus.
Cytomegalovirus is a less well-known infection which affects considerably greater numbers of babies than rubella.

In another example we use the BYU Corpus del español to combine the exact form *metros* within a distance of 5 positions from the number 100. Results include the expressions *100 metros libre* (100 meters free style) or *100 metros cuadrados* (100 square meters).

Table 2. Results of the search for the number 100 within 5 positions of distance from the form *metros*

terminó ayer su participación en Phoenix, Arizona, con el quinto lugar en los **100 metros** libre. [...]
) del lugar, con una área mínima de construcción de 200 **metros** cuadrados y **100 metros** cuadrados para parqueo. [...]

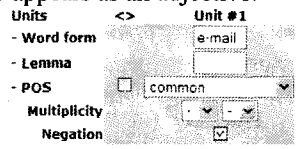
In this example, we use the BYU Corpus of Contemporary American English to combine the exact form *brake* within 3 positions of distance from the part of speech "verb". Results include expressions such as: *have a brake, have to brake, set the brake, released the hand brake,* etc.

The Hellenic National Corpus is the only corpus analyzed that allows the user to combine more than two elements noncontinuously without having to be familiar with the CQP interrogation language. Three forms, lemmas, numbers or parts of speech can be combined within 5 positions of distance.

4.1.6 Query probe filters

Filters add a search restriction to the query probe introduced, such as **part-of-speech filters**. For example, BwanaNet allows the user to search for forms or lemmas that may or not belong to a particular part of speech.

The following figure shows a query in the BwanaNet English Computing corpus. We introduce the form *e-mail* with the exclusion of the part of speech "noun" (option *negation* below the box *POS*). The result is a concordance in which *e-mail* appears as an adjective.

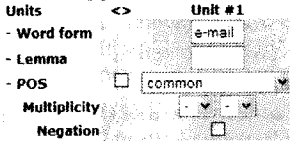


[...] it can be obtained through CD-ROM, **e-mail** server, [...].

[7] Many authors have studied linguistic patterns. See, for example, Sánchez [17], Faber et al. [8], López and Tercedor [13] or Meyer [14].

Figure 2. Search for a form or lemma with the exclusion of a particular part of speech

In contrast, if we introduce the form *e-mail* and limit the search to nouns, the result is a list of concordances in which the form *e-mail* appears as a noun.



By sending **e-mail** to clinton-info@campaign92.org, I was able to request press releases on foreign policy.
[...] suggestions on how to use everything from **e-mail** to remote databases, tutorials, lists of frequently asked questions, [...].

Figure 3. Search for a form or lemma as a particular part of speech

Part-of-speech filters are also available in the BNC, HNC, COMPARA corpus, and BYU corpora.

4.2 Query resource

The corpus resource is always the collection of texts that constitute that corpus. Nevertheless, depending on the query probe and outcome of a search technique, we can distinguish different types of electronic corpora[8]: monolingual corpora, aligned corpora and tagged corpora. Filters can also be used to restrict the corpus.

4.2.1 Monolingual corpora

An electronic corpus can be **monolingual,** i.e., all the texts in the corpus are written in the same language. In this type of corpora the query probe and outcome will always be in one specific language.

4.2.2 Aligned corpora

An electronic corpus can also be **aligned.** An aligned corpus is a parallel corpus composed of source texts and their translations. The introduction of query probes in monolingual and aligned corpora is usually the same but the query outcomes obtained vary.

In aligned corpora the query probe is normally introduced in one of the corpus languages, as in monolingual corpora. However, in aligned corpora, search results include the segments in one language, as well as equivalent segments in the second language. For example, when we introduce the form *run* in the English part of the COMPARA corpus, the concordances in English are given with the form *run* highlighted in bold. Equivalent segments in Portuguese appear next to the concordances of the form *run*. The equivalents of *run* (*montar, correr* and *comprar*) are not highlighted.

[8] Many authors have elaborated wider typologies of corpora in which all the types of corpora are described. See, for example, Corpas [6], Zanettin [24] and Sinclair [20]. In this paper, we have limited our classification of corpora according to what can we introduce in the corpus (the probe) and what can we obtain (the outcome).

Table 3. Results of the search of the form *run* in the aligned corpus COMPARA

I said to Nizar, «You could probably **run** a little rental business [...]»)	Em resposta, sugeri ao Nizar: «Talvez pudesse montar um negócio de aluguer [...]»
We had to **run** for a train once at Euston: [...]	Uma vez em Euston tivemos de correr para apanharmos um comboio: [...]
Neither of our families could afford to **run** a car in those far-off days.	Naquele tempo nem a minha família nem a dela tinham dinheiro para comprar carro.

Some aligned corpora also allow the user to introduce query probes simultaneously in both corpus languages. The COMPARA corpus offers an alignment restriction option that allows the user to introduce a query probe in one language with the condition that its equivalent segments contain another query probe. In the following example we introduce the form *run* in the English part of the COMPARA corpus, and the lemma *correr* in the Portuguese part. The corpus retrieves concordances of the form *run* in English whose equivalent segments in Portuguese include the lemma *correr*. In the concordances both the form *run* and all the forms of the lemma *correr* are highlighted in bold.

Table 4. Results of the simultaneous search for the form *run* in English and the lemma *correr* in Portuguese in the aligned corpus COMPARA

We had to **run** for a train [...]	[...] tivemos de **correr** para apanharmos um comboio [...]
If they had broken into a **run**, [...].	Se tivessem desatado a **correr**, [...].
But I feel we **run** a grave risk by doing so.	Mas eu acho que **corremos** um risco grave se o fizermos.

4.2.3 Tagged corpora

Corpora may either be tagged or not, and if they are, they may be tagged at different levels. In **POS-tagged corpora** all the words are tagged with their part of speech. Grammatical expressions can only be introduced in tagged corpora. In **lemmatized corpora** all the words in the corpus are tagged with their lemma. Lemmas can only be introduced in lemmatized corpora.

4.2.4 Query resource filters

The corpus search can be restricted to one section of the corpus using filters, such as thematic field, text type, geographic area, author, date, and text area.

The **thematic field filter** limits the corpus to sections of a selected thematic field. The BwanaNet, CQP edition of the BNC, CREA, and CORDE corpora offer this option. The **text type filter** limits the search to texts of a specific genre. This filter is available in the CREA, CORDE, CQP edition of the BNC, and BYU corpora. The **geographic area filter** limits the search to texts from a specific language area. For example, in the COMPARA corpus the search can be restricted to Portuguese from Angola, Portugal, Brazil and Mozambique, or to English from South Africa, United Kingdom or the United States. The CQP edition of the BNC also offers a geographic area filter.

The **author filter** limits the search to texts published by one or more authors. This filter is offered in the COMPARA, CREA and CORDE corpora. The **date filter** limits the search to texts published on a specific date or within a time period, and is a feature of the COMPARA, the CQP edition of the BNC, CREA, CORDE and BYU corpora. The **text area filter** limits the search to titles, lists, tables, etc. BwanaNet offers this filter. The CQP edition of the BNC allows the user to search in titles and keywords.

4.3 Query outcome

When an electronic corpus is queried, the user can select different types of query outcomes depending on the result he/she desires. These query outcomes may be a list of monolingual or aligned concordances, a list of words, a list of synonyms, a list of collocates or a list of clusters.

4.3.1 List of concordances

Concordances are the contexts in which the query probe appears. Most of the corpora provide concordances in an easy to read format called KWIC (key words in context), which means that the query probe is highlighted in the center of the context. Depending on the query resource used in a search technique, lists of concordances can be monolingual or bilingual.

4.3.1.1 List of monolingual concordances

Monolingual corpora can generate lists of monolingual concordances, i.e., lists of contexts in one language. Monolingual concordances are mainly used to observe a word in context. For example, if we look for the concordances of the lexical expression *Prime Minister* in the BNC we access contexts in English where this expression is used.

Another function of concordances is to find a word by searching for words that appear in a nearby context. For example, we can search in the BYU-OED Oxford English Dictionary corpus to find a word that refers to "a case where an archer holds arrows". In this case, we introduce the lemma *arrow* within 9 positions of distance to the part of speech "noun". The corpus retrieves concordances of nouns appearing near the forms of *arrow*; one of these nouns is the word *quiver*.

Table 5. Concordances of the lemma *arrow* within 9 positions of distance from the noun *quiver*

A gaily-painted **quiver**, full of **arrows**
He could draw an **arrow** from his **quiver** [...]

4.3.1.2 List of bilingual concordances

Aligned corpora can generate lists of bilingual concordances, which are lists of contexts in one language with equivalent contexts in another language. Bilingual concordances allow the user to decide on a more reliable translation equivalent because both the query probe in the source language and its equivalent in the target language are situated in a context that can be compared with the context of the translation, thus allowing the translator to verify equivalence.

In the following example we introduce the form *play* in the COMPARA aligned corpus and search for its concordances. Depending on the context, *play* is translated in Portuguese as *tocar* (when it refers to a music instrument), *jogar* (when it refers to a sport) or *fazer a* (when it refers to playing a role).

Table 6. List of concordances of the form *play* with its equivalents in Portuguese (highlight in Portuguese concordances added)

«[...] and not being able to **play** the piano.»	«[...] e à incapacidade de **tocar** piano.»
Joe wanted to switch partners and **play** the best of three sets, [...]	Joe queria trocar de parceiros e **jogar** de novo, uma melhor de três, [...]
([...] he likes to **play** the father in our relationship.)	([...] gosta de **fazer a** figura paterna no nosso relacionamento.

4.3.2 Word lists

There are two types of word lists. One type includes the most frequent words in a corpus. The other is a list of keywords, which are extracted by comparing the word frequency lists of two corpora; the result is a list of words that are typical of one corpus, which are different from the other corpus[9].

Word lists can provide a useful overview of the specific terminology in a field. Of the corpora analyzed, BwanaNet provides lists of words with the option *isolated tokens*. The lists in BwanaNet may be of forms, lemmas or parts of speech. The BYU corpora also generate word lists. In these corpora, the user must introduce a part of speech and the corpora generate word lists that are tagged with that part of speech. The CQP edition of the BNC generates word or lemma frequency lists and allows the user to limit the lists introducing word patterns or using part-of-speech filters. This corpus also generates lists of keywords comparing the frequency lists of the whole BNC, the written BNC, and the spoken component of the BNC.

For example, if we generate a list of lemmas in the BwanaNet English Economy corpus, the first lemmas in the list are, logically, general language words, mainly prepositions and articles, since these are the most frequent words in every corpus. However, the sign = appears at the top of the list, as a typical component of economic texts. Other words from this field, such as *rate, market, price, good, capital, investment,* etc, also appear near the beginning.

4.3.3 List of synonyms

Some corpora have incorporated semantically-based searches. This option allows the user to find synonyms for the word introduced. Of the corpora analyzed in this study, only the BYU corpora provide this option.

[9] These list types are extracted from specialized corpora which are compared with general language corpora, known as *reference* corpora.

In the following example, we use the BYU Corpus of Contemporary American English to search for synonyms of the form *beautiful*, by introducing [=beautiful]. A list of synonyms is provided: *wonderful, attractive, striking, lovely, handsome,* etc. The frequency of each synonym in the corpus and access to concordances of the synonyms are given. We can also compare the frequency and distribution of the synonyms in the corpus by text type and dates.

4.3.4 List of collocates

A collocate is a word that frequently appears near another word. Lists of collocates can be useful to access the words in the context of a term without having to read all its concordances. This function helps to speed up the search process.

In all the corpora, collocates of a word can be seen by reading all the contexts of that word. However, of the corpora analyzed, only the BYU corpora generate lists of word collocates in which the part of speech of the collocate is specified. For example, if we search in the BYU Corpus of Contemporary American English for the noun collocates of the form *television*, the retrieved list of collocates includes: *radio, news, show, cable, network, station, series,* etc.

Collocates of word synonyms can also be accessed. For instance, the BYU Corpus del español lists the nouns that appear near the synonyms of *sucio* (dirty); retrieved collocates include the words *pocilga* (pigsty) or *tugurio* (hovel or dive).

4.3.5 List of clusters

Clusters are sequences of two or more words that are frequent in a corpus. Various query probes can be introduced in a search for clusters. We may choose not to specify a query probe and only specify the number of words we want the cluster to have (two or more). We can specify a word that must appear in the cluster, for instance *mesita* (table). We can also specify the grammatical sequence of the cluster, for example clusters of "noun+adjective+adjective". Words and parts of speech that must be included in the clusters can also be specified, for example "mesita+preposition+adjective".

Lists of clusters can provide a useful overview of how terminology is frequently combined in a field. They can also be used to find a word if we know other words it is frequently combined with, or a typical construction in which the word appears.

Of the corpora analyzed, BwanaNet generates two-word clusters without specifying a query probe. The BYU corpora retrieve clusters specifying words or parts of speech that must appear in the clusters. CREA and CORDE generate clusters specifying one or more words that must appear in the clusters. For example, in the CREA corpus we can search for clusters of three words that include the word *mesita* (table). The retrieved list includes the following clusters: *mesita de noche* (bedside table), *mesita de madera* (wooden table), *mesita de luz* (lamp table), *mesita del teléfono* (telephone table), etc.

5. Conclusion

This study has shown how search techniques can vary from one corpus to another. Within the context of translator training in the use of corpora, there is a need to systematize the search techniques that can be used in electronic corpora. The classification of search techniques provided in this paper, focusing on the query probe, resource and outcome, attempts to meet that need. These three elements have been considered to explore the range of search possibilities corpora offer.

Table 7. Classification of search techniques in corpora

QUERY PROBE	QUERY RESOURCE	QUERY OUTCOME
- Lexical expression - Grammatical expression - Numbers - Hybrid expression - Non-continuous combination of expressions ○ Probe filters	- Monolingual corpora - Aligned corpora - Tagged corpora ○ Resource filters	- List of monolingual or bilingual concordances - Word list - List of synonyms - List of collocates - List of clusters
SEARCH TECHNIQUES		

Although our search technique classification is subject to further additions and variations, it has two main applications. First, it will help us to reflect on the most useful search techniques for translators, thus enabling us to consider improvements in corpora to adapt these resources to translators needs. Second, it may serve as a guide in teaching translation students search techniques in electronic corpora.

6. Corpora examined

British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at <http://www.natcorp.ox.ac.uk/> [20/05/09].

Bwananet, Programa de explotación del corpus técnico del IULA (Universitat Pompeu Fabra). Available at < http://brangaene.upf.es/bwananet/indexes.htm> [22/05/09].

COMPARA. Available at < http://www.linguateca.pt/COMPARA/> [19/05/09].

Davies, M. (2002-). *Corpus del Español* (100 million words, 1200s-1900s). Available at <http://www.corpusdelespanol.org> [16/05/09].

Davies, M. (2004-). *BYU-BNC: The British National Corpus.* Available at <http://corpus.byu.edu/bnc>[16/05/09].

Davies, M. (2007-). *TIME Magazine Corpus* (100 million words, 1920s-2000s). Available at <http://corpus.byu.edu/time> [16/05/09].

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA)*: 385 million words, 1990-present. Available at <http://www.americancorpus.org> [16/05/09].

Davies, M. (2009-). *BYU-OED: The Oxford English Dictionary.* Available at <http://corpus.byu.edu/oed> [16/05/09].

Davies, M. and M. Ferreira. (2006-). *Corpus do Português* (45 million words, 1300s-1900s). Available at <http://www.corpusdoportugues.org> [16/05/09].

Hellenic National Corpus (HNC). Available at < http://hnc.ilsp.gr/en/info.asp> [19/05/09].
Real Academia Española: *Corpus de referencia del español actual* (CREA) *Corpus de referencia del español actual.* Available at <http://www.rae.es> [01/05/09].
Real Academia Española: *Corpus diacrónico del español* (CORDE). Available at <http://www.rae.es> [01/05/09].

7. References

[1] Alcina Caudet, A. (forthcoming). "Metodología y tecnologías para la elaboración de diccionarios terminológicos onomasiológicos", in A. Alcina Caudet *Terminología y sociedad del conocimiento.* Bern, Peter Lang.

[2] Alcina Caudet, A. (2003). "La programación de objetivos didácticos en Terminótica atendiendo a las nuevas herramientas y recursos", in N. Gallardo San Salvador *Terminología y traducción: un bosquejo de su evolución.* Granada, Atrio.

[3] Bowker, L. (1998). "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study", *Meta* 43(4): 631-651.

[4] Bowker, L. (2000). "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources", *International Journal of Corpus Linguistics* 5(1): 17-52.

[5] Bowker, L. and J. Pearson (2002). "Working with Specialized Language. A practical guide to using corpora", London/New York, Routledge.

[6] Corpas Pastor, G. (2004). "Localización de recursos y compilación de corpus vía Internet: aplicaciones para la didáctica de la traducción médica especializada. Manual de documentación y terminología para la traducción especializada", in C. Gonzalo García and V. García Yebra. Madrid, Arcos/Libros: 223-274.

[7] Davies, M. (2005). "The advantage of using relational databases for large corpora. Speed, advanced queries, and unlimited annotation", *International Journal of Corpus Linguistics* 10(3): 307-334.

[8] Faber, P., C. López Rodríguez and M. I. Tercedor Sánchez (2001). "Utilización de técnicas de corpus en la representación del conocimiento médico", *Terminology* 7(2): 167-197.

[9] Fraser, J. (1999). "The Translator and the Word: The Pros and Cons of Dictionaries in Translation", in G. Anderman and M. Rogers *Word, Text, Translation. Liber Amicorum for Peter Newmark.* England, Multilingual Matters.

[10] Gallardo San Salvador, N. and A. de Irazazábal (2002). "Elaboración de un vocabulario multilingüe del campo temático de la siderurgia", in A. Alcina Caudet and S. Gamero Pérez *La traducción científico-técnica y la terminología en la sociedad de la información.* Castellón, Publicaciones de la Universitat Jaume I: 189-198.

[11] Hull, D. (2001). "Software tools to support the construction of bilingual terminology lexicons", in D. Bourigault, C. Jacquemin and M.-C. L'Homme *Recent Advances in Computational Terminology.* Amsterdam/Philadelphia, John Benjamins: 225-244.

[12] Kraif, O. (2008). "Extraction automatique de lexique bilingue: application pour la recherche d'exemples en lexicographie", in F. Maniez, P. Dury, N. Arlin and C. Rougemont *Corpus et dictionnaires de langues de spécialité.* Bresson, Presses Universitaires de Grenoble.

[13] López Rodríguez, C. I. and M. I. Tercedor Sánchez (2008). "Corpora and Students' Autonomy in Scientific and Technical Translation Training", *The Journal of Specialised Translation*, 9. Available at <http://www.jostrans.org/issue09/art_lopez_tercedor.pdf>. [25/05/09].

[14] Meyer, I. (2001). "Extracting knowledge-rich contexts for terminography: a conceptual and methodological framework", in D. Bourigault, C. Jacquemin and M.-C. L'Homme: *Recent Advances in Computational Terminology.* Amsterdam/Philadelphia, John Benjamins: 279-302.

[15] Montero Martínez, S. and P. Faber Benítez (2008). *Terminología para traductores e intérpretes.* Granada, Tragacanto.

[16] Nesi, H. (1999). "A User's Guide to Electronic Dictionaries for Language Learners", *International Journal of Lexicography* 12(1): 55-66.

[17] Sánchez Gijón, P. (2003). *Els documents digitals especialitzats: utilització de la lingüística de corpus com a front de recursos per a la traducció.* Thesis available at < http://www.tdx.cbuc.es/ >. Barcelona, Universidad Autónoma de Barcelona.

[18] Schäffner, C. (1996). "Parallel Texts in Translation". *Unity in Diversity? International Translation Studies Conference.* Dublin City University. 9-11 May 1996

[19] Shreve, G.M. (2001). "Terminological Aspects of Text Production", in S.E. Wright and G. Budin *Handbook of Terminology Management. Volume 2. Application-Oriented Terminology Management.* Amsterdam/Philadelphia, John Benjamins.

[20] Sinclair, J. M. (1996). "EAGLES Preliminary recommendations on Corpus Typology, EAG-TCWG-CTYP/P". Available at <http://citeseer.ist.psu.edu/cache/papers/cs/21540/ftp:zSzzSzftp.ilc.pi.cnr.itzSzpubzSzeagleszSzcorporazSzcorpustyp.pdf/eagles-preliminary-recommendations-on.pdf > [01/05/09]

[21] Snell-Hornby, M. (1988). *Translation Studies. An Integrated Approach.* Amsterdam/Philadelphia, John Benjamins.

[22] Varantola, K. (1998). "Translators and their use of dictionaries", in B. T. S. Atkins *Using Dictionaries.* Tübingen, Niemeyer: 179-192.

[23] Zanettin, F. (1998). "Bilingual Comparable Corpora and the Training of Translators", *Meta* 43(4): 616-630.

[24] Zanettin, F. (2002). "Corpora in Translation Practice", in *Proceedings of the First International Workshop on Language Resources (LR) for Translation Work and Research.* Las Palmas de Gran Canaria.