9

The Per-Fide Corpus: A new Resource for Corpus-Based Terminology, Contrastive Linguistics and Translation Studies

José João Almeida, Sílvia Araújo, Nuno Carvalho, Idalete Dias, Ana Oliveira, André Santos and Alberto Simões

1. Introduction

The Per-Fide project is a joint collaboration between researchers at the Department of Informatics and the Institute of Arts and Humanities at the University of Minho, Portugal. The acronym Per-Fide stands for Portuguese (P) in parallel with 6 languages: English (E), Russian (R), French (F), Italian (I), German/Deutsch (D) and Spanish/Español (E). First, we expound on the role of the Per-Fide project within the context of existing corpora that include the Portuguese language in its different variants – namely, European Portuguese, Brazilian Portuguese and Portuguese spoken in African countries (Angola, Mozambique, Guinea-Bissau, Cape Verde, Sao Tome and Principe). The idea of creating a multilingual parallel corpus project in which Portuguese assumes a pivotal role arose primarily due to the fact that the majority of online corpora that include Portuguese are either monolingual or bilingual. Furthermore, these corpora focus mainly on one specific text type. Consequently, the few multilingual parallel corpora that include Portuguese consist of a relatively small Portuguese subcorpus[1] and provide limited search facilities mainly due to the fact that the Portuguese texts have not been morphologically tagged and/or syntactically annotated. Our second goal in this chapter is to provide an overview of the design criteria for the development of tools and resources in the various stages of the Per-Fide corpora construction process, focusing particularly on automation, validation, generalization and resource sharing. Here, a brief description of the workflow components involved in the pre- and post-alignment phases will be included. Finally, we draw attention to several practical applications of the current features of the Per-Fide corpus in translation practice and contrastive linguistic studies, focusing on the use and potential of probabilistic translation dictionaries and the role that parallel corpora can play in translating idioms.

2. The Per-Fide corpus in the context of Natural Language Processing

In discussing the current status of the computational processing of the Portuguese language, particular mention must be made of the work developed by the Language Resource Center for Portuguese, Linguateca (see Santos, in this volume). Most of the corpora compiled by Linguateca are monolingual, such as the CETEMPúblico corpus and its Brazilian counterpart CETEMFolha, two large corpus collections of articles from the Portuguese newspaper *Público* and the Brazilian newspaper *Folha de S. Paulo*, respectively. The most noticeable parallel corpus project being developed by Linguateca[2] is the Portuguese–English Parallel Translation Corpus COMPARA (Frankenberg-Garcia and Santos, 2002). COMPARA is a bidirectional parallel corpus of English and Portuguese literary text extracts. The majority of Portuguese source texts were written by Portuguese or Brazilian authors, but the corpus also contains texts by Mozambican and Angolan authors. The original English texts were written by authors from the United Kingdom, the United States and South Africa. In some cases, more than one translation of the same source text has been included in the corpus and each corpus text has been annotated with simple text-related metadata consisting of the following elements: author, translator, title, publishing and copyright information, extract start and end page numbers, and the number of tokens, words and types. The

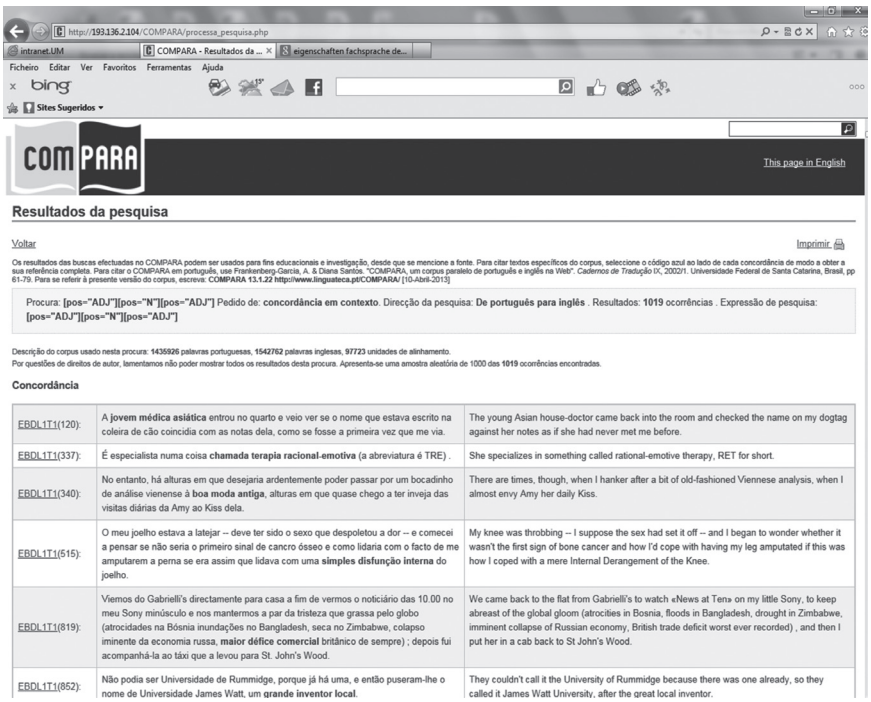

Figure 9.1 The COMPARA search interface and query results

COMPARA corpus can be queried via the DISPARA interface (Santos, 2002), which provides simple and advanced search facilities.[3] To formulate a simple query, users need only define the search direction (Portuguese–English or English–Portuguese) and enter the search term. In simple query mode, all texts in the collection will be searched. The advanced query features include searching by text(s), author(s), publication date and varieties of Portuguese or English. Furthermore, users can specify the query type and, consequently, how the query results are to be presented (see Figure 9.1).

The parallel concordance query displays and highlights all occurrences of a specific word or expression in the source language and its translation equivalent in the sentence-level contexts in which they appear. Since both the Portuguese and English texts are part-of-speech tagged, the parallel concordance query can be further refined by using the part-of-speech tags option. Since part-of-speech-annotated corpora allow users to perform more complex and sophisticated searches involving pattern-based queries, to take advantage of this feature, users must be somewhat familiar with the corpus query syntax,[4] a powerful information retrieval and corpus analysis tool roughly defined in terms of regular expressions consisting of an attribute and its value: [attribute = "value"]. For example, if we want to look for three-word Portuguese sequences beginning with any adjective, followed first by a noun and then by an adjective, the query can have the following form, where 'pos' stands for any part-of-speech tag: [pos = "ADJ"][pos = "N"][pos = "ADJ"]. This pattern captures sequences like *jovem médica asiática* (young Asian doctor), *principais nações industriais* (top industrial nations) and *pequenas tarefas domésticas* (humdrum domestic tasks).

The COMPARA search interface allows for two types of frequency distribution queries:

- text-specific frequency search options, which map the frequency of each word form of a given lemma; and
- corpus-specific frequency search options, which show how the search term is distributed across the texts in the corpus by mapping the search term to textual sources, authors, variety of English or Portuguese and original or translated text. Frequency data for verb tense, person, number and gender are available only for Portuguese.

Moreover, the COMPARA advanced search mode includes a filter option that allows for the specification of alignment constraints. For example, if we want to find all occurrences of the Portuguese word *bonito* that have been translated as 'nice' in the English text collection, this mechanism for narrowing down the search results will exclude the instances in which other translation equivalents, such as 'beautiful,' 'good looking,' 'handsome' and 'pretty' occur.

In terms of the multilingual corpora that include Portuguese, particular emphasis must be placed on the OPUS project, 'a growing multilingual corpus of translated open-source documents available on the Internet' (Tiedemann and Nygaard, 2004, p. 1183). Currently, the OPUS multilingual search interface[5] (Tiedemann, 2012) consists of 16 specialized subcorpora drawn from heterogeneous-specific domains: legislative and parliamentary texts (European Constitution and European Parliament

Proceedings), economic and financial documents (European Central Bank), medical texts (European Medicines Agency), technical computer-related texts/manuals (PHP scripting language, OpenOffice software suite) and subtitle collections (OpenSubtitle.org corpus). European Portuguese is included in seven and Brazilian Portuguese in three of the 16 subcorpora.

Generating a parallel concordance via the OPUS search interface begins with the selection of a subcorpus and a search language. The resulting query panel allows us to perform simple and more complex searches as well as control the concordance output. If we are interested in a bilingual or multilingual query, then one or multiple target languages must be selected. The hits for the search language and the selected target languages will be displayed in the same concordance results window. As such, OPUS is particularly suited to researchers who wish to carry out contrastive linguistics studies in more than two languages. Advanced search options (e.g., lemma, part of speech, phrase structure information) are currently not available for most of the languages of the subcorpora, including Portuguese. For example, in the medical subcorpus, only three of the 22 languages have been lemmatized and part-of-speech tagged: English, French and Italian. As is the case with COMPARA, the OPUS search facility can be optimized using the corpus query syntax, as shown in Figure 9.2.

It is important to note that, in general terms, these two corpus projects follow the same query syntax, but differences occur concerning the attribute and value tagsets[6] used for morphosyntactic annotation and lemmatization. To give a simple illustration,

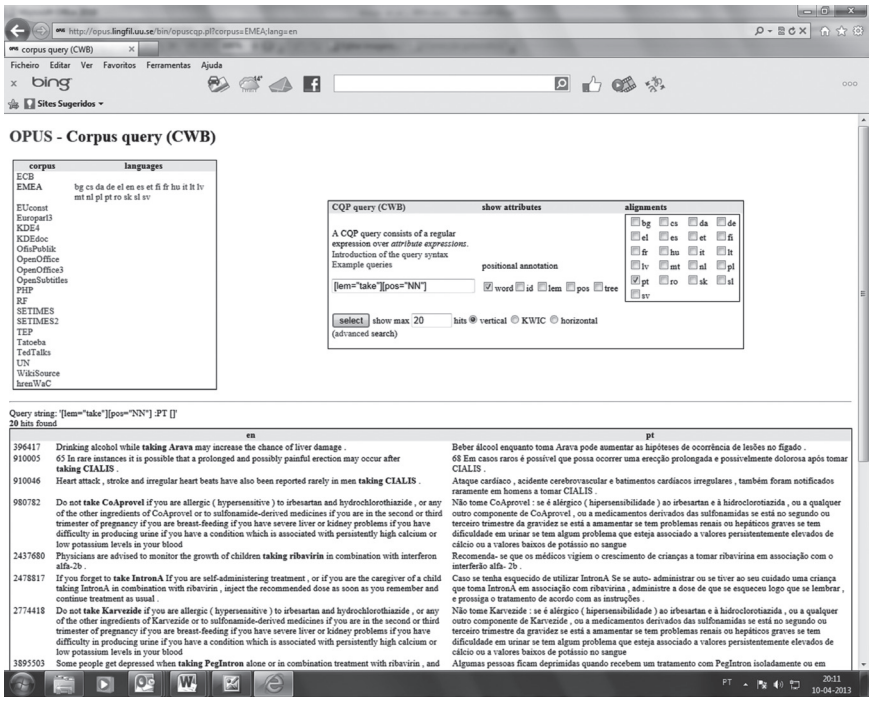

Figure 9.2 The OPUS search interface and query results

suppose we want to find occurrences of the lemma 'take' (as a support verb) (Gross, 1998), followed by a noun, such as 'take care,' 'take offence,' 'take place,' and 'take part.' In COMPARA, this can be achieved using the following query: [lema = "take"][pos = "N.*"]. The same result can be obtained in OPUS using the structure [lem = "take"] [pos = "NN"]. Note the differences between the attributes lema/lem and the tags used to annotate the grammatical class of the tokens: N/NN for nouns. The OPUS interface offers several options for displaying search results: the vertical parallel concordance output and the horizontal KWIC (KeyWord in Context) concordance output. The former displays the sentence-level contexts in which the search term occurs and the corresponding target language sentences in parallel vertical alignment. The latter shows from 5 to 15 words on either side of the keyword/term and the corresponding target language translations in parallel horizontal alignment. If more context is needed to the left and/or right of the query term, the context option allows the user to specify the size of the context in terms of the number of sentences or paragraphs on either side of the term. As with COMPARA, OPUS users can also specify alignment constraints.

One of the main concerns of the OPUS project is to provide the research community with multilingual corpus-based resources and corpus processing tools, such as language-specific taggers and parsers. The parallel corpora prepared in XML Corpus Encoding Standard (XCES) format are freely downloadable from the OPUS project website. This component, which aims to make resources and tools publicly available, is also a key feature of the Per-Fide project.

As noted in the introduction, the Per-Fide project grew out of the need to develop significant multilingual corpus-based resources in which the Portuguese language in its different variants plays a pivotal role. The Per-Fide parallel corpora are bidirectional and Portuguese is always either the search or target language in combination with 6 other languages: Spanish, Russian, French, Italian, German and English. As the resulting language pair combinations (PT ↔ ES/RU/FR/IT/DE/EN) are not commonly found in existing multilingual parallel corpora, interesting contrastive linguistic studies can be performed, including the comparison of lexical, semantic and syntactic patterns between these languages. Another distinctive property of the Per-Fide project is that it covers a wide range of text types: religious texts (main sources: the Vatican, Comboni Missionary Community, Taizé Community), literary texts, official documents and legal texts (JRC-Acquis, EuroParl, EurLex), journalistic texts and technical texts (economics and finance, technology, health and medicine, social sciences, philosophy, tourism, gastronomy). Needless to say, one of the most complicated tasks in the text-selection process is obtaining copyright clearance for both originals and translations. This is particularly true for literary texts.[7]

The heterogeneous nature of the texts also raises key issues concerning the appropriate document classification scheme and metadata handling. Choosing or developing a classification scheme is not a straightforward matter, due in part to the fact that some texts can be classified as belonging to different categories. For example, more often than not, texts with a spiritual and religious content can be considered borderline cases between fiction and non-fiction/literary and religious texts. Furthermore, it was clear from the outset that the project classification scheme had to be flexible enough to incorporate new text types. Preparatory work undertaken

to develop a project-specific classification system included the study of document classification schemes, such as the Universal Decimal Classification (UDC; McIlwaine, 2000) and the UNESCO Thesaurus.[8]

Another main goal of the Per-Fide project is to provide each text with a bibliographic record containing not only basic metadata elements (e.g., title, author, editor), but also more specific items, such as language, translator, text type and literary period. For this purpose, we decided to use the metadata encoding scheme developed by the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 2002). Given that not all project members involved in the text selection and classification process were familiar with TEI, a web interface was designed containing a bibliographic data entry form to generate a TEI header automatically for each text.

3. Corpus Processing Pipeline

In what follows, we describe the design goals that underlie the development and management of the Per-Fide project tools that comprise the Per-Fide corpora pipeline:

- Automation: making all of the corpora processing tasks as systematic and automatic as possible ensures the rapid processing of new texts and reprocessing of existing corpora when a new tool is developed and incorporated or a bug is fixed.
- Validation: much time and effort have been dedicated to developing tools that compute metrics. These tools can be used to evaluate, to a certain extent, the quality of the generated resources.
- Generalization: the development of tools that are general enough to be used in other contexts is an asset-building strategy. Therefore, most of the Per-Fide tools are available to the community through an open-source licence. The overall quality of the tools developed will most certainly benefit from community feedback.
- Resource sharing: the Per-Fide project provides resources on at least three distinct levels. First, the project's web interface allows querying and browsing of the corpora collection and other relevant content, such as Probabilistic Translation Dictionaries (PTDs). Secondly, the available corpora, PTDs and other intermediate resources are also available for public download as files in standard formats. Finally, following the semantic web trends, a RESTful Application Programming Interface (API) (Fielding and Taylor, 2002) is also publicly available, providing queries and operations that allow the implementation of third-party tools that can easily be integrated with Per-Fide resources.

Preparing documents to be included in corpora and enriching the corpora obtained involve several different steps that result in a complex network of dependencies and conditional tasks determined by the original format and state of the documents, the type of resource extraction, and the intended use. The corpus pipeline[9] can be divided into two main phases: pre-alignment and post-alignment. The pre-alignment phase includes tasks such as text cleaning, normalization and the alignment task itself.

Post-alignment tasks include corpora tokenization, segmentation and tagging as well as the generation of derived resources such as the extraction of PTDs and terminology. This phase includes the process of making the corpora available for online querying and download.

Both the large amount of documents and the many alternative and conditional tasks involved reduce the feasibility of manual maintenance. The automation of these processes presents several interesting challenges.

This section describes the two main phases that comprise the corpus-building process. The tasks are currently in different stages of integration in the workflow, ranging from fully integrated and automated steps to prototypes that are still being tried, tested and tuned.

3.1 Pre-alignment phase

The alignment process takes as its input pairs of plain text files. These files must resemble each other very closely in terms of size and structure. If this is not the case, the performance of the alignment tool will decrease.

3.1.1 Cleaning documents

Text documents retrieved for the purpose of being automatically processed often present several types of 'noise,' such as structural residue (e.g. page numbers, headers and footers, footnotes), text encoding and mark-up syntax (e.g. notation used for sections, paragraphs and sentences), which are obstacles to any further use of the texts in the corpus. This is particularly true for the conversion of documents available in portable document format (PDF) to plain text. For example, mathematical formulae and tables can render parts of the resulting textual document completely illegible. In fact, for some documents, the simple conversion of the document to plain text format can be difficult depending on which tool generated the original PDF file. In order to reduce the noise in these documents and its impact on subsequent processing, documents are pre-processed with the Text-Per-Fide-BookCleaner, a tool designed to clean and normalize unwanted elements (Santos and Almeida, 2011). This tool makes use of an ontology of document structure elements to detect and remove unwanted parts.

3.1.2 Finding pairs of documents

One of the challenges of preparing files for alignment is finding, within a large collection, pairs of files to be aligned – that is, pairs of files where each is a translation of the other in a distinct language. Depending on the origin of the candidate documents, it is sometimes possible to extract information for the pairing of the documents from their names or universal resource locators (URLs). This typically happens when the document pairs are retrieved from the same source, like a website, using crawling mechanisms (Almeida et al., 2002). However, when the documents come from a variety of distinct sources, different naming conventions might have been used, which means that methods for finding candidate pairs for alignment must

rely on the contents of the documents. This is particularly the case with literary texts, where the translations in the different languages are obtained from different sources.

A method for solving this problem consists in comparing documents based on language-independent elements (LIEs) (Santos, 2011). Examples of such elements are year references (e.g., 1755) and proper names (e.g., Hamlet). The set of LIEs in every file is extracted and the sets compared to each other. Files presenting a high percentage of LIEs in common are proposed as candidate pairs for alignment. This approach is similar to that used for the word alignment of parallel corpora (Tiedemann, 2003). Although this method will help organize and detect document pairs from the set of documents collected, some prior organizational efforts will greatly contribute to achieving document pairing. Therefore, the collected documents were added into a collaborative repository (SVN) and stored in a hierarchical tree of directories, making the task of detecting document pairs easier.

3.1.3 Synchronizing documents

Text-alignment tools are very sensitive to differences in the documents resulting from the insertion and/or deletion of text and the inversion of the order of paragraphs or sentences. It is quite common to see entire sections of books, such as biographical notes and prefaces to a given edition, absent in the translations. Such cases often render the entire alignment unusable: once the aligner tool desynchronizes, it is very difficult to synchronize later on in the alignment process.

Document synchronization, then, is a process of aligning two documents at the section level (Santos et al., 2012). This process can be used either to define hard anchor points (which the aligner can use for synchronization purposes) or to split the original pair of documents into smaller parallel chunks. It also allows for the identification of non-matching sections that can be removed later.

Another alternative for defining anchor points consists in the use of bilingual dictionaries or similar resources, like the Unambiguous-Concept Translation Sets (UCTSs) (Santos et al., 2012). As will become clear later on, the advantage of this approach is that users do not need to identify the document sections, only a set of synchronization points. The alignment itself is carried out using the easy-align tool that is part of the Open Corpus Workbench package (Evert and Hardie, 2011). The Per-Fide workgroup is currently analysing the alignment quality of another aligner, namely HunAlign (Varga et al., 2005).

3.2 Post-alignment phase

Once the texts have been aligned, they can be given as input to the next workflow component available in the Per-Fide environment. The alignment process typically culminates in the generation of a parallel corpus, encoded using the Translation Memory Interchange (TMX) format. An example of two alignment entries in the TMX file follows. Note that neither the document header nor footer is shown. The extract in Figure 9.3 is taken from a corpus composed of free software internationalization messages.

```
    <tu>
      <tuv xml:lang="PT">
        <seg>Alternar breakpoint no cursor</seg>
      </tuv>
      <tuv xml:lang="EN">
        <seg>Toggle breakpoint at cursor</seg>
      </tuv>
    </tu>

    <tu>
      <tuv xml:lang="PT">
        <seg>Intervalo de gravação de sessão em
minutos</seg>
      </tuv>
      <tuv xml:lang="EN">
        <seg>Save session interval in minutes</seg>
      </tuv>
    </tu>
```

Figure 9.3 Pair of translation units in TMX format from a corpus of free software internationalization messages

As soon as these TMX files are added to the Per-Fide corpora collection, an automatic procedure defines the processing operations needed to build the commonly available resources. These instructions are stored in a standard Unix Makefile.[10] There are two main advantages for adopting this technology:

(a) Many operations described in the Makefile are computationally intensive and can take several days to compute without parallelization. Thus, one of the goals is to minimize the number of operations to be performed, making sure that an operation is only executed when required, either because the resource is not available, or a resource has changed and dependent resources need to be updated. Makefiles allow for the specification of dependencies in chains and the built-in rules verify whether a resource (file) really needs to be calculated.
(b) It is possible to take advantage of many other built-in features of Makefiles. One of these features is used to parallelize long operation sequences. This means that many computations can be performed simultaneously to save time. For example, one might divide the corpus into smaller chunks and annotate each chunk using a parallel process, thus making the full annotation process faster, much like the well-known MapReduce technique (Lin and Dyer, 2010).

Another major concern is that all the operations described need to be language-aware. The computed resources need to be created for every language and every language pair available in the TMX file. In practice, this is achieved by using a tool that manages all available resources and Makefiles. It also provides a web interface for the administration of the project environment. Moreover, it is general enough to be used in other contexts, outside the scope of this project. To implement this tool it was necessary to

devise a set of documents, specifically a corpus manifesto, that lists, for each corpus, all the files related to it.

3.2.1 Segmentation and Tokenization

Segmentation is the task of dividing text into sentences (or other segments), whereas tokenization is the task of dividing sentences into tokens. Although these tasks have a precision level of 99% on most tools, they are not straightforward processes. In Per-Fide, we decided to use a natural language processing library, named FreeLing (Padró, 2011), to perform segmentation and tokenization operations. Although FreeLing does not currently support all of the languages in the Per-Fide corpora, its syntax for defining new segmentation and tokenization tools is quite simple and extensible.

Once TMX files with the desired languages are available, tokenization can take place. This step is carried out earlier in the workflow as many subsequent operations can take advantage of tokenization. The tool developed for the tokenization of TMX files is aware of the TMX annotation and applies the correct tokenization module for that language.

3.2.2 Part-of-speech tagging

Corpora become particularly valuable resources when annotated (or tagged) with part-of-speech information. Here, again, the NLP library FreeLing was used. It includes two different part-of-speech (PoS) tagging approaches that can be used for this task: a method based on Hidden Markov Models (tagger training based on sequences of annotations and the prediction of the more probable annotation that follows, based on the possible PoS chosen by the morphological analyzer) and another based on the Relax algorithm (relaxation labelling is a family of energy-function-minimizing algorithms that change the labelling in accordance with a set of constraint rules; Manning and Schütze, 1999). Unfortunately, FreeLing does not include data for both methods for all supported languages, therefore the method was chosen depending on the language and the tagging data available. When both methods are available for a specific language, the Hidden Markov Model was chosen. FreeLing is also able to detect and tag different kinds of constructions, like proper names and locutions (multi-word terms). Finally, a Chart Parser is also available, allowing for the annotation of tree structure information, but this kind of annotation has not yet been included in our corpora. Currently, tools are being developed for the annotation of the TMX files. This is being done in such a way as to allow the annotated files to be encoded later in the IMS Corpus Workbench (see below).

3.2.3 Probabilistic Translation Dictionaries

Probabilistic Translation Dictionaries (PTDs) are translation dictionaries that map words from one language to a set of possible translations. Each of these translations has a probability value. Further details on how PTDs are computed can be found in Simões and Almeida (2003). A PTD entry example extracted from a Portuguese–English dictionary is shown in Figure 9.4.

$$codificada \begin{cases} codified\ (62.83\%) \\ uncoded\ (13.16\%) \\ coded\ (6.47\%) \\ \ldots \end{cases}$$

Figure 9.4 A Probabilistic Translation Dictionary (PTD) entry

In the corpus pipeline, PTDs are computed for every language pair available. In addition to computing one PTD for every language pair, an accumulated PTD is also compiled that aggregates all the PTDs calculated, in order to improve both the coverage and the quality of the dictionary. The different PTDs can then be browsed using the internet interface, or the files can be downloaded for offline processing.

3.2.4 IMS Corpus WorkBench indexing

The IMS Corpus WorkBench is undoubtedly the most widely used tool for managing and indexing corpora for fast querying. For each TMX file, we create a monolingual corpus for every available language, as well as a parallel corpus for each language combination.

This step is crucial for the web interface to be able to rapidly query the available corpora. When a corpus is queried, either through the web interface (by humans using a browser) or the web service (more oriented to machine–machine communication), all the resources required to answer the query have already been computed. This greatly improves the query response time, and even queries that return many millions of hits are displayed almost immediately.

3.2.5 Unambiguous-concept translation sets

The translation of certain terms or lexical units does not pose ambiguity problems; in principle, they are always translated the same way. We call these 'unambiguous concepts.' Examples include:

- proper names (London_{en} ~ Londres_{pt}; Oporto_{en}, Porto_{en} ~ Porto_{pt});
- technical terminology (file_{en} ~ ficheiro_{pt} ; folder_{en},directory_{en} ~ pasta_{pt},directoria_{pt});
- possible synonyms (wolfram_{en}, tungsten_{en} ~ volfrâmio_{pt});
- morphological agreement constraints that need to be kept (Israel_{pt} ~ Израиль_{ru} (nominative or accusative case), Израилем_{ru} (instrumental case), Израиля_{ru} (genitive case), Израилю_{ru} (dative case)); and
- months, seasons, weekdays, numerals, cardinals, etc.

Unambiguous-concept translation sets (UCTSs) can be exported/extracted from resources, produced manually, or extracted automatically from PTD files. They can be used for such tasks as partial synchronization or alignment as well as the assessment of the alignment process (in the translation sector, terminology is also used for quality assurance).

3.2.6 Bi-word sets

A bi-word set (BWS) is a collection of strongly related word pairs. Each bi-word '$word_{L1}$, $word_{L2}$' tells us that $word_{L2}$ is a possible translation of $word_{L1}$ (although there may be other translation candidates). The main difference between UCTSs and BWSs is the notion of (un)ambiguity. In BWSs:

- terms can be ambiguous;
- relations are unidirectional; and
- one term can appear in more than one entry.

In UCTSs, on the other hand:

- lexical units are expected to be translated by a term belonging to a small set of well-defined concepts;
- each term in a UCTS list can only be found in exactly one UCTS; and
- relationships are bidirectional.

BWSs include pairs of words whose relation is not so strong as with UCTSs: $rest_{en}$ = $descanso_{pt}$, $rest_{en}$ = $descansar_{pt}$, $rest_{en}$ = $repouso_{pt}$, $rest_{en}$ = $pausa_{pt}$, $pause_{en}$ = $pausa_{pt}$, $break_{en}$ = $pausa_{pt}$

UCTS and BWS can be used for the analysis of translation (or alignment) quality. Consider, for example, the UCTS that defines the translation of 'Oporto' as *Porto*. If, after extracting a translation dictionary from a corpus, *Porto* is missing from the translations of 'Oporto', it can be deduced that something went wrong in the alignment process.

BWSs are more useful during the alignment process. They can be used as clues (soft anchor points) for the aligner tool. See, for instance, Tiedemann (2003) for a discussion on clue-based alignment procedures.

3.2.7 Evaluation, metrics and quality

The automatic evaluation of the alignment process and the derived resources is particularly challenging when dealing with large amounts of files and data as it is difficult to identify what metrics to use to infer their quality. Nonetheless, some elements can be measured and provide clues to evaluate the quality of the resource.

One of these, which refers to the evaluation of TMX files, is the metric known as the percentage of 1:1 correspondences (i.e. a single sentence aligned with one single sentence) versus other types of correspondences. Although non-1:1 correspondences can occur in correctly aligned texts, a high percentage of these usually indicates a low-quality alignment.

Another method of evaluating alignments is to check for the presence of UCTSs in translation units, as illustrated in the previous section. If a term from an UCTS appears in one language segment from a translation unit then one of the accepted translations should appear in the aligned language segment. If not, it is highly probable that the translation unit has been translated incorrectly. Even without the same level of confidence as with UCTSs, this same kind of approach can be performed using BWSs or PTDs.

3.2.8 Using resources to improve our tools

As a corollary to the previously mentioned design goals concerning generalization and resource sharing, our tools are often implemented as clients of each other – namely, resources generated by some tools can be used by other tools in order to improve results. For example, although the UCTS extraction requires corpora alignment, the alignment can take advantage of extracted UCTSs for a better alignment quality. Therefore, one can align a corpus, extract UCTSs and use them to re-align the same corpus.

The UCTSs generated from PTDs can be used in the above-mentioned process of document synchronization to split the texts before the alignment as well as to synchronize the alignment tools.

3.2.9 Query interface

Once all of the resources have been calculated and made available, they can be immediately queried by any user. This can be done using the Per-Fide query internet interface.[11] The project environment also provides a web service that applies a set of operations, via an easy-to-use RESTful public interface (API), to available resources that can be incorporated into other tools to build more complex applications. This web service provides a set of well-defined online operations for querying project resources. The query results are provided in a set of well-defined XML schemas. This is a useful component for other tools that want to take advantage of the resources described and keep up to date with the new resources being built and/or updated. In addition to all of these query options, all of the linguistic resources are available for download, either for offline operations or for building new resources.

4. Applications of the Per-Fide Corpus in Cross-linguistic Research

There is an increasing research interest in corpora and their applications and the potential for development in this area is immense. As stated by Granger (2010, p. 7), any field that relies on the analysis of two or more languages can benefit from corpus-based cross-linguistic research. In addition to the undeniable utility of parallel concordances in translation studies, bilingual lexicography and machine translation (Granger et al., 2007), their potential for second-language learning is enormous and there is a significant body of literature demonstrating how language learners can benefit from their use (e.g., Aston, 2001; Granger, 2003; Sinclair, 2004; Frankenberg-Garcia, 2005).

Students at the beginner and intermediate levels are still very dependent on dictionaries. The use of concordances as a tool for language learning at the beginner's level can be motivating and rewarding for learners because this tool can provide contextualized examples that encourage such students to explore the meanings and uses of words in authentic contexts (cf. St. John, 2001). Portuguese learners of French who look up *démarche* in a Portuguese–French bilingual dictionary, for example, will

encounter several possible translations for the word (*modo de andar*, *passo*, *attitude*, *comportamento*, *modo de pensar*, *procedimento*, *diligências*, *trâmites*, etc.) and might find it difficult to choose which term to use. If they look up *démarche* in the L1–L2 direction of a Portuguese–French parallel corpus like Per-Fide, they will not only be able to see different ways in which *démarche* has been rendered in Portuguese, but also the different contexts in which each term was used (see Examples 1a through 7b).

(1a) *[…] la communauté internationale est invitée à suivre cette <démarche>.*
(1b) *[…] a comunidade internacional é convidada a seguir esta <iniciativa>.*
(2a) *Les analyses prospectives et socio-économiques représentent une partie importante de la <démarche>.*
(2b) *Uma parte importante da <abordagem> é constituída por análises prospectivas e socioeconómicas.*
(3a) *Une telle <démarche> se heurterait cependant à deux inconvénients majeurs.*
(3b) *Um <procedimento> deste tipo levanta, contudo, dois inconvenientes significativos.*
(4a) *La Région flamande aurait suivi la même <démarche> et fixé ses propres objectifs de qualité.*
(4b) *a Região da Flandres teria feito a mesma <diligência> e fixado os seus próprios objectivos de qualidade.*
(5a) *Il est important que la Cour adopte une <démarche> cohérente pour décider d'exercer ou non sa compétence.*
(5b) *É importante que o Tribunal de Justiça adopte uma <posição> coerente quando decidir se deve ou não considerar-se competente.*
(6a) *[…] la <démarche> de la Commission n'appelle aucune réserve […]*
(6b) *[…] a <atitude> da Comissão não levanta quaisquer reservas […]*
(7a) *Une <démarche> anormale et des chutes ont été des événements indésirables très fréquemment rapportés avec olanzapine.*
(7b) *Os efeitos adversos muito frequentes associados com o uso da olanzapina neste grupo de doentes, foram perturbações na <marcha> e quedas.*

It is clear that, used as a complement to (or instead of) monolingual or bilingual dictionaries, parallel concordances can help learners understand foreign words they do not know as well as the contexts in which the words are appropriate (Frankenberg-Garcia, 2005, p. 191). For the verb *implementar*, the dictionary Infopédia[12] offers the following equivalents in French: *accomplir, exécuter* and *implémenter*. The bilingual concordance depicted below, which we extracted from the Per-Fide corpus, illustrates the different possibilities of translating the Portuguese collocation (cf. Iriarte Sanromán, 2001; Grossmann and Tutin, 2003) *implementar medidas* into French. Note that none of the French verbs proposed by Infopédia appear in the bilingual concordance (see Examples 8a through 13b).

(8a) *Existem procedimentos para desenvolver e <implementar medidas> de controlo de riscos.*
(8b) *Il existe des procédures pour élaborer et <instaurer des mesures> de maîtrise des risques.*

(9a) *[…] os Estados-Membros deverão <implementar medidas> eficazes de acompanhamento e controlo.*
(9b) *[…] les États membres devraient <mettre en place des mesures> efficaces de suivi et de contrôle.*
(10a) *Insiste-se aqui na necessidade de <implementar medidas> destinadas a preservar a quantidade dos recursos naturais.*
(10b) *On insiste ici sur la nécessité de <mettre en œuvre des mesures> destinées à préserver la quantité des ressources naturelles.*
(11a) *[…] a Opel Nederland considerou necessário <implementar medidas> de neutralização em Outubro e em Dezembro de 1996.*
(11b) *[…] Opel Nederland a jugé utile de <prendre des mesures> correctives en octobre et en décembre 1996.*
(12a) *Se se tornar necessário <implementar medidas> fiscais para alcançar os objectivos acordados, então, em minha opinião, esta via terá fracassado.*
(12b) *S'il devait s'avérer nécessaire d'<introduire des mesures> fiscales pour atteindre les objectifs convenus, cette voie serait alors pour moi un échec.*
(13a) *Cada parte tem o direito de adoptar e <implementar medida>s mais rigorosas do que as enunciadas nas disposições da presente convenção.*
(13b) *Chacune des parties contractantes a le droit d'adopter et d'<appliquer des mesures> plus rigoureuses que celles qui sont énoncées dans la présente convention.*

Whereas no single case of the French verb *implémenter* followed by the phrase *des mesures* was found in the Per-Fide corpus, the English translation of the Portuguese locution *implementar medidas* is almost exclusively translated as the combination of the verb 'implement' and the noun 'measures' (see Examples 14a through 15b).

(14a) *Insiste-se aqui na necessidade de <implementar medidas> destinadas a preservar a quantidade dos recursos naturais.*
(14b) The emphasis here is on the necessity of <implementing measures> to preserve the quantity of natural resources.
(15a) *Cada parte tem o direito de adoptar e <implementar medidas> mais rigorosas do que as enunciadas nas disposições da presente convenção.*
(15b) Each Contracting Party has the right to adopt and <implement measures> being more stringent than those resulting from the provisions of this Convention.

The fact that parallel concordances provide not only linguistic equivalents, but also the contexts in which different terms are equivalent (*prendre* / *introduire* / *appliquer* / *mettre en œuvre* / *mettre en place … des mesures*), which can help learners decide which term is appropriate in a specific context. This tool can be especially helpful when learners or translators have to deal with idiomatic expressions for which there are no simple, direct translations available in their mother tongue.

The following examples, taken from the Per-Fide corpus with French as the source language, demonstrate that – when confronted with an idiomatic phrase such as *couper les cheveux en quatre* (to split hairs) – translators can draw upon a number

of different translation alternatives which, in this case, might be more or less synonymous (see Examples 16a through 19b).

(16a) *Les chercheurs sont des gestionnaires, des ingénieurs, des collectionneurs, des <coupeurs de cheveux en quatre>, ou des artistes.*
(16b) *Os investigadores podem ser gestores, engenheiros, coleccionadores, <picuinhas>, fantasiadores ou artistas.*
(17a) *Monsieur le Président, il ne faut pas <couper les cheveux en quatre>, comme disent les Français.*
(17b) *Senhor Presidente, não devemos <perder-nos em subtilezas>, como dizem os franceses.*
(18a) *Au Conseil de ministres, je dirai: arrêtez de <couper les cheveux en quatre>.*
(18b) *Ao Conselho de Ministros direi: <Não se prendam com coisas insignificantes>.*
(19a) *C'est essentiel afin que le Parlement européen ne devienne pas un lieu où l'on <coupe les cheveux en quatre> mais un lieu que la Commission et le Conseil prennent au sérieux.*
(19b) *É essencial para que o Parlamento Europeu não seja um local de <debate fútil> e para que a Comissão e o Conselho o levem a sério.*

These examples clearly show that idioms are 'one of the most relevant manifestations of the creative potential of any language, as evidenced by the richness of their images, the originality of their metaphors as well as the variety and malleability of their structure' (Alvarez, 2007, p. 160). Indeed, it is not always easy to grasp the metaphorical nuances of such phrases, as demonstrated in Examples (20a) and (20b), in which the proposed translation was too literal and therefore unable to render the meaning of the original structure.

(20a) *Le commissaire a évoqué la nécessité d'un mouvement proactif dans le sens de la prestation de services, mais nous n'allons pas commencer à <couper les cheveux en quatre>; au contraire, nous devons faire quelque chose de productif.*
(20b) *O Senhor Comissário falou da necessidade de medidas proactivas dirigidas à prestação de serviços, mas não podemos dedicar-nos todos <a cortar o cabelo uns aos outros>; pelo contrário, temos de produzir algo também.*

Idioms are unquestionably a critical area of languages in general, inasmuch as their global meaning cannot be apprehended by aggregating the individual meaning of each constituent. When combined, isolated lexemes generate new meaning (Mejri, 1997), which results in the construction of syntagmatic structures whose figurative value refers to a specific reality with a particular meaning. Thus, it comes as no surprise that the translation of such structures should pose such a multitude of challenges.

Since Google Translate is one of the most widely used translation resources, being used on a fairly regular basis by translators, we decided to test the quality of the translation of idioms provided by this machine translation system. As can be seen in Examples (21a) through (21e), the French idiom *couper les cheveux en quatre* (to split hairs) has been incorrectly translated into Portuguese, Spanish, English and German (besides the incorrect translation of the expression *couper les cheveux en quatre*, Example (21e) presents an incorrect translation of the phrase *comme disent les Français*).

(21a) French: *Monsieur le Président, il ne faut pas <couper les cheveux en quatre>, comme disent les Français.*
(21b) Portuguese: *Sr. Presidente, <não cortar os cabelos>, como dizem os franceses.*
(21c) Spanish: *Señor Presidente, <no corte los pelos>, como dicen los franceses.*
(21d) English: Mr. President, <do not cut the hairs>, as the French say.
(21e) German: *Herr Präsident, <nicht schneiden die Haare>, wie die Französisch Wort.*

Google Translate returns literal translations of the lexical units that make up the French idiom. In translating the idiom *poser un lapin* (to stand someone up), Google Translate does not provide a literal translation of the elements that comprise the French expression (*poser un lapin à quelqu'un*, literally 'to put a rabbit to someone') in any of the four target languages (see Examples 22a through 22e), but the results leave much to be desired. Note that the surrounding context does not point the system toward a more appropriate translation, as would be expected.

(22a) French: *Je l'ai attendue tout l'après-midi. Elle n'est pas venue à notre rendez-vous. Elle m'<a posé un lapin>.*
(22b) Portuguese: *Esperei toda a tarde. Ela não veio ao nosso encontro. Ela <se levantou>.*
(22c) Spanish: *Esperé toda la tarde. Ella no vino a nuestro encuentro. <Se puso de pie>.*
(22d) English: I waited all afternoon. She did not come to our rendezvous. She <stood up>.
(22e) German: *Ich wartete den ganzen Nachmittag. Sie wollte nicht zu unserem Treffpunkt kommen. Sie <stand auf>.*

Thus, it becomes clear that the ability to query corpora is, beyond any doubt, a feature that needs to be introduced in training courses for translators and other professionals in related fields.

The Per-Fide search interface allows for simultaneous L1 and L2 queries. This kind of search can be useful, on the one hand, for determining if a word in L1 corresponds to another word in L2, such as if *casa* (house) can be translated as 'box' and, if so, in which contexts; on the other hand, it can be used to identify the various equivalent terms of the word in L2, such as the word *casa* corresponding to 'box', 'home', 'house', 'place', 'section', etc. In order to narrow down the search in L1 to a particular concept (e.g., *casa* referring to a type of housing facility), we can choose corresponding terms, such as 'home' or 'house' in L2, and all the occurrences featuring these last two terms will be retrieved (PT: *casa* – EN: home / house; see Table 9.1).

The search query involving the word *casa* can be refined if we wish, for example, to determine lexical patterns in which the word *casa* is followed by the preposition *de*. The Per-Fide query interface will return the instances of *casa de* as shown in Table 9.2.

The results of the simple search *casa* and the more refined search *casa de* enable us to see that the Portuguese multi-word lexical unit *casa de férias* has two different English translation equivalents: 'holiday home' and 'resort home'. It would be an interesting task to use the Per-Fide search facilities to look into the semantic nuances of the English units, which is beyond the scope of this chapter.

Table 9.1 Search for *casa*

Portuguese	English
casa	**home**
casa particular	private home
casa natal	native home
casa de morada de família	family home / matrimonial home
casa de férias	holiday home
[…]	[…]
casa	**house**
casa das máquinas	wheel house
casa familiar-tipo	typical family house
casa provincial	provincial house
casa do clero	priest's house
casa de leilões	auction house
[…]	[…]

Table 9.2 Occurrences of *casa de* in the Per-Fide corpus

Portuguese	English
casa de fim-de-semana	weekend-house
casa de férias	resort-home
casa de hóspedes	boarding house
casa de acolhimento	guest house
[…]	[…]

When searching for multi-word lexical units containing words connected by prepositions, it is possible to specify which prepositions we wish to include in the search as well as their various contracted forms, such as *máquina* ('machine') followed by the preposition *de* (of) or the contractions *do*, *da*, *dos* and *das* ('of' plus a definite article). The following formulae can be used to perform this search:

(a) *máquina* d.
(b) *máquina* (*de*|*da*|*do*|*dos*|*das*)

In search query (a), the period operator (.) matches any single character. Therefore, only occurrences that feature the word *máquina* followed either by the preposition *de* or the singular contracted forms *da* or *do* will be retrieved. In query (b), the disjunction operator (|) expresses an alternative. In this case, instances of the word *máquina* followed by all the specified singular and plural contracted forms will be matched. Both queries (a) and (b) will only retrieve occurrences that include the word *máquina* followed by the preposition *de* and the specified contracted forms (see Table 9.3).

Regardless of the fact that the Per-Fide corpus is still under construction and we are currently working on the part-of-speech tagging, it is already possible, as demonstrated

Table 9.3 Search results for occurrences with *máquina* (*de* | *da* | *das* | *do* | *dos*)

Portuguese	English	French
máquina de escrever	typewriter	*machine à écrire*
máquina de lavar loiça	dishwasher	*lave-vaisselle*
máquina de lavar roupa (para uso doméstico)	(household) washing machine	*lave-linge (ménager)*
máquina de barbear eléctrica sem cabeça	electric shaver with the head removed	*rasoir électrique sans tête*
máquina de depilar com motor eléctrico incorporado	hair-removing appliance with self-contained electric motor	*appareil à épiler à moteur électrique incorporé*
máquina de escolha de notas com retalhadora integrada	banknote sorting machine with an integrated shredder	*machine de tri équipée d'un broyeur intégré*
[…]	[…]	[…]

in the previous examples, to run simple single-word or multi-word search queries. Moreover, the corpus offers a supplementary resource for corpus querying: PTDs. By generating a PTD, the query system provides a paradigmatic family of functional equivalents along with the respective percentage of direct correspondence between the source term and various possible target terms within the selected corpus (see the following example of PTDs in the EuroParl corpus) and in all the corpora included in Per-Fide (see Figure 9.5).

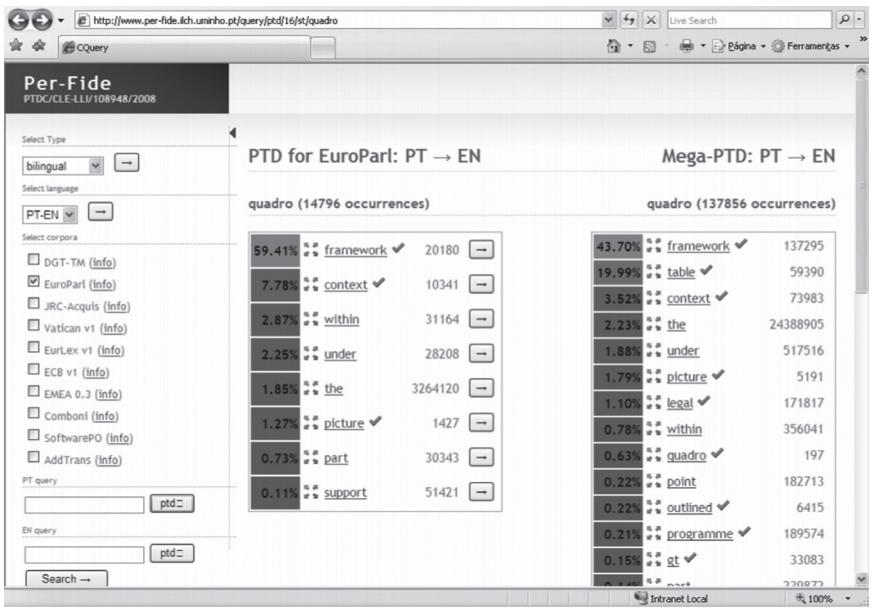

Figure 9.5 Micro- and Mega-PTD for the word 'frame'

The option 'ptd' was selected to initiate the search for the English word 'frame,' which, as can be seen in Figure 9.5, has several equivalents in Portuguese. The results obtained with this option allow users to visualize the translation alternatives of the source term in different contexts, which they access by clicking on the arrow located alongside each term listed in the PTD. PTDs can be used to compile terminological lists, which is potentially useful for the production of bilingual glossaries or dictionaries. It goes without saying that the list of terms generated by the PTD becomes highly valuable to translators if enhanced with the study of the term in context. In other words, the work of the translator can benefit significantly from switching between the isolated terms given by the PTD and their contextualized bilingual concordance. Users can begin by choosing the bilingual concordance without even looking at the PTD (as we saw in the simple query of idioms). In the bilingual concordance, users must skim through all of the occurrences, which might reach the hundreds, in order to identify the term that best suits the context of the word they wish to translate. Thus, users who activate the PTD as a query option can increase the efficacy and output efficiency of their work by immediately circumscribing the range of alternatives at their disposal to translate a given expression. Hence, it is essential to expand the corpus both qualitatively (text types) and quantitatively (amount of bi-texts) so that the terms listed in the PTD are representative of as many usage contexts as possible.

Mega-PTDs aim to complement the data supplied by the PTD of a single corpus (micro-PTD) and can encompass contexts that the micro-PTD was unable to retrieve. In the Mega-PTD, it is possible to depart from an L1 to an L2 and revert to L1. For example, one of the equivalents of 'frame' is the term *quadro*; if we click on the latter,

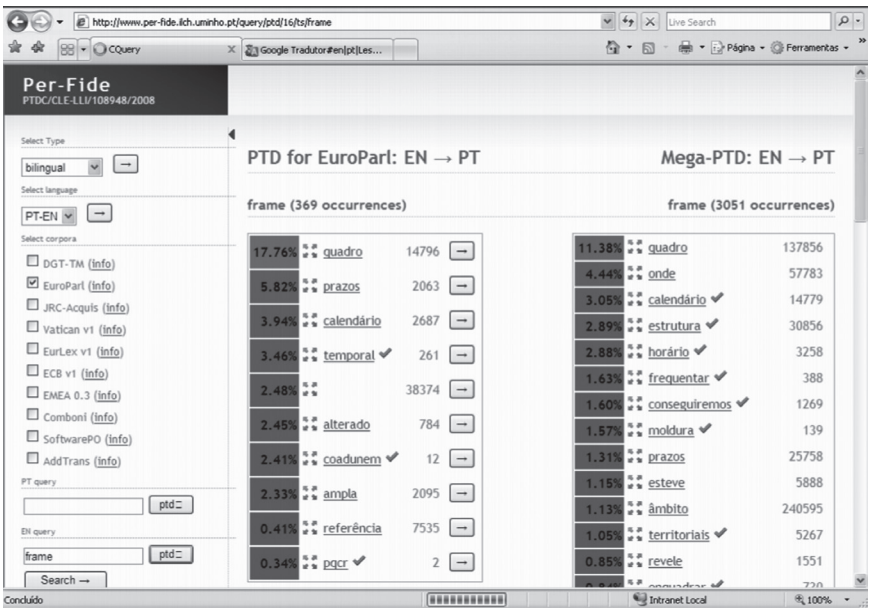

Figure 9.6 Micro- and Mega-PTD of *quadro*

Table 9.4 Micro-PTD of *casa* in different subcorpora of the Per-Fide corpus

Comboni	**Vatican**	**DGT**	**ECB**
house 68.02% home 9.13%	house 52.36% home 29.67%	box 65.55%	mint 52.23%

we will view the respective micro-PTD in the source language (L1), where we can once again have access to its translation equivalents and their contexts, as can be seen in Figure 9.6.

It thus becomes clear that PTDs, by allowing for the alternation between L1 and L2, have a cyclical nature that can be automatically activated either through the micro- or mega-PTD by clicking on the four centrifugal arrows.

It is interesting to note in Table 9.4 that the results obtained with the micro-PTDs might diverge depending on the type of corpus under analysis.

The functional equivalents of *casa* ('house' / 'home') become more specific when our query involves more technical corpora linked to the financial and economic field: *Casa* might not correspond to 'home' or 'house' and might be translated, for example, as 'mint' (*Casa da Moeda*) in the European Central Bank (ECB) corpus or as 'box' in the European Commission Directorate-General for Translation (DGT) corpus. This clearly illustrates the importance of building corpora based on a diversified text typology, as can be seen from the fact that the level of the technicality of terms gradually increases in some types of corpora featuring specialized language.

5. Concluding remarks

The Per-Fide corpus sets itself apart from other corpora mainly due to the number of languages involved and the central role played by Portuguese. Furthermore, it will be made freely available to the research community for searching and downloading, along with the terminological and lexicographic material produced in the context of this project. As observed by Kraif (2006, p. 15), both alignment and bilingual concordance tools still remain largely underexplored. Indeed, some of their features, such as the automatic extraction of bilingual lexicons (PTDs) or the query process based on morphosyntactic annotation and lemmatization, are unknown to many students, linguists and translators. The Per-Fide Project has organized a series of workshops demonstrating the potential of these resources and tools in different research domains. Our mission has been to help different target groups work efficiently with corpus-based instruments and apply corpus querying methods in their research and professional activities.

Acknowledgements

The Per-Fide Project is supported in part by a grant (Reference No. PTDC / CLELLI / 108948 / 2008) from the Portuguese Foundation for Science and Technology, and it is co-funded by the European Regional Development Fund.

We would like to thank all contributing authors, translators, publishers, and institutions for their generosity in allowing us to include their texts in the Per-Fide corpus.

References

Almeida, J. J., Simões, A. and Castro, J. A. (2002), 'Grabbing parallel corpora from the web'. *Procesamiento del Lenguaje Natural*, 29, 13–20.

Alvarez, M. L. O. (2007), 'As expressões idiomáticas nas aulas de ELE: Um bicho de sete cabeças?', in I. González Rey (ed.), *Les expressions figées en didactique des langues étrangères*. Fernelmont: Proximités E.M.E, pp. 159–79.

Aston, G. (ed.) (2001), *Learning with corpora.* Houston: Athelstan.

Evert, S. and Hardie, A. (2011), 'Twenty-first century Corpus Workbench: Updating a query architecture for the New Millennium'. Paper presented at *Corpus Linguistics 2011*, University of Birmingham, UK.

Fielding, R. T. and Taylor, R. N. (2002), 'Principled design of the modern Web architecture'. *ACM Transactions on Internet Technology*, 2, (2), 115–50.

Frankenberg-Garcia, A. (2005), 'Pedagogical uses of monolingual and parallel concordances'. *ELT Journal*, 59, (3), 189–98.

Frankenberg-Garcia, A. and Santos, D. (2002), 'COMPARA, um corpus paralelo de português e de inglês na Web'. *Cadernos de Tradução*, 9, (1), 61–79.

Granger, S. (2003), 'The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research', *TESOL Quarterly,* 37, (3), 538–46.

—(2010), 'Comparable and translation corpora in cross-linguistic research. Design, analysis and applications'. *Journal of Shanghai Jiaotong University*, 2, 14–21.

Granger S., Lerot J. and Petch-Tyson S. (eds) (2007), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Beijing: Foreign Language Teaching and Research Press.

Gross, M. (1998), 'La fonction sémantique des verbes supports'. *Travaux de Linguistique*, 37, 25–46.

Grossmann, F. and Tutin, A. (2003), *Les collocations: Analyse et traitement*. Amsterdam: De Werelt.

Iriarte Sanromán, Á. (2001), *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*. Braga: University of Minho.

Kraif, O. (2006), 'Qu'attendre de l'alignement de corpus multilingues?', in *Revue Traduire, 4e Journée de la traduction professionnelle*, 210, 17–37.

Lin, J. and Dyer, C. (2010), *Data-Intensive Text Processing With MapReduce*. San Rafael, CA: Morgan & Claypool Publishers.

McIlwaine, I. (2000), *The Universal Decimal Classification: A Guide to its Use*. The Hague: UDC Consortium.

Manning, C. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mejri, S. (1997), *Le figement lexical. Descriptions linguistiques et structuration sémantique*. Tunis: Publications de la Faculté des lettres Manouba.

Padró, L. (2011), 'Analizadores Multilingües en FreeLing'. *Linguamática*, 3, (2), 13–20.

Santos, A. (2011), *Contributions for building a Corpora-Flow system*. Master's thesis, University of Minho.

—(2002), 'DISPARA, a system for distributing parallel corpora on the Web', in N. Mamede and E. Ranchhod (eds), *Advances in Natural Language Processing (PorTAL 2002)*, Berlin/Heidelberg: Springer-Verlag, pp. 209–18.

Santos, A. and Almeida, J. J. (2011), 'Text::Perfide::BookCleaner, a Perl module to clean plain text books', Paper presented at *27th Conference of the Spanish Society for Natural Language Processing (SEPLN 2011)*, University of Huelva, Spain.

Santos, A., Almeida, J. J. and Carvalho, N. (2012), 'Structural alignment of plain text books', in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'2012)*. CD-ROM

Simões, A. M. and Almeida, J. J. (2003), 'NATools – a statistical word aligner workbench'. *Procesamiento del Lenguaje Natural*, 31, 217–24.

Sinclair, J. (ed.) (2004), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Sperberg-McQueen, C. M. and Burnard, L. (eds) (2002), *Guidelines for Text Encoding and Interchange*. Oxford: University of Oxford, Humanities Computing Unit.

St. John, E. (2001), 'A case for using a parallel corpus and concordancer for beginners of a foreign language'. *Language Learning & Technology*, 5, (3), 185–203.

Tiedemann, J. (2003), 'Combining Clues for Word Alignment', in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 339–46.

—(2012), 'Parallel data, Tools and Interfaces in OPUS', in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. CD-ROM

Tiedemann, J. and Nygaard, L. (2004),'The OPUS corpus – parallel and free', in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*. CD-ROM

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. and Nagy, V. (2005), 'Parallel corpora for medium density languages', in *Proceedings of the RANLP 2005*, pp. 590–6.

Notes

1 Although it is beyond the scope of this article to provide an exhaustive list of the parallel corpora which include a Portuguese subcorpus, mention should be made of the following projects. (i) The Oslo Multilingual Corpus (OMC) consists of an English–Norwegian–Portuguese translation subcorpus containing 15 English original fiction texts and their translations into Norwegian and Portuguese. Due to copyright reasons, access to the OMC is only available to researchers and graduate students at the universities in Oslo and Bergen. For details of the OMC, see http://www.hf.uio.no/ilos/english/services/omc/. (ii) The Linguistic Corpus of the University of Vigo (CLUVI) includes four small parallel subcorpora of Portuguese as source or target language: TURIGAL, a 1.3-million-word corpus of Portuguese–English tourism texts; the 900,000 word corpus of English–Portuguese literary texts; PALOP, a 600,000

word corpus of Portuguese–Spanish postcolonial literature; and PEGA, a 70,000 word corpus of Portuguese–Galician literary texts. For further information on CLUVI, see http://sli.uvigo.es/CLUVI/index_en.html.

2 For further information on the activities being conducted and the resources made available by the Linguateca, see http://www.linguateca.pt/.

3 http://www.linguateca.pt/COMPARA/.

4 In order to realize the full potential of electronic corpora, most of today's linguists depend on the availability of specialized software tools. The IMS Corpus Workbench (CWB: http://cwb.sourceforge.net/) is a widely used architecture for corpus analysis, originally designed at the IMS, University of Stuttgart. The central component of the Corpus Workbench (cf. Evert and Hardie, 2011) is the corpus query processor CQP. Its query language allows sophisticated searches both for individual words and lexico-grammatical patterns.

5 http://opus.lingfil.uu.se/bin/opuscqp.pl.

6 For a detailed account of the attribute and value tagsets used for morphosyntactic annotation and lemmatization in COMPARA and OPUS, see the handout from the workshop Como pesquisar em corpora (How to query corpora) in the scope of the I International Per-Fide Conference on Corpora and Translation: http://per-fide.ilch.uminho.pt/site.pl/workshop.pt.

7 For a regularly updated list of the collaborators and the texts included in the Per-Fide corpora, see http://per-fide.ilch.uminho.pt/.

8 For a detailed account of the UNESCO Thesaurus, see the UNESCO website http://www2.ulcc.ac.uk/unesco/.

9 In computer science, a pipeline usually refers to a sequence of operations where data are handled from one operation to the next – in this specific context, operations over corpora.

10 A simple file to describe how compilation or other operations over files are executed.

11 http://perfide.ilch.uminho.pt/query.

12 Porto Editora is the leading educational publisher in Portugal, specializing in educational manuals, dictionaries, and multimedia products both online and offline. The lexicographic resources are available from the service Infopédia: (http://www.infopedia.pt/).