Introducing the *Per-Fide* Project: Parallelizing Portuguese with six different Languages

Sílvia Araújo, Ana Correia, Ana Oliveira, Alberto Simões
University of Minho

Abstract: In this paper we present the *Per-Fide* project[1], aimed at the construction of parallel corpora mapping the Portuguese language to six other languages - English, Russian, French, Italian, German and Spanish - in various domains including literary, journalistic and religious texts. First we will focus on the corpus design criteria and its main features, particularly those that distinguish this corpus from existing parallel corpora. Secondly, we will discuss the challenges of elaborating a typology of text-types for the religious domain and problems associated with the encoding of the texts belonging to this category. To conclude, we will demonstrate how the *Per-Fide* Corpus can be used in contrastive and translation studies with a case study of pronominal causative constructions in a French-Portuguese contrastive perspective.

Introductory note

The *Per-Fide* project is a joint venture between the Computer Science Department and the School of Humanities of the University of Minho (Braga, Portugal) intended to establish a research environment for translation and literary studies, lexicography, contrastive linguistics, and language teaching.

1. Project goals

We aim to compile a set of parallel corpora between the Portuguese language and six other languages. Although the term parallel is subject to some controversy, we have chosen to adopt the interpretation offered by McEnery & Xiao (2007), according to which "a [parallel] corpus contains source texts and their translations." We plan to include the same text in as many languages as possible, which accounts for the multi-directionality of the corpus. This was also a concern for the ACTRES parallel corpus[2], and in that regard Marlén Izquierdo *et al.* (2008: 35) argue that "given the translation nature of the corpus, being [multi]-directional would improve the quality and representativeness of the results obtained from such a tool when embarking on a translation project." We believe that the *Per-Fide* Corpus will be an added value to both Corpus Linguistics and translation studies based on a contrastive perspective not only because of the languages it represents but also because of the multiple directionality possibilities that it will offer.

It should be noticed, however, that the scope of this project goes beyond corpora compilation. That is merely considered to be the initial step. One of the main concerns of the members of the *Per-Fide* project is that all the resources collected and produced (not only the corpora but also translation dictionaries or bilingual terminology) be made publicly available for query and download.

1.1. Language range

The *Per-Fide* Corpus will establish a relation between Portuguese and six other languages, namely Spanish, Russian, French, Italian, German and English (Pt, Es, Ru, Fr, It, De, En: *Per-Fide*). The corpus will contain various language combinations in which the Portuguese language (in its different varieties: European Portuguese, Brazilian Portuguese and African Portuguese) is part of, either as source language or as target language. We are, nonetheless, aware of the problem of unknown translation direction, which tends to occur in multi-language environments such as the European Legislation.

One of the main features that distinguish the *Dutch Parallel Corpus* from other existing corpora is the “Dutch kernel” (Paulussen *et al*., 2007: 1). Similarly, Portuguese will be the pivotal language of the *Per-Fide* Corpus (Araújo *et al*., 2010).

1.2. Text typology

The developed corpora will contain original texts in the seven languages and their translations into as many of the other six languages as possible. Whenever possible, we will try to produce parallel corpora with more than two languages. Specifically, the corpora will consist of contemporary novels and short stories with a strong focus on the works of Portuguese authors; religious texts[3] (mainly Encyclicals, Letters and Angelus from the Vatican website); journalistic articles (*Le Monde Diplomatique*, *Le Courrier International*); judicial texts (European Community Law and international agreements) and technical texts (instructions and operating manuals, norms, standards and directives, technical texts and specialised documentation in the fields of automotive industry, electronics, telecommunications, computer science, standardisation, pharmaceutical industry and medicine/health sciences).

2. Describing the corpus compilation process

The first stage of the project concerns text selection, classification, digitisation and copyright clearance. Once the texts are in electronic format, each one will be enriched with a detailed descriptive header conforming to the TEI standard[4]. Metadata plays a key role in organising the ways in which a language corpus can be processed. It records the interpretive framework within which the components of a corpus were selected and are to be understood. Such information can also be very helpful to the user seeking to determine the potential relevance of the resource.

The corpus will then be tagged with morphosyntactic information. While the search for word concordances can be useful, there is a huge relevance on lemma or Part-of-Speech (POS) concordances. All corpora will be lemmatised and annotated with POS using language-specific morphological analysers[5]. The next stage is the sentence alignment process. We will take the source text as our starting point and align this text with its translations in the different languages at paragraph and sentence level. As it is known, technical texts have a relatively natural alignment. Literary text alignment is non-linear, leading to the need for a quality verification of the parallel text alignment process, which should not be completely manual. Alignment metrics will be calculated and used to detect suspicious alignments[6]. At present, we have succeeded in aligning the texts obtained from the Vatican website[7] as well as several world literature classics, and as a result, two experimental interfaces[8] have been created to access these corpora, which can be used for research purposes in linguistic studies. We have come across various obstacles that compromise the accuracy of the alignment: the disproportionate length of the bitexts and the amount of unnecessary meta-information in the retrieved documents. Therefore, as far as the alignment process is concerned it is necessary to accomplish a set of cleaning tasks, removing copyright data (while it is preserved in the meta-information TEI file), removing page numbers, headers, etc..

After the sentence alignment, we intend to process these parallel corpora to extract probabilistic translation dictionaries (Simões & Almeida, 2003), translation examples (Simões & Almeida, 2007) and bilingual terminology (Guinovart & Simões, 2009). These resources are undoubtedly a valuable contribution to lexicography and the translation process, whether it is manual or automatic.

The compiled corpora will be made available for download in TEI (Erjavec, 1999), XCES (Ide *et al.*, 2000) and TMX (Savourel, 2005) formats, and will be available for on-line querying, relying on a user-friendly search interface with a diversity of search and querying options for a wide range of research interests.

The related resources, such as terminology and dictionaries, will be available for download in suitable open formats, and their query will be integrated in the on-line interface, making it an integrated tool to consult all resources from the project.

3. The contribution of the *Per-Fide* Corpus to the Natural Language Processing of Portuguese

The expression of the Portuguese language in existing corpora is quite limited. One of the key aspects of the *Per-Fide* Corpus is the preponderant role assigned to Portuguese and its Brazilian and African varieties as either source or target language. Some monolingual projects in Portuguese have been developed by the Language Resource Center for Portuguese (*Linguateca*[9]) such as *CETEMPúblico* (Rocha & Santos, 2000) and *CETENFolha*, which are journalistic corpora for European and Brazilian Portuguese, respectively. Another project of journalistic corpora compilation has been developed by *Linguateca* in collaboration with the Portuguese and French editions of *Le Monde Diplomatique* (Correia, 2006). The texts provided, which correspond to the articles published between 1999 and 2002, have been aligned at sentence level and the result is available at http://linguateca.di.uminho.pt/nat/nat.pl under the description LMD-PT-FR. This collaboration project with *Le Monde Diplomatique* is now being resumed in order to integrate additional and more recent texts in French and Portuguese as well as to extend this compilation work to the remaining five languages of the project. *Linguateca* has other smaller corpora, including classic literature, oral and political corpora. The Center for Linguistics of the University of Lisbon built the *Reference Corpus of Contemporary Portuguese* (Nascimento, 2000). The *Corpus do Português*, a 45 million word corpus, which was created by Mark Davies (Brigham Young University) in collaboration with Michael Ferreira (Georgetown

University) includes texts from the XIVth to the XXth Century pertaining to European Portuguese and Brazilian Portuguese. While available for querying on the Internet, these last two corpora are not available for download. One of the main concerns of the *Per-Fide* Corpus is to make all the texts available for query and download, thus making the corpus available to the entire research community.

The neglect of the Portuguese language in multilingual parallel corpora is noticeable. However, since Portugal joined the European Union in 1986 the number of parallel corpora related to judicial domains has indeed undergone remarkable evolution. Well-established parallel corpora including the Portuguese language are *EuroParl* (Koehn, 2005) and *JRC-Acquis* (Steinberger *et al*., 2006). These corpora are compiled automatically for all European Languages on the basis of legislation for the European Union. One of the problems users face when accessing and searching these corpora is the fact that the file formats found online require specific software and alignment scripts in order to be consulted. In addition to this limitation, these corpora are noisy and have a low alignment quality. The quality standard of the *Per-Fide* Corpus will be guaranteed by automatic techniques created to ensure alignment quality metrics. Also, concordance query will be made on the basis of bitexts. Aligning a text in all seven languages of the project would necessarily compromise alignment quality, as the process of translation can remove, concatenate or add new sentences. Nonetheless, that would not be easy to achieve because in most cases we will not be able to obtain the same text in all the relevant languages.

As for the literary domain, *Linguateca* developed *COMPARA* (Garcia & Santos, 2003), which is a bi-directional Portuguese-English corpus. Portuguese also appears in the *OSLO-Multilingual Corpus* only as a target language in the language combination English-Norwegian-Portuguese. This trilingual sub-corpus comprises a small number of texts, 15 in total, in the form of extracts containing 10,000 to 15,000 words. Our goal is, precisely, to produce yet more parallel literary texts, adding a set of other languages to align Portuguese to.

Ultimately, however, there is still a pressing need for more corpora involving the Portuguese language as well as different domains of knowledge. In our project, we have included a variety of religious and technical texts. The typology of judicial texts included in the above-mentioned corpora can be extended to include other judicial text types, namely: international agreements, jurisprudence texts and other relevant documents. Some of these documents can be found and obtained automatically from the *EUR-Lex* website. One of our objectives is to extend the existing legislative parallel corpora on the basis of these texts available online in order to offer a more updated and larger database of judicial documents.

4. The *Per-Fide* Corpus and the Text Encoding Initiative

The Text Encoding Initiative (TEI) consortium has developed a set of guidelines for the representation of text in digital form, providing a methodology for the encoding of machine-readable texts. These standards have proved particularly useful in the fields of humanities, social sciences and linguistics. In compiling the *Per-Fide* Corpus, we intend to use the structure developed by the TEI to annotate meta-information.

4.1. An Independent header

The initial concept underlying the TEI was the reproduction of a series of descriptions and statements that could provide the electronic document (i.e. the machine-readable version of the original document) with a title page that can be compared to the title page of any printed text. The information described and stated in the TEI header refers to both its bibliographical and non-bibliographical aspects. The TEI header allows the annotation of the electronic text according to a particular model as if it were a metadata bank (Giordano, 1995).

In this project, all the texts integrated in the corpus are annotated with meta-information, and its description is organised according to the standards developed by the TEI. The meta-information is stated and described in an independent header, which is to be attached to the corresponding electronic document. Thus, the independent header contains all the relevant information that identifies the electronic document to which it is attached. It should be noticed, however, that our goal was never to apply the TEI standard to the body of text. The current logistical constraints prevented us from pursuing such endeavour. Thus, we focused on the elaboration of a header-document, which is attached to the corresponding electronic document, where it would be possible to store all the descriptive information and relevant statements that serve as an electronic title page, thus defining the structure of the document it refers to. In accessing this independent header, it becomes possible to create catalogues or indexes (e.g. index of 'sacred books') by domain, author, language, year, etc.. Furthermore, it helps to enhance the corpus document retrieval system.

To better suit the particular needs and goals of the *Per-Fide* project, we have created a customised version of the original TEI header proposal. The first section is the *file description*, which contains a bibliographical description of the electronic file and includes an item called *source description* for document provenance. The second section concerns the *profile description,* which contemplates more specific aspects of the language(s) used - <langUsage> - as well as an element specifically concerning text classification - <textClass>. Below you will find a segment illustrating the structure of the TEI header:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI.2>
  <teiHeader id=“321">                              <!-- Sequential ID value -->
    <fileDesc>
      <titleStmt>...</titleStmt>                    <!-- Document title information -->
      <extent>3540 words</extent>                   <!-- Document size -->
      <publicationStmt>...</publicationStmt>        <!-- Document copyright information -->
      <notesStmt>...</notesStmt>                    <!-- Miscellaneous notes -->
      <sourceDesc>...</sourceDesc>                  <!-- Document origin -->
    </fileDesc>
    <profileDesc>...</profileDesc>                  <!-- language, document classification, etc. -->
  </teiHeader>
</TEI.2>
```

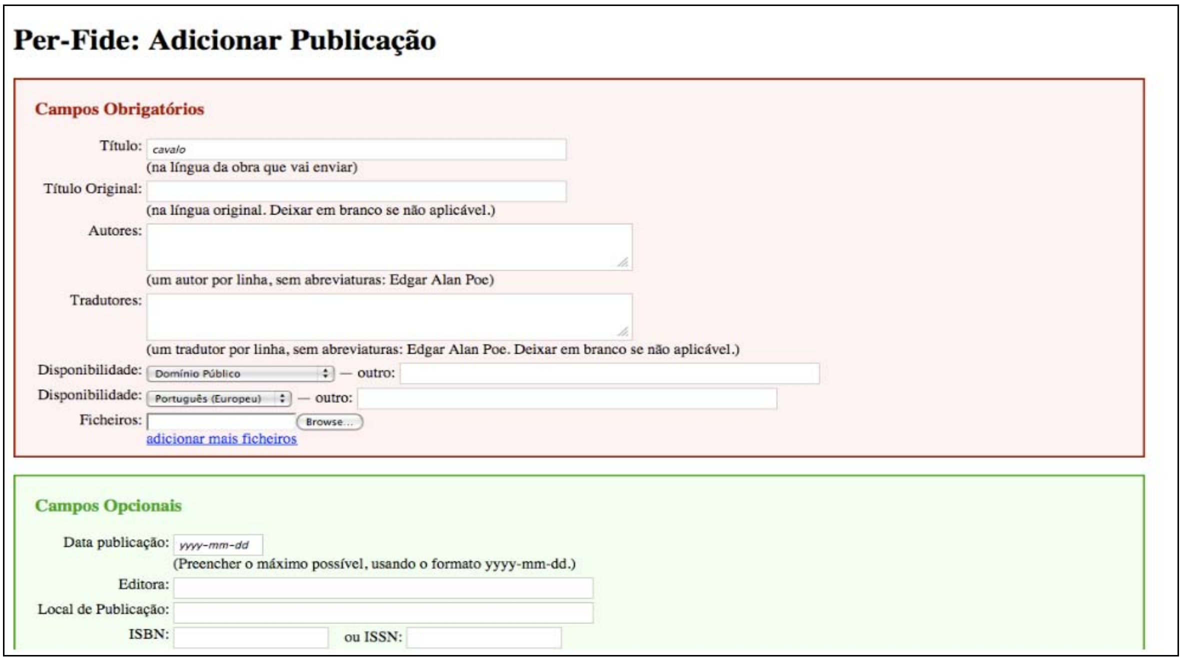
[1] TEI header segment

As we can see, the TEI header includes different kinds of information; not only the standard bibliographic data (e.g. document title, list of authors, publishing company, publication year, etc.) but also non-bibliographical information such as document copyright, languages and even a set of classificatory keywords (the next section will focus on this problem, explaining how the *Per-Fide* project is planning to proceed in terms of document classification). Indeed, the amount of information that can be annotated with recourse to the TEI header is noteworthy. Nevertheless, the TEI header was developed as a proposal, hence a model which is susceptible of adaptation: on one hand, for the document to be considered TEI-compliant, only some items of this panoply of possible elements are mandatory; on the other hand, it is also possible to expand the set of elements originally contemplated by the TEI proposal according to nature of each project. It is beyond the scope of this project to describe the full TEI header structure. Nonetheless, we invite the reader to visit the full specification that is available online.[10]

It should be noted that the TEI header of the *Per-Fide* Corpus includes elements that cannot be identified at every instance. As an example, in analysing the texts from the Vatican website, it is not possible to determine the source language of each text. In such cases, the source language will be explicitly annotated as ‘unknown’.

In some other circumstances other kind of data will be missing, as the publisher name (for a document in the public domain, for example). In these cases, instead of filling in the field with the ‘unknown’ string, that element will simply be ignored and excluded from the resulting XML file.

The texts for compilation will be gathered by all the members of the research team, some of which are not familiar with the TEI and, for that reason, unprepared to fill in the header using XML format as required. Consequently, in order to assist the process of creating the TEI header, a web interface was developed where the members of the project can fill in the relevant information concerning each text they submit for integration in the *Per-Fide* Corpus, subsequently uploading the respective file:



Per-Fide: Adicionar Publicação

Campos Obrigatórios

Título: cavalo
(na língua da obra que vai enviar)
Título Original:
(na língua original. Deixar em branco se não aplicável.)
Autores:
(um autor por linha, sem abreviaturas: Edgar Alan Poe)
Tradutores:
(um tradutor por linha, sem abreviaturas: Edgar Alan Poe. Deixar em branco se não aplicável.)
Disponibilidade: Domínio Público — outro:
Disponibilidade: Português (Europeu) — outro:
Ficheiros:
adicionar mais ficheiros

Campos Opcionais

Data publicação: yyyy-mm-dd
(Preencher o máximo possível, usando o formato yyyy-mm-dd.)
Editora:
Local de Publicação:
ISBN: ou ISSN:

[2] Web interface for uploading files

This interface is structurally identical to the TEI header, which means that the fields to be filled in correspond to the elements stated in the header. The interface then automatically generates the TEI header file, and stores both the document and its meta-information into a collaborative version control system.

5. Text classification: the problem of multiple-domain texts

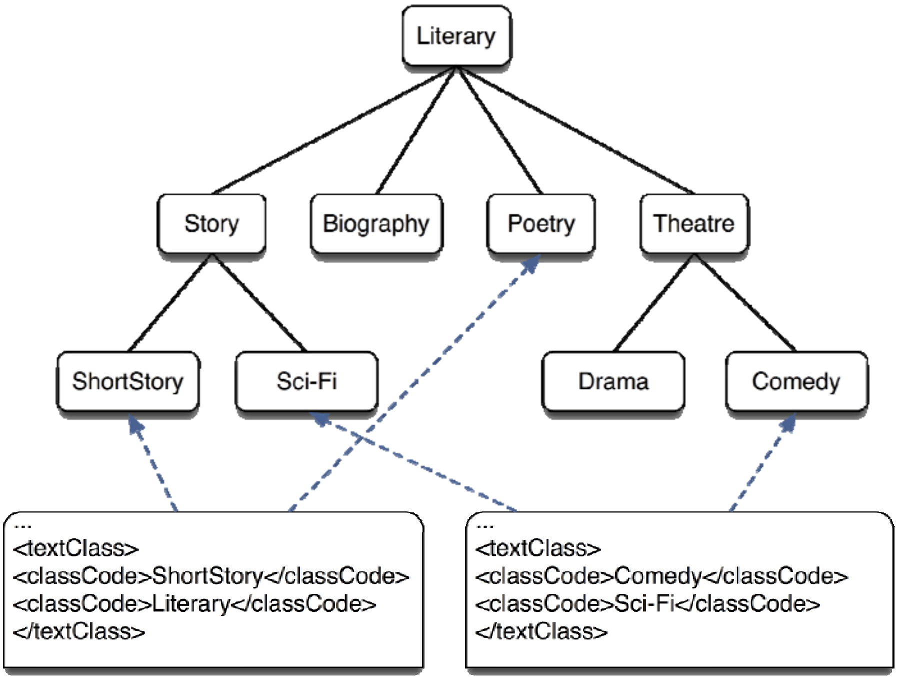
The classification of texts into different categories is not a simple task, and the terminology used to identify those categories may be open to question. However, the difference between the categories and text identification is dependent on the observation of certain characteristics. It is thus fair to assume that in obtaining the characteristics of a particular text, we will be left with a variety of elements which could raise questions about the belonging of certain texts to a particular category. This kind of text, which will be referred to as 'hybrid', can motivate different perspectives regarding their classification.

José Saramago[11]'s last work, *Caim*[12] (2009), clearly illustrates the need of classifying 'hybrid' texts. On the one hand, it is considered to be a novel, hence belonging to the literary domain[13]. On the other hand, however, on *Wikipedia* it is regarded as part of the *Books Critical of Religion* category. Therefore, we have decided to use a strategy that allows us to record this inter-relation of concepts in the various domains. This strategy should simultaneously result in the standardisation of the terminology to be used. In order to represent and structure the different domains, we have resorted to a methodology that provides the principles necessary to group concepts of a similar nature into wider categories where all terms are hierarchically inter-related and associated with one another. The use of an ontology seemed to be the most appropriate not only for the normalisation of terms, for concept representation and text indexation but also because it enables the user to find information within a particular domain in the process of document search and selection. The thesaurus offers a similar approach to that of the ontology. However, the difference between these two methodologies resides in the fact that the ontology is an extensible system that allows the relation between terms to be re-arranged and/or extended whenever necessary. As we are unable to predict the kind of texts that we will come across with during the compilation stage of the corpus, it is essential to rely on a classification system that can be rectified and ameliorated over time according to the type of texts submitted for integration in the corpus.

The hierarchical organisation of a list of words in an ontology can be defined in different ways. Nevertheless, the usual method to define class hierarchy is the relation between more generic terms – Broader Term (BT) – and more specific ones – Narrower Term (NT).

The lack of standardising ontologies for document classification has led us to initiate a comparative research project to create an outline based on international document classification schemes such as the *Universal Decimal Classification* (UDC) and virtual libraries (thesauruses) in the areas of Biblioteconomy and Information Science such as the *UNESCO Thesaurus*. In elaborating this ontology, we were able to hierarchically structure thematic domains that will be used to classify the texts that will integrate the *Per-Fide* Corpus.

As a final note, it is relevant to explain how the TEI header will play together with this ontology. As previously explained, the ontology is a classification system made up of a more or less complex network of hierarchical relations between terms. In as much as it provides a classificatory structure and nomenclature to help fill in the section for document classification in the TEI header, it is fair to assume that the ontology serves as a basis for the TEI header. As defined in the Thesaurus ISO 2788[14], the identification of each term of the ontology must be unique so that, in analysing it, it becomes easier to locate a specific term and consequently identify its superior context of classification. Taking into consideration the possibility of reorganising the ontological structure, by adding or removing hierarchical tiers, we have chosen to state in the TEI header only the most specific, hence lower, tier of the ontology. In the TEI header, this element is stated as <textClass>. Below is a part of the ontological structure, demonstrating how it co-relates with the TEI header:



```
...
<textClass>
<classCode>ShortStory</classCode>
<classCode>Literary</classCode>
</textClass>
```

```
...
<textClass>
<classCode>Comedy</classCode>
<classCode>Sci-Fi</classCode>
</textClass>
```

[3] The ontology and the TEI header

The advantage of using the ontology in conjunction with the TEI header for document classification resides in the fact that it enables us to extend the conceptual and hierarchical relations between terms at any given stage of the classification process without it affecting the structure of the TEI header.

6. The French construction *se faire* + *infinitif* and their functional equivalents in Portuguese

In the present chapter we wish to demonstrate the use of a parallel corpus in translation studies based on the case study of the pronominal causative constructions in French *(se faire* + *inf)* and in Portuguese *(fazer-se* + *inf)*. As previously mentioned[15], a FR-PT parallel corpus composed of literary texts aligned at sentence level was compiled[16]. This corpus-based case study will enable us to simultaneously detect the similarities between *se faire* and *fazer-se* as well as their functional differences, which account for a reduced degree of equivalence in translations (Araújo, 2008).

6.1. Degree of agentivity of the syntactic subject in the *se faire* construction

The structure *se faire + inf* has been the object of diverging analyses. Some authors (Labelle, 2002 ; Riegel et *al.,* 1994) consider it a form of passive. Others (Tasmowski & Van Oevelen, 1987) advocate a unitary perspective, defending that the structure *se faire + inf* remains essentially causative despite its similarity with the passive construction. There are yet other authors (Kupferman, 1995) who refuse the unitary perspective in favour of a binary analysis according to which there are two pronominal causative constructions - one is causative and the other is passive.

Most authors who have studied this pronominal causative construction put forward a common semantic argument - the subject of *se faire + inf* is to some extent responsible for the process marked by the infinitive verb. In (1a) *l'enfant s'est fait renverser par une voiture*, the verb *faire* necessarily involves the syntactic subject in a more active way rather than would be the case in a canonical passive sentence such as (1b) *l'enfant a été renversé par une voiture*. Despite the strong syntactic affinity, these examples differ in terms of semantics: in (1a) *l'enfant* is an active patient, while in (1b) *l'enfant* is a passive patient (cf. on this matter Gaatone, 1983; Novakova, 2008).

6.2. A contrastive analysis of Portuguese translational equivalents of *se faire + infinitif*

Se faire + inf entails a double semantic role of the syntactic subject, who is simultaneously patient and instigator (voluntary or involuntary) of the process implied by the infinitive verb. Subsequently, we will systematise the functional uses of *se faire*, and analyse the respective translational equivalents in Portuguese.

6.2.1. Voluntary instigator role

The action marked by the infinitive is intentionally originated by the subject, who becomes the object of an action or state that was provoked by himself/herself:

> (2a) Depuis trois mois, il l'enveloppait dans l'irrésistible filet de sa tendresse. Il la séduisait, la captivait, la conquérait. Il *s'était fait aimer* par elle, comme il savait *se faire aimer*. Il avait cueilli sans peine son âme légère de poupée. [*Bel Ami,* Guy de Maupassant]

In Portuguese, it is possible to resort to literal translation in these cases as in (2b), which was translated from the example above:

> (2b) Havia três meses que a envolvia na irresistível rede da sua ternura. Seduzia-a, cativava-a, conquistava-a. *Tinha-se feito amar* por ela como só ele sabia *fazer-se amar*. Apossara-se sem dificuldade da sua alma ingénua de boneca. [*Bel Ami,* translation by Cesar Oldemiro]

It would seem that intentionality is the *sine qua non* for the use of *fazer-se* but the fact remains that it is not valid in every situation that conjures up the notion of intentionality. Among other translation strategies, translators often resort to the canonical passive structure rather than *fazer-se* in order to convey the meaning of *se faire* with teleonomic value. In (3a) the Portuguese translator opted for the passive voice to translate *se faire* because, even in a teleonomic context, the subject also possesses a non-agentive role. However, in associating the passive structure with the volitive verb (*querer*), the translator also manages to preserve the notion of intentionality:

> (3a) Swann *se fit conduire* dans les derniers restaurants; c'est la seule hypothèse du bonheur qu'il avait envisagée avec calme; [*Un amour de Swann,* Marcel Proust]
>
> (3b) Swann quis *ser conduzido* aos últimos restaurantes; era a única hipótese de felicidade que encarara com calma; [*Amor de Swann,* translation by Miguel Serras Pereira].

In a similar context, another translator chose to resort to literal translation in order to express that same teleonomic meaning:

(4a) Arrivés près de Bologne, nos amis *se firent conduire* à travers champs sur la route qui de Florence conduit à Bologne;» [*Chartreuse de Parme*, Henri Beyle Stendhal]

(4b) Chegados perto de Bolonha, os nossos amigos *fizeram-se conduzir*, através dos campos, até à estrada que vai de Florença àquela cidade; [*Cartuxa de Parma*, translation by Adolfo Casais Monteiro]

Indeed when the subject acts as voluntary instigator, literal translation is justified. However, when that is not the case, as we will see in the next items, in order to deal with the limitations of the Portuguese language, translators must find solutions that comply with the following criteria (Lejeune & Araújo, 2003):

a) to preserve the thematic continuity of the text, keeping the syntactic subject in that same position;

b) to recover the role of the beneficiary marked in the source text by the pronoun *se*. This could be done explicitly or implicitly by means of co-textual elements;

c) to preserve the subjective nature underlying the process (subject intentionality in the case of teleonomy; subject responsibility in the case of antiteleonomy).

The translation of (5a) fulfills all three requisites. We would like to draw the reader's attention to the beneficiary role, which has been recovered in the translation by introducing the dative pronoun *lhe* in the completive sentence of (5b):

(5a) Asie *se faisait expliquer* le Palais qu'elle connaissait mieux que l'avocat ne le connaissait lui-même; [*Splendeurs et misères des courtisanes,* Honoré de Balzac]

(5b) Ásia *fazia com que* o advogado lhe *explicasse* o Palácio, que ela conhecia melhor do que o próprio advogado. [*Esplendores e misérias das cortesãs*, translation by Américo de Carvalho]

6.2.2. Involuntary instigator role

By means of his/her own behaviour (negligence, recklessness, provocation, etc.), the subject has non-intentionally brought upon him/herself a situation that is often detrimental. In the literature, this particular use of *se faire* supposes responsibility on the part of the subject:

> (6a) L'après-midi, justement, elle avait lu dans *Le Figaro* le compte rendu d'une séance de réunion publique, poussée au comique, dont elle riait encore, à cause des mots d'argot et de la sale tête d'un pochard qui *s'était fait expulser*. [*Nana*, Emile Zola]

The Portuguese *fazer-se* is intrinsically linked to the notion of intentionality. If we consider that it is unusual to admit that a human being is the voluntary instigator of an action that will negatively affect him/herself (Gaatone, 1983: 170, translation by the authors), it is understandable that the intentional meaning of *fazer-se* doesn't associate as spontaneously with verbs that mark violent or disagreeable actions[17] (Spang-Hanssen 1967: 140, translation by the authors). Thus, the translator naturally resorts to the passive construction as an alternative to *fazer-se*:

> (6b) Precisamente, nessa tarde, lera no *Fígaro* a reportagem de um comício feita num tom irónico e ria-se ainda, por causa das palavras ditas em calão e pelo comportamento ignóbil de um bêbado que acabara por *ser expulso***.** [*Nana*, translation by Carlos Loures]

When the indirect object is syntactically promoted to the role of the subject, it is not possible to resort to literal translation. In effect, the direct translation of *se faire* with *fazer-se* becomes an impossibility when *se faire* co-occurs with ditransitive verbs[18] to form a *passif complémentaire* (Bat-Zeev Shyldkrot, 1999: 67) or of the *destinataire* (J. François, 2000), allowing the indirect object to occupy the syntactic role of the subject. In Portuguese, *fazer-se* does not possess such syntactic flexibility, and for that reason, it is necessary to resort to a number of alternative translation strategies[19] such as converse verbs (*buy-sell*) in order to render these trivalent constructions that mark the transfer of one object from one situation to another.

With recourse to these verbs, it is possible to translate the thematisation of the indirect object induced by *se faire*:

(7a) Frédéric et Deslauriers marchaient au milieu de la foule pas à pas, quand un spectacle les arrêta: Martinon *se faisait rendre* de la monnaie au dépôt des parapluies; [*L'Education sentimentale,* Gustave Flaubert]

(7b) Frederico e Deslauriers caminhavam no meio da multidão quando um espectáculo os fez parar: Martinon *recebia* os trocos no depósito de guarda-chuvas; [*Educação sentimental*, translation by Alice Direito da Silva Santos]

(8a) Le lendemain matin Fabrice est parti pour la France, après *s'être fait donner* le passeport d'un de ses amis du peuple, un marchand de baromètres nommé Vasi. [*La Chartreuse de Parme,* Henri Beyle Stendhal]

(8b) No dia seguinte pela manhã Fabrício partiu para França, depois de *ter obtido* o passaporte dum dos seus amigos do povo, um negociante de barómetros chamado Vasi. [*A Cartuxa de Parma*, translation by Adolfo Casais Monteiro]

To the *se faire* construction may also correspond more specific verbs such as *pedir* (*ask*),*obrigar* (make) or *ordenar* (*tell*) that always express, to different extents, a manipulation of the subject of the infinitive verb by the subject of *se faire*, which can range from a mere request to plain coercion, with all the intermediary nuances (Khalifa 2004: 66, translated by the authors):

(9a) Enfin, après avoir donné l'ordre de fermer sa porte, il *se fit servir* son déjeuner dans le pavillon qui se trouvait à l'un des angles de son jardin. [*Splendeurs et misères des courtisanes;* Honoré de balzac]

(9b) Olhava constantemente para o jardim. Enfim depois de mandar fechar a porta, *pediu que* <u>lhe</u> *servissem* o almoço no pavilhão que se encontrava num dos ângulos do jardim. [*Esplendores e misérias das cortesãs*, traduction de Américo de Carvalho]

(10a) Le lendemain, vers une heure, il *se fit donner* le dernier coup de brosse par Mousqueton, et s'achemina vers la rue aux Ours, du pas d'un homme qui est en double bonne fortune. [*Les Trois Mousquetaires,* Alexandre Dumas]

(10b) No dia seguinte, por volta de uma hora, *ordenou* a Mousqueton *que* <u>lhe</u> *desse* a última escovadela e dirigiu-se para a rua Aux ours, caminhando como homem duplamente feliz. [*Os três mosqueteiros*, translation by Delfim de Brito]

(11a) Le notaire donna à la bonne des indications détaillées qu'elle *se fit répéter* plusieurs fois; [*Une vie,* Guy de Maupassant]

(11b) O notário prestou à criada indicações pormenorizadas que ela o *obrigou a repetir* várias vezes; [*Uma vida*, translation by Carlos Loures]

These verbs that express an *action sur autrui* (Le Goffic 1993: 260) usually accept a completive clause linked to the subjunctive (cf supra (9a) and (9b)). In adding the dative pronoun *lhe* to the completive clause in the examples (9b) and (10b), it becomes clear that the translator had every intention of preserving a mark of the beneficiary marked in (9a) and (10a) by *se*. The accuracy of the translation is necessarily compromised when the beneficiary role is not recoverable from the context. In fact, in (12b) and (13b), the translator chose to eliminate any reference to the beneficiary:

(12a) Emma *se fit servir* à dîner dans sa chambre, au coin du feu, sur un plateau. [*Madame Bovary,* Gustave Flaubert]

(12b) Emma *mandou servir* o jantar no quarto, no canto do fogão, sobre um tabuleiro. [*Madame Bovary*, translation by Fernanda Ferreira Graça]

(13a) Le lendemain, quand il fut debout (vers deux heures environ, il avait dormi tard), Rodolphe *se fit cueillir* une corbeille d'abricots. [*Madame Bovary,* Gustave Flaubert]

(13b) No dia seguinte, quando se levantou (cerca das duas horas, porque se deitara tarde), *mandou apanhar* um cesto de damascos. [*Madame Bovary*, translation by Fernanda Ferreira Graça]

In (14b) and (14c), the result of the translation fails to convey the fatalist dimension conferred by the use of the *se faire* construction in the source text:

(14a) Se jeter à genoux pour demander la grâce de Julien, devant la voiture du roi allant au galop, attirer l'attention du prince, au risque de *se faire* mille fois *écraser*, était une des moindres chimères que rêvait cette imagination exaltée et courageuse. [*Le Rouge et le Noir,* Stendhal]

(14b) Lançar-se de joelhos para pedir o perdão de Julien, perante a carruagem do rei indo a galope, chamar a atenção do príncipe com o perigo de *ser atropelada* mil vezes, era uma das menores quimeras com que sonhava essa imaginação exaltada e corajosa. [*O Vermelho e o Negro*, translation by Maria Eveline Monteiro]

> (14c) Deitar-se de joelhos para pedir o indulto de Julião diante do carro do rei a galope, chamar a atenção do príncipe arriscando-se mil vezes a *ser esmagada*, era uma das menores quimeras que sonhava aquela imaginação exaltada e corajosa. [*O Vermelho e o Negro*, translation by Maria Manuel e Branquinho da Fonseca]

6.2.3. *Se faire* with inanimate subjects

The *se faire* construction can also be used with inanimate subjects that are generally considered to be the origin of an effect or phenomenon whose manifestation causes perception (*se faire entendre, écouter, voir, sentir*, etc).

> (15a) La gelée avait si bien purifié l'air, durci la terre et saisi les pavés, que tout avait cette sonorité sèche dont les phénomènes nous surprennent toujours. La lourde démarche d'un buveur attardé, ou le bruit d'un fiacre retournant à Paris retentissaient plus vivement et *se faisaient écouter* plus loin que de coutume. [*La femme de trente ans,* Honoré de Balzac]

In the following example, however, it is the non-manifestation of an expected phenomenon that can be verified:

> (16) L'agriculture risque, elle, d'y retrouver un second souffle et les paysans une prospérité nouvelle, relançant ainsi un décollage économique qui *s'est fait trop attendre* jusqu'à maintenant. [*Le Monde Diplomatique*, Juin 1982]

The temporal manifestation of the phenomenon or event is secondary as it has already been the object of a previous construction generally marked by a definite article, which introduces the subject (*la gelée* in (15a)) of that previous construction. The phenomenon that ‘makes itself felt’ has a prior existence that makes it, or its consequences, potentially ‘sensitive’ to others (the sounds of the steps and the wagon can be heard quite distinctively due to the ice; see (15a)), which is why it is reasonable to assume there is a teleonomic context, albeit in a broader sense.

The translation of example (15a) demonstrates that the marker *fazer-se* provides the most probable option for translation in Portuguese when *se faire* is associated with inanimate subjects:

(15b) A geada tinha purificado extraordinariamente o ar, endurecido a terra e invadido as lajes de tal modo que tudo possuía aquela sonoridade seca cujos fenómenos nos surpreendem constantemente. O passo pesado de um bêbado atrasado ou o ruído de um fiacre regressando a Paris ressoavam fortemente e *faziam-se ouvir* até mais longe que de costume. [*A mulher de trinta anos*, traduction de Dóris Graça Dias]

However, a comparison between French originals and the Portuguese translation will show that the passive construction built with *ser PP* in (15c) or the pronominal passive in (17b) are widely used as alternatives for translating *se faire*, notwithstanding the fact it would be syntactically correct to use *fazer-se*:

(15c) […] O passo pesado de um bêbado retardatário ou o ruído de uma carruagem de regresso a Paris ecoavam mais vivamente e *eram ouvidos* mais longe do que de costume. [*A mulher de trinta anos*, traduction de Carlos Loures]

(17a) De nouvelles allées et venues *se firent entendre* [*Le Comte de Monte Cristo*, Alexandre Dumas]

(17b) *Ouviram-se* novas idas e vindas [*O Conde de Monte Cristo*, traduction de Adelino dos Santos Rodrigues]

In the translated segment (17b), the inversion of the canonical subject (*novas idas e vindas*) – verb (*ouviram-se*) order develops into an effect of focalisation on the verb that was chosen as the departure point – theme of the predicative relation. In the source text the subject (*de nouvelles allées et venues*) funtions as the departure point of the predicative relation.

To sum up, the French *se faire* + *inf* is indeed more grammaticalised than the Portuguese *fazer-se* + *inf*. While the French construction has developed both agentive and passive meanings, in Portuguese the use of *fazer-se* with non-agentive meaning is rare or inexistent as can be seen from the synoptic table below, which visually describes the areas of semantic intersection between *se faire* and *fazer-se*:

	FRENCH	PORTUGUESE
	DIRECT OBJECT	
A) Voluntary Instigator	*SE FAIRE* + INF (*représenter*, *comprendre*, etc.)	*FAZER-SE* + INF (*representar*, *entender*, etc.)
B)Involuntary Instigator	*SE FAIRE* + INF (*violer*, *renverser*, etc.)	**FAZER-SE* + INF (*violar*, *atropelar*, etc.)
C) Inanimate Subject	*SE FAIRE* + INF (*écouter*, *sentir*, etc.)	*FAZER-SE* + INF (*ouvir*, *sentir*, etc.)
INDIRECT OBJECT		
D) Voluntary Instigator	*SE FAIRE* + INF (*livrer* /*teindre* OD, etc.)	**FAZER-SE* + INF (*entregar* /*pintar* OD, etc.)
E) Involuntary instigator	*SE FAIRE* + INF (*voler/confisquer* OD, etc.)	**FAZER-SE* + INF (*roubar/confiscar* OD, etc.)

[4] Synoptic table: *se faire* / *fazer-se*

The grey colour highlights the cases that allow for literal translation of *se faire* with *fazer-se*. The remaining cases demonstrate the constraints of the Portuguese language, which force translators to find alternative solutions. These solutions can only be accurately systematised with recourse to corpus analysis.

7. Concluding remarks

In this paper, we presented the corpus compilation project, *Per-Fide*, which is now taking its first steps. There are still some elementary issues that need to be addressed such as the acquisition of copyright clearance, particularly for the literary texts, and the codification system to be used for the Russian language, which is different than the one used for western European languages.

The *Per-Fide* Corpus sets itself apart from other existing corpora mainly due to the number of languages involved and the strong focus on the Portuguese language. Furthermore, it will be made freely available for query and download to the research community as well as the terminological and lexicographic material produced in the scope of this project.

From the case study described in section 6, it becomes quite clear that the contribution of corpus linguistics to analyse, systematise and explain the choices intuitively made by translators is quite remarkable. The results of this kind of corpus-based contrastive study are unquestionably solid foundations for the creation of resources and methods for translation teaching and the training of translators.

There is an increasing research interest in corpora and their applications, and the potential for development in this area is immense. The output achieved during the three-year lifespan of the *Per-Fide* project could ramify into multiple lines of research. As an example, adding a comparable feature to this corpus would certainly improve its usability, bridging the gap between contrastive and translation studies.

Acknowledgements

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

Notes

[1] This paper aims at introducing the *Per-Fide* Corpus in terms of its main features and applications. More information on the development of the project as well as its team members and collaborators is available at: http://per-fide.ilch.uminho.pt

[2] The ACTRES parallel corpus is an English-Spanish translation corpus built at the Department of Modern Languages at the University of León (Spain) in collaboration with the Department of Culture, Language and Information Technology at the University of Bergen (Norway). For further information: http://actres.unileon.es

[3] The classification of religious texts is far from consensual. We will address this matter in item 5 of this paper.

[4] TEI is an XML standard for document structure and meta-information. For further information consult http://www.tei-c.org

[5] For the purpose of morphosyntactic annotation, we intend to use the parser *Palavras* (Bick, 2000).

[6] There is a variety of tools available for sentence-level alignment. Some of the most recent are *EasyAlign* which is part of the *IMS Corpus Workbench* (Christ *et al.*, 1999), *HunAlign* (Varga *et al*., 2005) or the *Clue Aligner in PLUG* (Tiedemann, 2003).

[7] For further information, please consult http://www.vatican.va

[8] The religious interface is available at: http://natura.di.uminho.pt/jjbin/perfide1; The literary interface is available at: http://natura.di.uminho.pt/jjbin/perfide2

[9] For further information: http://www.linguateca.pt

[10] The full TEI specification is available at http://www.tei-c.org/Guidelines/P4/html/index.html, and the TEI header description available at http://www.tei-c.org/Guidelines/P4/html/HD.html

[11] José de Sousa Saramago (1922-2010). Portuguese Nobel Prize in Literature, 1998.

[12] In Portuguese newspaper, *O Público*: … According to the Old Testament f the Bible, Cain was the first born son of Adam and Eve, who killed his younger brother, Abel, in a jealous fit after realising that God favoured him. (…) The author stressed the fact that he is by no means a writer of religious books, it is a subject that interests him merely because it has since ancient times been in the minds of men and in the history of mankind. … (in http://www.publico.pt/Cultura, translated by the authors).

[13] Saramago's wife, Pilar del Rio, claims that this last novel by Saramago (…) is literature in its purest state. (in http://www.josesramago.org, translated by the authors)

[14] For further information on this standard, visit the International Standard Organisation (ISO) at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7776

[15] Using French as a source language in this case study can be explained by its significantly recurrent and varied use of the constructions with *se faire*. Thus, it is interesting to observe the translation process from French into Portuguese.

[16] See endnote 8 on the literary corpus interface.

[17] It should be noted that this association between *fazer-se* and detrimental actions is possible whenever there are co-textual elements that point to intentionality on the part of the subject as in the following examples: *(a) Norbert saurait se faire tuer comme ses aïeux, c'est aussi le mérite d'un conscrit... Le marquis tombadans une rêverie profonde: Et encore se faire tuer, dit-il avec un soupir, peut-être ce Sorel le saurait-il aussi bien que lui...* [*Le Rouge et le Noir*, Stendhal]*; (b) Norbert saberia fazer-se matar como os seus antepassados, é também o mérito de um conscrito... O marquês caiu num profundo devaneio: E ainda fazer-se matar, dizia para consigo, talvez este Sorel o saiba tão bem como ele...* [*O Vermelho e o Negro*, translation by Maria Eveline Monteiro].

[18] The ditransitive verbs used with this construction are verbs of transfer (*se faire retirer, transmettre, livrer, prêter, remettre, rembourser, restituer*), donation (*se faire attribuer, distribuer, offrir, donner, offrir, servir/ravir*), and communication (*se faire conseiller, dédicacer, annoncer, dicter, notifier, raconter, ordonner*).

[19.] Some of the alternative translation strategies are: *Fazer (com) que + Subjunctive; fazer + Infinitive, mandar + Infinitive* (with recovery of the IO in the completive sentence or without mention to the IO, which can sometimes be recovered from the context)*;* passage of the instigator to agent*;* canonical passive with *ser* + PP; diathesis inversion (transformation into active); symmetrical construction.

References

Araújo, S., A. Simões, J. J. Almeida and I. Dias (2010). "Apresentação do projecto *Per-fide*: paralelizando o Português com seis outras línguas". *Linguamática*, 2:2, 71-74.

Araújo, S. (2008). *Entre l'actif et le passif: se faire/fazer-se. Syntaxe, sémantique et pragmatique comparées français-portugais*. PhD dissertation, University of Minho, Braga, Portugal.

Bat-Zeev Shyldkrot, H. (1999). "Analyse sémantique d'une forme passive complémentaire: *se laisser*". *Langages*, 135, 63-74.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Christ, O., B. M. Schulze, A. Hofmann and E. König, (1999). *The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual.* Institute for Natural Language Processing. University of Stuttgart.
Available at:
http://www.ims.uni_stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.

Correia, A. (2006). "Colaboração na constituição do corpus paralelo *Le Monde Diplomatique* (FR-PT)". Internship report, University of Minho, Braga, Portugal.

Erjavec, T. (1999). "A TEI encoding of aligned corpora as translation memories". In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora '99*, 49–60.

François, J. (2000). "Désémantisation verbale et grammaticalisation: *(se) voir* employé comme outil de redistribution des actants". *Syntaxe & Sémantique*, 2, 159-175.

Gaatone, D. (1983). "Le désagréable dans la syntaxe". *Revue romane*, 18:2, 161-174.

Garcia, A. F. and D. Santos (2003). "Introducing COMPARA, the Portuguese-English parallel translation corpus". In F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translation Education*. Manchester: St. Jerome Publishing, 71-87.

Giordano, R. (1995). "The TEI header and the documentation of electronic texts". *Computers and the Humanities*, 29:1, 75-84.

Labelle, M (2002). "The French non canonical passive in se faire". In Haraguchi, Shosuke, Bohumil Palek and Osamu Fujimura (eds.) *Proceedings of Linguistics and Phonetics*. Tokyo: Charles University Press and Meikai University.

Le Goffic, P. (1993). *Grammaire de la phrase française*, Paris: Hachette.

Guinovart, X. G. and A. Simões (2009). “Parallel corpus-based bilingual terminology extraction”. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence.*
Available at: http://alfarrabio.di.uminho.pt/~albie/publications/tia09.pdf

Ide, N., P. Bonhomme and L. Romary (2000). “XCES: an XML-based encoding standard for linguistic corpora”. In *Proceedings of the Second International Language Resources and Evaluation Conference.* Paris: European Language Resources Association.

Izquierdo, M., K. Hofland and Ø. Reigem (2008). “The ACTRES parallel corpus: an English-Spanish translation corpus”. *Corpora*, 3:1, 31-41.

Khalifa, J-C. (2004). *Syntaxe de l’anglais. Théories et pratique de l’énoncé complexe*. Paris: Ophrys.

Koehn, P. (2005). “EuroParl: A Parallel Corpus for Statistical Machine Translation”. In *Proceedings of MT-Summit*, 79-86.
Available at: http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf

Kupferman, L. (1995). “La construction passive en *se faire*”. *Journal of French Language Studies*, 5, 57-83.

Lejeune, P. and S. Araújo (2003) “Os equivalentes funcionais em português das construções francesas *se faire + INF e* se voir + INF/*PP*”. In *Actas do XIX Encontro da Associação Portuguesa de Linguística*, 213-226.

McEnery, A. M. and R. Z. Xiao (2007). “Parallel and comparable corpora: What is happening?” In G. Anderman and M. Rogers (eds) *Incorporating Corpora. The Linguist and the Translator*. Clevedon: Multilingual Matters, 18-31.

Nascimento, M. F. B. (2000). “O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito”. In E. Gärtner and C. Hundt (orgs.) *Estudos de Gramática Portuguesa I.* Frankfurt am Main: TFM, 185-200.

Novakova, I. (2008). “La construction se faire+Vinf: analyse fonctionnelle”. Colloque RSL, *Représentations du sens linguistique*, IV, 28-30 mai 2008, Helsinki, Finlande.

Paulussen, H., L. Macken, J. Trushkina, P. Desmet and W. Vandeweghe (2006). “Dutch Parallel Corpus: a multifunctional and multilingual corpus”. *Cahiers de l'Institut de Linguistique de Louvain*, 32: 1-4, 269—285.

Rocha, P. A. and D. Santos (2000). “CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa”. In M. G. Nunes (ed.) *V Encontro para o processamento computacional da língua portuguesa escrita e falada* (PROPOR 2000). São Paulo: ICMC/USP, 43-52.

Riegel, M., J-C. Pellat and R. Rioul (1994). *Grammaire méthodique du français*, Paris, PUF.

Savourel, Y. (2005). “TMX 1.4b Specification”. *Localisation Industry Standards Association* (LISA).
Available at: http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm

Simões, A. and J. J. Almeida (2003). “NATools – A Statistical Word Aligner Workbench”. *Procesamiento del Lenguaje Natural*, 31, 217-224.

Simões, A. and J. J. Almeida (2007). “Parallel Corpora based Translation Resources Extraction”. *Procesamiento del Lenguaje Natural*, 39, 265-272.

Spang-Hanssen, E. (1967). “Quelques périphrases passives du français moderne”. *Revue Romane*, 1, *Actes du 14 e Congrès des Romanistes Scandinaves dédiés à Holger Stan.* Copenhague: Akademisk Forlang, 139-147.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga (2006). “The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages”. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (*LREC'2006*). Genoa, Italy.
Available at: http://arxiv.org/ftp/cs/papers/0609/0609058.pdf

Tasmowski-De Ryck, L. and H. Van Oevelen (1987). “Le causatif pronominal”. *Revue romane,* 22:1, 40-58.

Tiedemann, J. (2003). “Combining clues for word alignment”. In *Proceedings of the 10th Conference of the European Chapter of the ACL* (EACL03). Budapest, Hungary.
Available at: http://urd.let.rug.nl/~tiedeman/Uplug/

Unesco (1983) *Thesaurus de l'Unesco: liste structurée de descripteurs pour l'indexation et la recherche bibliographiques dans les domaines de l'éducation, de la science, des sciences sociales, de la culture et de la communication.* Paris: Unesco.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy (2005). “Parallel corpora for medium density languages”. In Proceedings of *Recent Advances in Natural Language Processing* (RANLP 2005), 590–596.