Chapter 3
Corpus-Based Contrastive Study of Verbal Tenses

3.1 Dealing with Corpus Data

The corpus linguistics field has flourished over the last 50 years, mainly due to linguists' growing interest in having objective, quantifiable and reproducible data, and in using computers, which has without a doubt facilitated the use of large, and very large, corpora. For example, Kolaiti and Wilson (2014) carried out a corpus-based investigation of the unitary account from Relevance Theory, and lexical pragmatics in particular, in which narrowing, approximation and metaphorical extension are explained within the same model. They argue that:

> Corpus-based evidence provides a valuable complement to more traditional methods of investigation, by helping to sharpen intuitions, develop and test hypotheses and reduce the possibility of intuitive data being mere artefacts of the linguist. (Kolaiti and Wilson 2014, 211)

They point to the fact that corpus studies are a valuable source of inspiration for theorists in Relevance Theory, who are mainly concerned with the mental processes that enable the hearer to infer the speaker's meaning. This is primarily due to the fact that corpus work has forced 'us to consider examples that we might not have come up with ourselves, helping to sharpen and test our hypotheses, and raised new intriguing questions' (Kolaiti and Wilson 2014, 212).

Defining a corpus can be an easy and a difficult task at the same time, because of the numerous factors to consider, such as the type of text, the size, the purpose of creation, the way in which it can be analysed, etc. The well-known description of a corpus as "a body of naturally occurring language" (McEnery et al. 2006, 4) is largely accepted by the corpus linguistics community, as well as other domains that use corpora, such as empirical pragmatics, Natural Language Processing, Machine Translation and Translation Studies (Baker 1993, 1995).

The main features of corpora are that they have *finite size* (which may change over time, but which, in general, is pre-established such that construction criteria

C. Grisot, *Cohesion, Coherence and Temporal Reference from an Experimental Corpus Pragmatics Perspective*, Yearbook of Corpus Linguistics
and Pragmatics, https://doi.org/10.1007/978-3-319-96752-3_3

like balancing can be applied), constitute a *representative sample* of the variety or varieties of the language analysed, and represent the *standard reference*. Corpora have been compiled for many different purposes, and as such have different kinds of design, and include texts of different natures. Another definition of a corpus is that it constitutes an empirical basis by which to identify the elements and patterns of the structure of a language in order to analyse variation, for example, or that it can be analysed distributionally to check how often and where a particular phonological, lexical, grammatical or pragmatic feature occurs.

Corpora were used in linguistics before the development of computers, but around the early 1960s it was computer use that gave an enormous boost to corpus linguistics, by reducing the time taken to create, use and analyse a corpus, and greatly increasing the size of databases. The definition of a corpus can thus be modified as follows: "a corpus is a collection of texts in an electronic database" (Kennedy 1998, 3), and therefore a collection in *machine-readable* form. This feature allows for semi-automatic and automatic compilation and analysis. As far as size is concerned, corpora are becoming larger and larger, as they can be tagged, compiled and analysed automatically. The most important aspect to take into account when doing corpus work is that the corpus type and size must be appropriate for the research goal (Gries 2009).

In the last 20 years, cross-linguistic studies have used more and more multilingual corpora, which has helped the revitalization of research in this domain the same time. Aijmer and Altenberg (1996, 12) indicate some of the benefits of corpus-based study in language comparison:

- They give new insights into the languages compared, giving the researcher language-specific and language-universal information, as well as information on typological and cultural differences;
- They illustrate differences between source texts (authentic texts) and their translations, and between native and non-native texts;
- They can be used in numerous domains, such as language teaching, translation, typology, semantics, pragmatics, Natural Language Processing and Machine Translation, among others.

Aijmer and Altenberg note that there is a difference—in use and purpose—between *comparable* corpora and *translation* corpora. Comparable corpora contain original texts in a certain language, and specify that the texts share broad criteria, such as stylistic genre, domain, purpose of creation, time of creation, etc. Their main advantage is that they present natural language in use, and have the property of being *authentic*. The most difficult problem with using these corpora is knowing what to compare (e.g. relating forms which have similar meanings and pragmatic functions in the languages compared, as suggested by Johansson 1998) and to what extent comparisons might yield insights.

Translation corpora contain texts that are intended to express the same meaning, and have the same discourse functions in the languages considered (Johansson 1998). Dyvik (1998) suggested that translations reveal semantic features of the source language. She argues that translation is a linguistic activity where the translator evaluates meaning relations between expressions in an objective manner (as opposed to research which aims to develop or test a theory about relations of meaning between expressions). Translations thus provide objective linguistic data. For this reason, translation equivalence has been considered the best basis for comparison, and was used for a long time as the main principle for the construction of a *tertium comparationis* (Krzeszowski 1990) in CA. As a method used in linguistic research, corpus data have numerous advantages, as well as a series of limitations.

A great advantage of corpus data is that they allow both *qualitative* and *quantitative* analyses. In qualitative analyses of data, all sentences are treated with equal attention, and the results cannot be generalized, as they are limited to the sample of language analysed. In addition, no attempts are made to assign frequencies to the linguistic features identified in the data. In quantitative analyses, frequencies are assigned to linguistic features identified in the data; features are classified, counted and summarized. A basic step in quantitative analysis of data is to classify sentences or items according to a certain schema, and then count how many items (called *tokens* or *occurrences*) are in each group of the classification schema (called *types*). The result of this process is a *distribution* of the tokens in the corpus (McEnery and Wilson 1996).

Another advantage of working on corpora is that they form an empirical basis for researchers' intuitions. Intuitions are the starting point of any study, but can be misleading; sometimes, a few striking differences might lead to hazardous generalizations. Moreover, the results of analyses of quantifiable data allow not only generalizations (by way of statistical significance tests) but also predictions (by way of statistical analyses such as correlations[1] or multiple regression models,[2] which are often used when investigating a phenomenon as complex as language).

Corpus work is appealing when the researcher is concerned with a descriptive approach to the linguistic phenomenon considered, as well as the study of language in use, given the fact that the cotext is provided in the corpus. Corpora permit monolingual and cross-linguistic investigations. Furthermore, corpus work allows the researcher to uncover what is probable and typical, on the one hand, and what is unusual about the phenomenon considered, on the other hand.

Another advantage is that data from corpora can be annotated (enriched) with syntactic, semantic and pragmatic information, which allows more complex analyses of the corpus. Annotation is the practice of adding interpretative linguistic information to a corpus (Leech 2005), and thus enriching the original raw corpus. From this perspective, adding annotations to corpora provides additional value, and thus increases their utility (McEnery and Wilson 1996; Leech 2004). Firstly, annotated

[1] Correlation is a monofactorial statistical method, which investigates the relation between one independent variable (the predictor) and one dependent variable (the phenomenon of interest). Correlation does not obligatorily involve causality between the two variables (they can only be associated), and can be used only when the relationship is linear (Baayen 2008; Gries 2009).

[2] Multiple regressions are multifactorial statistical methods, which investigate the relation between several independent variables (predictors) and one dependent variable, as well as their interactions. The relation between independent variables and the dependent variable can be linear or non-linear (cf. Gries 2009, Baayen 2008).

corpora are useful both for researchers who make the annotations, and for other researchers, who can use them for their own purposes, modify them, or enlarge them. Secondly, annotated corpora allow both manual and automatic analysis and processing of the corpus; the annotations themselves often reveal a whole range of uses which would not have been practicable had the corpus not been annotated. Thirdly, annotated corpora allow an objective record of analysis which is itself open to future analysis, with decisions being more objective and reproducible. Due to automatic corpus analysis, annotated corpora are often used for training of Natural Language Processing and Machine Translation tools, such as automatic classifiers (Meyer et al. 2013).

One of the major issues with using translation corpora relates to their very nature, given the translation process and the way in which the source language can create a bias affecting the target text (the so-called *translationese* of Gellerstam 1996). Secondly, Baker (1993, 1995) points out that translated texts use *translation universals*, which are defined by Lefer (2009), quoting Laviosa (2002), as features of a translated language which are independent of the source language, such as simplification, explicitation and normalization. Thirdly, translated texts can only be compared to their original texts, and not to others. Another methodological concern when working with translation corpora is that they need to be aligned (at sentence or phrase level) and processed by parallel concordancers. As Lefer (2009) notes, alignment can be time-consuming, because automatic alignment requires manual control and correction for complete accuracy of data. Most parallel concordancers, such *ParaConc*, offer automatic pre-alignment tools. Two examples of well-known and well-used parallel corpora are *Europarl*[3] and *Hansard*.[4] Europarl is a corpus extracted from the proceedings of the European Union Parliament. It includes versions in 23 European languages, and the 1996 version contained 20 million words (Koehn 2005). The Hansard corpus is a bilingual corpus (English-French) of the proceedings of the Canadian parliament.

Other difficulties include the lack of multilingual corpora for less widespread languages, and the predilection for 'form-based research' where there is interest in a specific grammatical form (Granger 2003). These difficulties may require researchers to carry out their research manually, including building their corpus themselves, and annotating it if they are interested in phenomena other than a specific grammatical form, such as semantic or syntactic categories. Moreover, when researchers are interested in infrequent phenomena,[5] there will not be enough occurrences in the corpus. Difficulties are also encountered when investigating phenomena which are not lexically expressed, such as world knowledge used in inferences, as well as the cognitive basis of language. This is one reason why corpus data are more and more often combined with other types of data, such as experimental data.

[3] http://www.statmt.org/europarl/

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

[5] For example, Grivaz (2012) studied causality in certain pairs of verbs, in a very large corpus and with human annotation experiments, and found that less frequent pairs had a high causal correlation while very frequent pairs had a small causal correlation.

Translation Spotting

Translational spotting or *transpotting* is a technique that makes use of the translation of a specific word or linguistic expression in order to distinguish its meaning and disambiguate between its senses. This method has been used not only for content words (Dyvik 1998; Noël 2003) but also for discourse relations (Behrens and Fabricius-Hansen 2003) and connectives (Zufferey and Cartoni 2012; Cartoni et al. 2013). The term *translation spotting*, coined by Véronis and Langlais (2000), initially referred to the automatic extraction of a translated equivalent in a parallel corpus. Translation spotting consists in detecting the translation of a particular word or expression in the target text, as shown by the examples of connectives in Table 3.1 (Cartoni et al. 2013).

Table 3.1 is an example of the investigation of the usages of the English connective *since*, carried out in translation corpora. The second column contains the translation of the original English sentence into French. The third column contains the linguistic expressions or types of linguistic expression used in French when translating the English *since*, called *transpots*. The idea behind this analysis is that French transpots provide information regarding the diverse contextual usages of English *since.*

Véronis and Langlais point out the difficulty of automatically spotting the words or sequences of words from the target language when there is no one-to-one correspondence between the source and the target language. Automatic spotting results have errors, and the aim of researchers working in Natural Language Processing is to reduce the number of errors as much as possible. For this reason, other researchers (Cartoni et al. 2013; Grisot and Moeschler 2014) performed the spotting manually, in order to get fully accurate data. Cartoni and colleagues agree that, despite the fact that translations do not reproduce the source language faithfully and have a number of inherent features (Baker 1993), they still can shed light on the source

Table 3.1 Example of translation spotting for the connective *since*

	English sentence	French sentence	Transpot
1.	In this regard, the technology feasibility review is necessary, *since* the emission control devices to meet the ambitious NOx limits are still under development.	À cet égard, il est. nécessaire de mener une étude de faisabilité, étant donné que les dispositifs de contrôle des émissions permettant d'atteindre les limites ambitieuses fixes pour les NOx sont toujours en cours de développement.	étant donné que
2.	Will we speak with one voice when we go to events in the future *since* we now have our single currency about to be born?	Parlerons-nous d'une seule voix lorsque nous en arriverons aux événements futurs, puisqu'à présent notre monnaie unique est. sur le point de voir le jour?	puisque
3.	In East Timor an estimated one-third of the population has died *since* the Indonesian invasion of 1975.	Au Timor oriental, environs un tiers de l population est. décédée depuis l'invasion indonésienne de 1975.	depuis
4.	It is 2 years *since* charges were laid.	Cela fait deux ans que les plaintes ont été déposées.	paraphrase

language. They suggest that theoretical insights developed according to analysis of a parallel corpus should be validated by monolingual experiments.

The theoretical idea behind translation spotting is that similarities and differences in translation can reveal semantic features of the source language (Dyvik 1998; Noël 2003). Dyvik's idea is that the activity of translation is one of the very few cases where speakers evaluate meaning relations between expressions in an objective manner, without doing so as part of some kind of meta-linguistic, philosophical or theoretical reflection. From this perspective, he suggests using translation corpora as a basis for semantic analyses. This method presupposes the existence of a *translational relation* between two languages. There are two aspects to be distinguished before determining a translational relation. The first is information regarding *parole*[6] and textual *token* items; the second is information about *langue* and *type* items. In the first case, translation choices are motivated only by reference to the particular text and its circumstances, whereas in the second case, translation choices are predictable and reflect translation correspondence relations between words and phrases, seen as types rather than textual tokens. According to Dyvik, it is on this second aspect of language that a translational relation should be built. A translational relation consists of a series of properties or, more precisely, a series of senses shared partially by the linguistic expressions standing in that translational relation. Translational relations can be identified using the translation spotting technique. In particular, in Cartoni and colleagues' study, the English connective *since* is translated into French by four linguistic expressions (three connectives and a paraphrase). In a sentence completion task experiment, Cartoni and colleagues showed that the four French translation possibilities form two clusters: a causal sense (for *étant donné que* and *puisque*); and a temporal sense (for *depuis* et *cela fait X que*). The translational relation of *since* and its transpots in French consist therefore of two properties or senses partially shared by these linguistic expressions.

Translational relations reflect partial semantic equivalences between words and expressions in different languages. Therefore, they are concrete tools for developing cross-linguistic semantic representations. A semantic representation groups together a set of linguistic expressions, across languages, which fall within the denotation of the representation (Dyvik 1998). Such cross-linguistically valid semantic representations are useful for improving the results of several Natural Language Processing tasks, such as machine translation systems, multilingual dictionaries and concordances.

Cross-Linguistic Transfer of Properties

Cross-linguistic transfer of properties is a novel technique that makes use of the notion of translational relation and its properties. My suggestion is that translation corpora permit the cross-linguistic transfer of semantic and/or pragmatic information. Samardzic (2013) also made use of this novel methodology to investigate the

[6] The well-known linguist Ferdinand de Saussure was the first to make the distinction between *parole* and *langue*, where the former refers to acts of language of individual people, and the latter refers to language as an abstract entity, proper to a linguistic community.

translation equivalents of a range of English light verb constructions into several languages. Unlike other European languages, Slavic languages encode Aspect morphologically. She applies the aspectual representation obtained in an English-Serbian cross-linguistic setting to classify English verbs into event duration classes.

In an experiment, native speakers of English were asked to judge Simple Past tokens with respect to Aspect and its two values *perfective* vs. *imperfective*. Participants found the task extremely difficult, and they had a very low agreement rate. The cross-linguistic transfer of properties method was therefore used in order to create human-annotated data (i.e. Simple Past tokens) with aspectual information. A native speaker translated the data into Serbian, and identified the contextual value of Aspect for each Simple Past token. Based on the assumptions related to translation corpora, this aspectual information was transferred back to the initial English source (see Sect. 4.3.3).

3.2 Bilingual Corpus: English-French

For the specific needs of this research, parallel (also called *translation*) corpora consisting of texts of four registers have been assembled. The qualitative and quantitative analyses of the corpora were carried out in two steps: (a) an initial, monolingual step, in order to see which verbal tenses occur in the corpus and to calculate their frequency in the source language; and (b) a secondary, bilingual step, in order to identify the verbal tenses used as translation possibilities in the target language for a certain tense in the source language, as well as to calculate their frequency. Analysis of frequency of tenses in the source language provided information about the verbal tenses that might be problematic candidates for machine translation systems. The assumption of this procedure is that frequent erroneously translated verbal tenses decrease the quality of the translated text more than infrequent incorrectly translated verbal forms. In this chapter, I will describe the corpus and provide the results of its analysis.

The purpose of this chapter is to describe the corpus work which is the first layer of the empirical work presented in this thesis. This research is partly based on parallel or translation corpora, consisting of texts written in English and their translations into three target languages. Three corpora have been built. The first and the second corpora are bilingual, consisting respectively of texts written in English and their translations into French, and texts written in French and their translations into English. The third corpus is multilingual, consisting of texts written in English and their translations into French, Italian and Romanian. All texts have been randomly selected, and belong to four stylistic registers: literature,[7] journalistic, legislation

[7] A detailed presentation of the texts from the four registers is available in the Appendix section. Some of the corpora were aligned at the sentence level by other researchers and are available for download online, and some were created during the COMTIS research project (typed and aligned manually by the author).

and EuroParl (Koehn 2005).[8] The literature register consists of the texts of several novels, either written in English and translated into French, or written in French and translated into English. The English-French-Italian-Romanian corpus consists of randomly selected passages from “Alice in Wonderland” by Lewis Carroll and their translations into the target languages. The journalistic register consists of texts from several newspapers, all of which have an online version. For the multilingual corpus, all texts were randomly selected from the Press Europ website[9] and aligned manually. The legislation register consists of randomly selected law texts from the multilingual JRC-ACQUIS parallel corpus and the EuConst Corpus.[10] The EuroParl register consists of the transcription of parliamentary debates. The language used in the EuroParl corpus is spoken but transcribed, therefore presenting features of both spoken and written language.

The purpose of the monolingual analysis is to identify frequent and less frequent verbal tenses, whereas the purpose of the cross-linguistic analysis is to identify *translation divergences*, i.e. each verbal tense which is consistently translated into the target language by more than one verbal tense. In Sects. 3.2, 3.3 and 3.4, I will describe the corpora and provide the results of the corpus analysis.

3.2.1 Monolingual Analysis

The English-French bilingual corpus consists of texts in English and their translations into French, belonging to four different stylistic genres according to the following proportions: literature 15%; journalistic 16%; legislation 38%; and EuroParl 31%. The corpus contains 1670 occurrences of predicative verbal tenses, occurring in a total of 725 sentences.[11] A total of 1281 predicative verbal tenses are considered,[12]

[8] The EuroParl corpus is a collection of the proceedings of the European Parliament from 1996 to 2011. It is available online at http://www.statmt.org/europarl/

[9] The translation of the journalistic articles into all the languages officially spoken in the European Union is available online at http://www.voxeurop.eu

[10] The JRC-ACQUIS corpus was collected by the Language Technology team of the European Commission’s *Joint Research Centre* (JRC) in the context of the *Exploiting parallel corpora in up to 20 languages* workshop, held in Arona, Italy, on 26 and 27 September 2005. The EuConst corpus is a parallel corpus collected from the European Constitution (Tiedemann 2009).

[11] I use the word sentence to refer to a chunk of text, consisting of one or more complex clauses. Since verbal tense is a referential category and its meaning is underdetermined (as argued in Sect. 2.3), contextual and cotextual information is needed to determine its meaning. Therefore, the segmentation was performed manually, in order to decide the size of the text chunks relevant to determining the meaning of a verbal tense.

[12] The tenses under consideration are several tenses from the indicative mood: the simple present and past tenses, the present perfect and the past perfect, the present continuous and the past continuous.

Table 3.2 Verbal tenses by register in the English-French bilingual corpus

Register	No. of sentences	No. of verbal tenses	No. of verbal tenses considered	% of verbal tenses considered	% of verbal tenses not considered
Literature	118	255	232	14%	1%
Journalistic	155	275	228	14%	3%
EuroParl	136	512	403	24%	7%
Legislation	316	628	418	25%	13%
Total	725	1670	1281	77%	23%

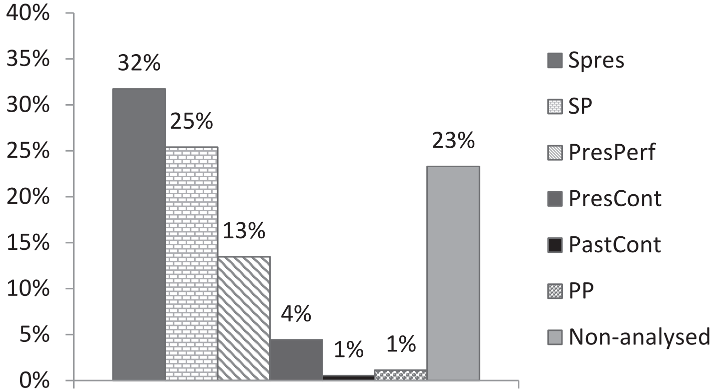


Fig. 3.1 Frequency of English verbal tenses in the English-French bilingual corpus

representing 77% of the verbal tenses occurring in the corpus, as shown in Table 3.2. The remaining 23% of verbal tenses have not been considered[13] in the analysis.

Figure 3.1 illustrates the frequency of verbal tenses analysed[14] in the English-French bilingual corpus, where the most frequent tenses are the Simple Present (32%), the Simple Past (25%) and the Present Perfect (13%), as opposed to the much less frequent past progressive, present progressive and past perfect verbal forms. This figure shows the unequal occurrence of verbal tenses in a corpus containing a total of 1670 predicative verbal forms relating to different stylistic registers. One possible explanation for the higher frequency of these verbal tenses is that they are highly context-dependent, and their interpretation depends on various contextual hypotheses.

Figure 3.2 presents the frequency of the three most frequent verbal tenses in each register. It shows that the Simple Past is the preferred tense in the literature register

[13] The non-analysed tenses are other tenses from the indicative mood (present perfect continuous and past perfect continuous, and all future tenses), English verbal tenses with conditional and subjunctive readings, and modal verbs.

[14] Legend: SP = Simple Past, PresPerf = Present Perfect, PresCont = Present Continuous, Spres = Simple Present, PastCont = Past Continuous, PP = Past Perfect and Non-analysed.

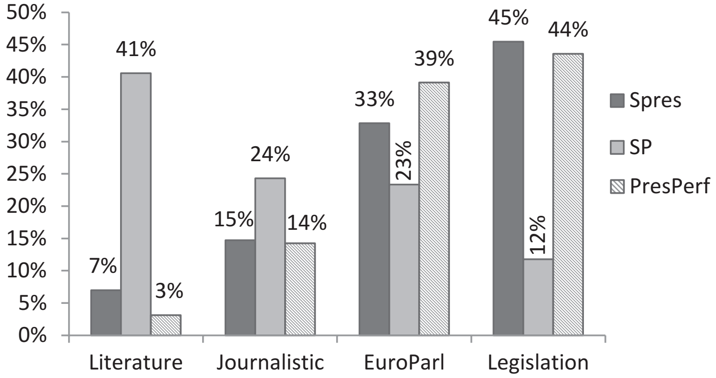


Fig. 3.2 Frequency of English tenses by register

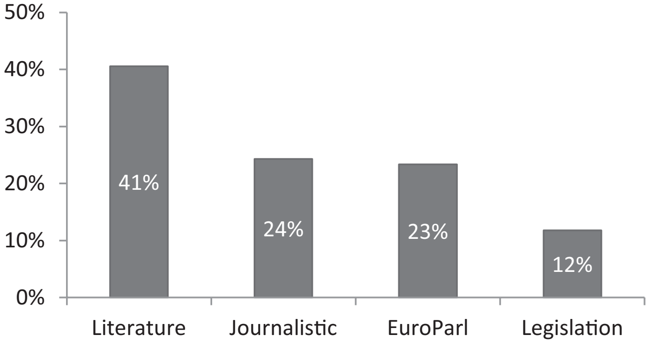


Fig. 3.3 The distribution of the SP by register

(representing 41% of the predicative tenses used) and in the journalistic register (24%). The Simple Present and the Present Perfect are used more frequently than the Simple Past in the EuroParl and legislation registers. The Simple Present and the Present Perfect show a similar distribution in the journalistic (15% and 14% respectively), EuroParl (33% and 39% respectively) and legislation registers (45% and 44% respectively), in contrast to the Simple Past.

These distributions are not surprising. Firstly, the Simple Past is preferred in narratives, instructing the addressee to order eventualities temporally with respect to one another. Secondly, the legislation register is a prospective and deontic register, and the Simple Present is a verbal tense appropriate for the expression of these interpretations (like the French Présent). The journalistic and EuroParl registers consist of mixed types of texts (small narratives, comments, descriptions, etc.).

Figure 3.3 shows the distribution of all Simple Past occurrences by register. 41% of Simple Past occurrences come from the literature register, with the remaining 59% shared between the journalistic (24%), EuroParl (23%) and legislation (12%) registers.

Table 3.3 English-French translation possibilities

English	Spres	PP	PastCont	PresCont	PresPerf	SP
French	PRES 81%	PQP 58%	IMP 67%	PRES 85%	PC 68%	PC 33%
						IMP 29%
					PRES 13%	PS 18%
						PRES 5%
	Others 19%	Others 42%	Others 33%	Others 15%	Others 19%	Others 15%

To sum up, the monolingual analysis of this corpus reveals that the most frequent verbal tenses are the Simple present, the Simple Past and the Present Perfect. In Sect. 3.2.2, I will provide the results of the cross-linguistic analysis, which show which verbal tenses consistently have more than one translation possibility in French (i.e. are ambiguous for machine translation systems).

3.2.2 Cross-Linguistic Analysis

The cross-linguistic analysis was performed using the *translation spotting* method, in order to identify *translation divergences*. A translation divergence occurs where a verbal tense for which there are at least two translation possibilities in the target language which are much more frequent than all the other possibilities. The analysis revealed two translation divergences among the verbal tenses considered: namely, the Simple Past and the Present Perfect. The results from Table 3.3 indicate that each of the first four verbal tenses is consistently translated into French by one frequent verbal form. Explicitly, the Simple Present is most often translated by the Présent,[15] the Past Perfect is most often translated by the Plus-que-parfait,[16] the Past Continuous is most often translated by the Imparfait[17] and the Present Continuous is most often translated by the Présent.[18]

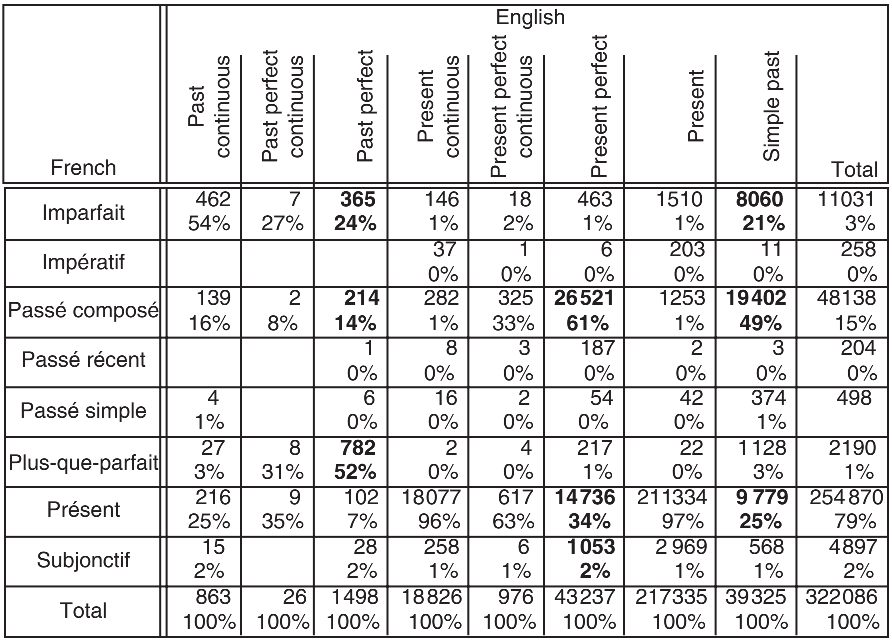
The Present Perfect is one of the two translation divergences identified. The Passé Composé is most often used; the Présent is much less used, though it is more

[15] The Others category consists of very infrequent cases, such as 0 translation (5%), present participle, past participle and modal verbs (2% for each form), conditional, future, Imparfait, Passé Composé, Passé Simple, infinitive and noun (1% for each form) and infinitive (0.2%), forming a total of 19%.

[16] The Others category consists of Imparfait (3 occurrences representing 16%), Passé Composé (2 occurrences representing 11%), subjunctive, participle and anterior past (1 occurrence representing 5% for each form), forming a total of 42%.

[17] The Others category consists of Plus-que-parfait, noun and the *était en train de* lexical construction (1% for each form representing 11%), for a total of 33%.

[18] The Others category consists of Imparfait (4 occurrences representing 5%), 0 translation (3 occurrences representing 4%), modal verbs (2 occurrences representing 3%), future and Passé Composé (1 occurrence representing 1% for each form), forming a total of 15%.



	English								
French	Past continuous	Past perfect continuous	Past perfect	Present continuous	Present perfect continuous	Present perfect	Present	Simple past	Total
Imparfait	462 54%	7 27%	**365** **24%**	146 1%	18 2%	463 1%	1510 1%	**8060** **21%**	11031 3%
Impératif				37 0%	1 0%	6 0%	203 0%	11 0%	258 0%
Passé composé	139 16%	2 8%	**214** **14%**	282 1%	325 33%	**26521** **61%**	1253 1%	**19402** **49%**	48138 15%
Passé récent			1 0%	8 0%	3 0%	187 0%	2 0%	3 0%	204 0%
Passé simple	4 1%		6 0%	16 0%	2 0%	54 0%	42 0%	374 1%	498
Plus-que-parfait	27 3%	8 31%	**782** **52%**	2 0%	4 0%	217 1%	22 0%	1 128 3%	2190 1%
Présent	216 25%	9 35%	102 7%	18077 96%	617 63%	**14736** **34%**	211334 97%	**9 779** **25%**	254 870 79%
Subjonctif	15 2%		28 2%	258 1%	6 1%	**1053** **2%**	2 969 1%	568 1%	4897 2%
Total	863 100%	26 100%	1498 100%	18826 100%	976 100%	43237 100%	217335 100%	39325 100%	322086 100%

Fig. 3.4 Distribution of the French translation labels for 322,086 English verb phrases in EuroParl

frequent than any other form.[19] Finally, the Simple Past is the most significant translation divergence. It is translated into French by four verbal tenses. The first three are past time tenses (Passé Composé, Imparfait and Passé Simple) and the fourth is the present tense (i.e. the Présent).[20]

This distribution was confirmed by Loáiciga et al. (2014), who automatically examined 322,086 finite English verb phrases from the EuroParl corpus and their translation into French. The verbal tenses expressing past time (shown in bold in Loáiciga et al.'s table, provided in Fig. 3.4) exhibit significant translation divergences. They found that the Simple Past and the Present Perfect translation divergences are statistically significant ($p < 0.05$). For example, the English Present Perfect (the seventh column in Loáiciga et al.'s table) can be translated into French either with a Passé Composé (61% of English-French pairs), a Présent (34%) or a

[19] The Others category consists of past participle (11 occurrences representing 5%), subjunctive (7 occurrences representing 3%), noun (8 occurrences representing 4%), 0 translation (4 occurrences representing 2%), Imparfait, *venir de*, past infinitive, anterior future, Plus-que-parfait (2 occurrences representing 1% for each form), participle and past conditional (1 occurrence representing 0.5% for each form), forming a total of 19%.

[20] The Others category consists of 0 translation (14 occurrences representing 3%), past participle, PQP, subjunctive (10 occurrences representing 2% for each form), conditional, past infinitive, noun and present participle (4 occurrences representing 1% for each form), past conditional, infinitive (2 occurrences representing 0.5% for each form) and *venir de* (1 occurrence representing 0.2%), forming a total of 15%.

subjunctive (2%). Similarly, the Simple Past (the ninth column) can be translated either by a Passé Composé (49% of pairs), by a Présent (25%), or by an Imparfait (21%).

My hypothetical explanation for this linguistic variation in the French forms used to translate each of the verbal tenses considered is that the most frequent translated tenses share semantics and pragmatics with the source verbal tense, and are predictable forms. In contrast, less frequent forms (included in the Others category in our analysis) are context-dependent—i.e. depend on the specific type of text, its purpose, the translator's personal choice, etc.—and are unpredictable forms. Following Dyvik (1998), I suggest that predictable forms are about *langue* and *type* items, whereas unpredictable forms are about *parole* and *token* items. As far as this thesis is concerned, I will deal only with the predictable forms of the Simple Past translation divergence.

The Simple Past translation divergence is interpreted as following: the Simple Past has several usages, corresponding to several French tenses used as its translation possibilities. The French tenses used to render the semantic and pragmatic meaning of the Simple Past are: the Imparfait; the Passé Composé; the Passé Simple; and the Présent. The Passé Composé is used most frequently in the EuroParl and journalistic registers, whereas the Passé Simple is used most frequently in the literature register, and the Présent in 10% of the cases in the legislation register, in order to create a certain effect in deontic contexts, as shown in Fig. 3.5. The variability indicated by this distribution shows that stylistic register is not a good predictor of the verbal tense used in the target language. For example, in the literature genre, the Simple Past is translated by an Imparfait in 44% of cases, and by a Passé Simple in 40%. Hence, establishing a translation rule based on a one-to-one correspondence is not possible.

Examples (439)–(441) depict the translation divergence of the English Simple Past: in (439) the Simple Past is translated by the French Imparfait; in (440) by the Passé Composé; in (441) by the Passé Simple; and in (442) by the Présent.

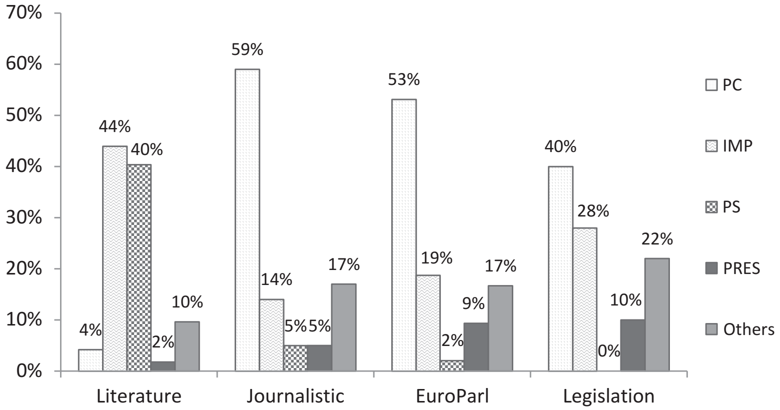


Fig. 3.5 Translation possibilities for the English Simple Past into French (column distribution)

(439) EN/SP: The atmosphere *had* more to do with the negative aspects of a great European project and vision than a positive promotion of what is deep and good about the European dream, and that is a disappointing feature of Nice. (EuroParl Corpus)
FR/IMP: 'L'ambiance *avait* plus à voir avec les aspects négatifs d'un grand projet et d' une grande vision pour l' Europe qu'avec une promotion positive de ce que le rêve européen a de profond et de positif, et c'est là un aspect décevant de Nice.'

(440) EN/SP: I welcome the consultation process and can assure colleagues that in my Member State the authorities *took care* to carry out a broad and meaningful consultation. (EuroParl Corpus)
FR/PC: 'Je me félicite du processus de consultation et je peux assurer mes collègues que les autorités de mon pays *ont pris soin* de mener une consultation vaste et significative.'

(441) EN/SP: Cyril had very little affection for him, and was only too glad to spend most of his holidays with us in Scotland. They never really *got on* together at all. (Literature Corpus)
FR/PS: 'Cyril avait fort peu d'affection pour lui, et n'était que trop heureux de passer l'essentiel de ses vacances avec nous en Ecosse. Ils ne s'*entendirent* jamais véritablement.' (Literature Corpus)

(442) EN/SP: Something else they *had in common* was that they either conflicted with existing legal instruments or duplicated them. (EuroParl Corpus)
FR/PRES: Ces initiatives *ont* également *en commun* que tantôt, elles sont en contradiction avec les instruments juridiques existants, tantôt, elles les dupliquent.

Corpus analysis reveals that there is a mismatch between theoretical descriptions of verb tenses and actual usages in corpora. Certain verb tenses that the theoretical literature predicts to be ambiguous for translation purposes, such as the English Past Continuous or Past Perfect, are infrequent in the corpus described in this section. Others, such as the English Simple Present and Simple Past, are ambiguous and frequent, and thus represent a significant translation divergence.

Regarding the theoretical description of the Simple Past in terms of the Reichenbachian coordinates S, R and E, the SP shares the same configuration only with the Passé Simple (E = R < S). Even though the Imparfait has the same configuration as the Passé Simple, Reichenbach (1947) emphasizes that the two verbal tenses are different: the first is extended (i.e. progressive), and the latter non-extensive. Moreover, the Passé Composé, which is the most frequent verbal tense used to translate the SP, has a different temporal configuration from the Simple Past—that is, E < R = S. Finally, the fourth tense used to translate the Simple Past is the Présent, which is described as E = R = S. There are two questions that arise at this point in the discussion. The first regards the relation between the source and target languages: what do the verbal tenses used in the target language reveal about the verbal tense used in the source language? The second question regards the factors which explain and predict this cross-linguistic variation. Several candidate features are tested experimentally in Chap. 4, wherein Sect. 4.4 provides a multifactorial analysis of the data.

3.3 Bilingual Corpus: French-English

3.3.1 Monolingual Analysis

The corpus consists of texts written in French and their translations into English, belonging to four different genres, according to the following proportions: literature 24%; journalistic 25%; legislation 21%; and EuroParl 31%. The corpus contains 1283 occurrences of predicative verbal tenses, occurring in a total of 603 sentences. A total of 1031 predicative verb tenses have been considered,[21] representing 80% of the verb tenses occurring in the corpus, as shown in Table 3.4. The remaining 20% of verbal tenses have not been considered[22] in the analysis.

Figure 3.6 illustrates the frequency of verbal tenses[23] in the corpus, where the most frequent are the Présent (37%), the Passé Composé (19%) and the Imparfait (14%), as opposed to the much less frequent Passé Simple and Plus-que-parfait (9% for the former and 3% for the latter).

Figure 3.7 presents the frequency of the analysed verbal tenses in each register. The Présent is the preferred tense in the journalistic and legislation registers (29% in the former and 26 in the latter), whereas the Passé Composé, Imparfait and Passé Simple are much less frequent (expect the Passé Composé, used in 14% of the cases in legislation). The distribution of these verbal tenses is more even in the literature and EuroParl registers.

Table 3.4 Verbal tenses by register in the French-English bilingual corpus

Register	No. of sentences	No. of verbal tenses	No. of verbal tenses considered	% of verbal tenses considered	% of verbal tenses not considered
Literature	162	305	275	21%	2%
Journalistic	172	320	220	17%	8%
EuroParl	180	392	332	26%	5%
Legislation	89	266	204	16%	5%
Total	603	1283	1031	80%	20%

[21] The tenses considered are several from the indicative mood, such as Imparfait, Passé Simple, Passé Composé, Présent and Plus-que-parfait.

[22] Non-analysed tenses are other tenses from the indicative mood, as well as other moods and modal verbs.

[23] Legend: PRES = Présent, PC=Passé Composé, IMP=Imparfait, PS = Passé Simple, PQP = Plus-que-parfait and Non-analysed.

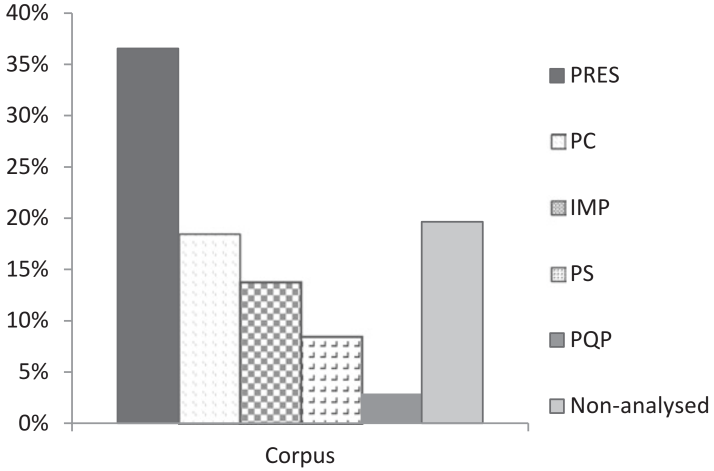


Fig. 3.6 Frequency of French verbal tenses in the French-English bilingual corpus

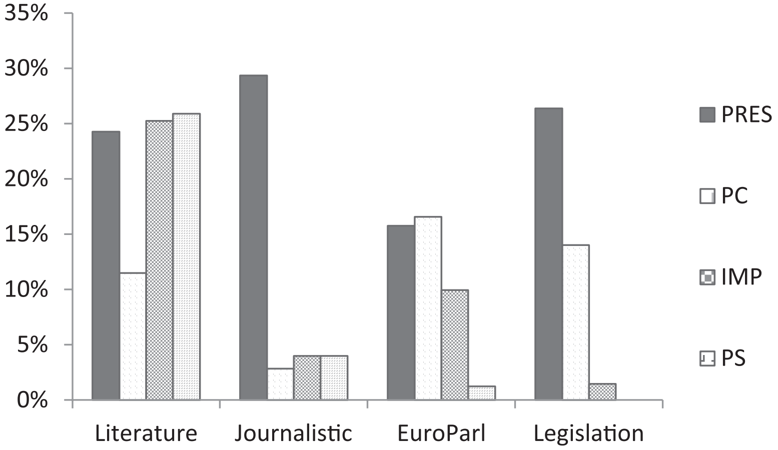


Fig. 3.7 Frequency of French verbal tenses by register

Figure 3.8 presents the distribution of each French verbal tense considered in the four registers. The Présent verbal tense occurs most often in the journalistic register (33%). 27% of the Présent tokens analysed occur in legislation, and 24% in EuroParl. Finally, 16% of the tokens come from the literature register. This distribution shows that the Présent is not specialized for any stylistic register. The Imparfait is often used in literature (44% of the Imparfait tokens) and EuroParl (41% of the Imparfait tokens). The Passé Simple and Passé Composé also seem to be stylistically specialized. In particular, most of the Passé Composé tokens occur in the EuroParl and legislation registers (51% and 27% respectively), whereas 72% of the Passé Simple tokens occur in the literature register.

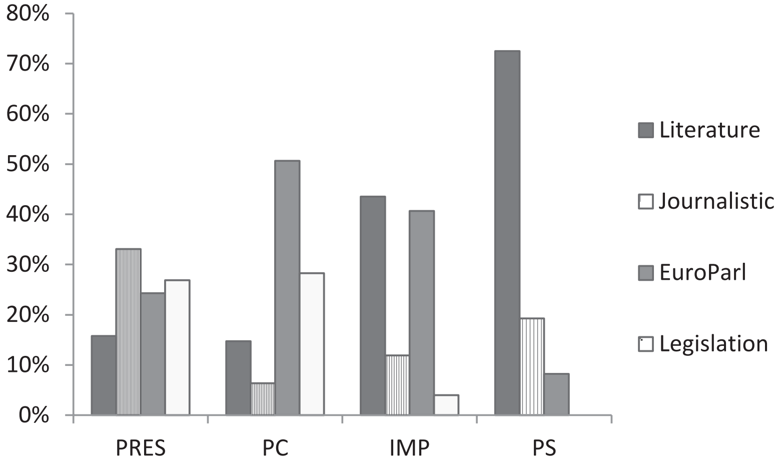


Fig. 3.8 Distribution of French verbal tenses in all registers

These results refute the predictions made in the literature—especially in the classical discourse analysis field—with respect to using verbal tenses exclusively in one or another stylistic register or type of discourse. For example, Weinrich (1973) predicts that the French Passé Simple is only used in texts of the *monde raconté* 'story' type (i.e. literature), as opposed to texts coming of the *monde commenté* 'commentary' type (i.e. journalistic, legislation and parliamentary discussion, among others), where other past time verbal tenses, such as the Passé Composé, are used.[24] Figure 3.8 indicates the Passé Simple is not exclusively used in the literary register, but also in the journalistic and EuroParl registers.

To sum up, the monolingual analysis of this corpus reveals that the most frequent verbal tenses are the Présent, the Passé Composé and the Imparfait. In Sect. 3.3.2, I will provide the results of the cross-linguistic analysis, which will show which verbal tenses consistently have more than one translation possibility in English.

3.3.2 Cross-Linguistic Analysis

The cross-linguistic analysis of the parallel corpora, performed by the translation spotting method, revealed two translation divergences among the verbal tenses considered: namely, the Passé Composé and the Plus-que-parfait. The results from Table 3.5 indicate that each of the first three verbal tenses considered (i.e. the Imparfait, Passé Simple and Présent) is consistently translated into English by one verbal form (i.e. the most frequent possibility for translation into the target

[24] For a critical discussion of discursive and textual theories regarding French verbal tenses, see de Saussure (2003).

Table 3.5 French-English translation possibilities

French	IMP	PS	PRES	PC	PQP
English	SP 82%	SP 93%	Spres 61%	SP 47.6%	SP 52.6%
				PresPerf 42.9%	PP 28.9%
					PresPerf 10.5%
	Others 18%	Others 7%	Others 39%	Others 9.5%	Others 7.8%

language). It can be seen that the Imparfait is most often translated by the Simple Past,[25] the Présent is most often translated by the Simple Present,[26] and the Passé Simple is most often translated by the Simple Past.[27] The Passé Composé is one of the two translation divergences identified for the French into English direction of translation. In particular, the Simple Past and the Present Perfect are far more frequent translation possibilities than the other forms.[28] Finally, the Plus-que-parfait is the second translation divergence. It is translated into English by three tenses: the Simple Past, Past Perfect and Present Perfect.[29]

Examples (443) and (444) illustrate the translation divergence of the French Passé Composé: in the former example, the Passé Composé is translated by the Simple Past, while in the latter, by the Present Perfect verbal tense.

(443) French/PC: Une chance à laquelle, comme l'*a dit* notre collègue Böge, nous devons maintenant donner une forme concrète. (EuroParl Corpus)
English/SP: 'An opportunity that must be given concrete shape, as the honorable Member Böge *said*.'

(444) French/PC: J'*ai volé* un peu partout dans le monde. Et la géographie, c'est exact, m'a beaucoup servi. (Literature Corpus)
English/PresPerf: 'I *have flown* a little over all parts of the world; and it is true that geography has been very useful to me.'

[25] The Others category consists of Past Perfect (12 occurrences representing 7%), *would* (7 occurrences representing 4%), Present Perfect, gerund (3 occurrences representing 2% for each form), Simple Present, Past Continuous, 0 translation, infinitive, past perfect continuous (2 occurrences representing 1% for each form), forming a total of 18%.

[26] The Others category consists of future (62 occurrences representing 13%, exclusively in the legislation register), Simple Past, Present Continuous, 0 translation (23 occurrences representing 5% for each form), Present Perfect (20 occurrences representing 4%), modal verbs (13 occurrences representing 3%), gerund (11 occurrences representing 2%), infinitive, past participle (4 occurrences representing 1% for each form), Past Continuous (2 occurrences representing 0.4%) and *would* (1 occurrence representing 0.2%), forming a total of 39%.

[27] The Others category consists of Past Perfect (4 occurrences representing 4%), modal verbs (2 occurrences representing 2%) and gerund (1 occurrence representing 1%).

[28] The Others category consists of Simple Present (7 occurrences representing 3%), Past Perfect, 0 translation (5 occurrences representing 2% for each form), Present Continuous, modal verbs (2 occurrences representing 1% for each form) and gerund (1 occurrence representing 0.4%), forming a total of 9.5%.

[29] The Others category consists of past participle, gerund and Simple Present (3% for each form), forming a total of 7.8%.

Examples (445)–(447) illustrate the translation divergence of the French Plus-que-parfait: in the first example, the Plus-que-parfait is translated by the Simple Past; in the second, by the Past Perfect; and in the third, by the Present Perfect.

(445) French/PQP: Dans les années 1570, le sang des protestants massacrés *avait* littéralement *ruisselé* dans les rues de Paris, et le conflit qui s'en était suivi avait déchiré le pays pendant des générations. (Journalistic Corpus)
English/SP: In the 1570s, Paris literally *flowed* with the blood of slaughtered Protestants, and the ensuing conflict tore the country apart for generations.

(446) French/PQP: Le père du jeune Fergusson, un brave capitaine de la marine anglaise, *avait associé* son fils, dès son plus jeune âge, aux dangers et aux aventures de sa profession. (Literature Corpus)
English/PP: 'Ferguson's father, a brave and worthy captain in the English Navy, *had associated* his son with him, from the young man's earliest years, in the perils and adventures of his profession.'

(447) French/PQP: De plus, ce n'est pas la première fois que j' interviens dans un parlement - y compris celui -ci - et jamais personne ne m' *avait accusé* de faire de la flibusterie, bien au contraire. (EuroParl Corpus)
English/PresPerf: 'Furthermore, this is not the first time I have spoken in a parliament – this not even the first time I have spoken in this one – and nobody *has* ever *accused* me of filibustering.'

As with the English into French direction of translation, corpus works reveals mismatches between theoretical descriptions of verbal tenses and their actual usage in human communication. The Passé Composé and Plus-que-parfait represent cases where theoretical description, with the help of Reichenbachian temporal coordinates, seems to need improvements. In particular, the French Passé Composé is described as having the same temporal configuration as the Present Perfect (i.e. $E < R = S$). In other words, the Passé Composé and the Present Perfect are expected to be in a one-to-one translation correspondence—i.e. to share the same semantic and pragmatic content. The corpus work described in this section provides evidence against this association, and questions the classical configuration suggested for the Passé Composé. A linguistic theory of the meaning of the French Passé Composé must explain cases where the Passé Composé is translated by a Simple Past, as well as cases where it is translated by a Present Perfect.

Another interesting case is the Plus-que-parfait, which, is considered to have the same temporal configuration as the English Past Perfect (i.e. $E < R < S$). However, corpus work reveals that the Past Perfect is only one of the three verbal tenses used to translate the Plus-que-parfait into English (in 29% of cases). As shown in Table 3.5, the Simple Past is used in 58% of cases, and the Present Perfect in 11%. As with the Passé Composé, theoretical semantics and pragmatics need to provide an explanation for the Plus-que-parfait translation divergence.

To sum up, Sects. 3.2 and 3.3 provided quantitative and qualitative analyses of verbal tenses and their usage in the source language, as well as their translation possibilities into a target language. Two directions of translation were considered: English into French, and French into English. Cross-linguistic analyses have indicated the most problematic translation divergences in each of the two directions of translation. In this thesis, one translation divergence is systematically investigated, namely the translation of the Simple Past into a target language. In order to increase the empirical basis of this research, two other Romance languages are added: Italian and Romanian. The results of multilingual corpus analysis are provided in the following section.

3.4 Multilingual Corpus

The multilingual corpus consists of texts written in English and their translations into French, Italian and Romanian. This kind of corpus is called a *parallel translations* corpus (Granger 2003). The main advantage of parallel translations corpora is that one can identify language-independent patterns—i.e. translators' systematic choices regarding the target languages when dealing with the same form in the source language. The multilingual corpus described in this section was built to identify language-independent patterns for the translation of the English Simple Past. In Sect. 3.4.1, I will describe how data were collected, and in Sect. 3.4.2 I will provide the results of the corpus analysis by target language.

3.4.1 Data Collection

The multilingual corpus was created with the specific purpose of analysing the translation of the English Simple Past into three target languages. The chosen languages belong to the same language family: i.e. the Romance languages. Within the family, however, they belong to different groups. As noted by Hall (1964), Romanian belongs to the Eastern group, whereas Italian and French belong to the Italo-Western group, which is further divided into Western Romance (Portuguese, Spanish, Catalan, Occitan and French) and Proto-Italian (Italian). This choice of language allows for the control of cross-linguistic variance due to structural differences between languages.

To guarantee comparability with the bilingual corpus (English-French, described in Sect. 3.2), the multilingual corpus consists of texts belonging to the same stylistic registers: literature; EuroParl; legislation; and journalistic.[30] The occurrences of the SP were randomly selected from the English texts, then aligned with their translations into French, Italian and Romanian. Regardless of language or stylistic register,

[30] A detailed presentation of the texts used for data collection is provided in the Appendix section.

Table 3.6 Description of the multilingual corpus

	Literature	EuroParl	Legislation	Journalistic
English – French/Italian	38%	19%	25%	18%
English – Romanian	39%	16%	26%	18%

Table 3.7 Translation possibilities for the Simple Past into French, Italian and Romanian in the multilingual corpus

	French	Italian	Romanian
Compound past	37%	33%	49%
Imperfect	24%	18%	15%
Simple past	16%	22%	18%
Present	8%	5%	5%
Others	16%	21%	13%

the texts are all parallel translations, other than the English-Romanian data from the EuroParl register. Since Romania joined the European Union later than France and Italy, the Romanian data in EuroParl are available only after 2004. Therefore, the English into French/Italian data consist of parallel translations where the English into Romanian data are a separate file. Table 3.6 provides the percentage, by register type, of Simple Past occurrences in the source texts. 513 occurrences of the Simple Past and their translations into three target languages (a total of 1281 sentences in the four languages) were analysed.

3.4.2 *Analysis and Results*

The corpus was analysed from a cross-linguistic perspective using the translation spotting method. The results from Table 3.7 indicate that all three target languages make use of the same verbal forms most frequently. In particular, the French data from the multilingual corpus are comparable to the French data[31] from the bilingual corpus, described in Sect. 3.3.2. The Italian data show that the Passato Prossimo accounts for 33% of cases, followed by the Passato Remoto at 22%, the Imperfetto at 18%, the Presente at 5% and, finally, several other linguistic forms included in the Others[32] category. In Romanian, the Perfectul Compus is by far the most frequent

[31] The French Others category consists of noun (12 occurrences representing 3%), 0 translation, past participle, Plus-que-parfait and subjunctive (10 occurrences representing 2% for each form), gerund, infinitive, rephrase (6 occurrences representing 1% for each form), and conditional (1 occurrence representing 0.2%), forming a total of 16%.

[32] The Italian Others category consists of past participle (17 occurrences representing 4%), noun, 0 translation, Trapassato prossimo (i.e. the pluperfect), subjunctive, rephrase (12 occurrences representing 3% for each form), gerund, infinitive (3 occurrences representing 1% for each form) and conditional (1 occurrence representing 0.7%), forming a total of 21%.

Table 3.8 Frequency of verbal tenses in French, Italian and Romanian by register

	Verbal tense	Literature	EuroParl	Legislation	Journalistic
French	Passé simple	**40%**	0%	0%	1%
	Imparfait	*35%*	*17%*	*14%*	*22%*
	Passé composé	10%	**45%**	**63%**	**49%**
	Présent	1%	17%	12%	9%
Italian	Passato Remoto	**55%**	0%	0%	9%
	Imperfetto	*28%*	*12%*	*14%*	*10%*
	Passato Prossimo	1%	**40%**	**64%**	**52%**
	Presente	0%	10%	7%	8%
Romanian	Perfectul Simplu	**45%**	0%	0%	0%
	Imperfectul	*29%*	*7%*	*5%*	*9%*
	Perfectul Compus	13%	**76%**	**75%**	**66%**
	Prezentul	1%	3%	11%	8%

verbal tense used (49%), followed by Perfectul Simplu (18%), Imperfectul (15%), Prezentul (5%), and other linguistic forms included in the Others[33] category.

Table 3.8 provides the frequency of each verbal tense considered in each register, for each target language. It can be seen that, for all three languages, and for each register, verbal tenses have similar distributions. In particular, the most frequent verbal tenses in the literature register are the simple past and the imperfect. In the EuroParl, legislation and journalistic registers, it is the compound past which is most frequently used, with the simple past almost non-existent. This distribution could be interpreted as a register specialization for the simple past, showing the complementarity of the two verbal tenses expressing past time. In each register, and for all three languages, the imperfect is the second most frequent tense. Based on these data, and on theoretical considerations, I suggest reducing the English Simple Past translation divergence to a three-way divergence: simple and compound past; imperfect; and the simple present.

This interpretation is also shown in Figs. 3.9, 3.10 and 3.11 for each of the target languages. From these figures, one can see that, in French, 99% of the Passé Simple occurrences are in the literature register, and the remaining 1% in the journalistic register. In Italian, 93% of the Passato Remoto occurrences are in the literature register, and the remaining 7% in the journalistic register. In Romanian, all occurrences of the Perfectul Simplu belong to the literature register. As for the compound past, it has the lowest frequencies in the literature register in all three languages—at its lowest in Italian, at 1%. Regarding the imperfect, most of the occurrences are in the literature register, in all three languages. The EuroParl, legislation and journalistic registers also make use of the imperfect, with a frequency of 12% in French, 10% in Italian and 7% in Romanian. Finally, the lowest frequencies of the simple present

[33] The Romanian Others category consists of noun (11 occurrences representing 3%), 0 translation, past participle (9 occurrences representing 2% for each form), Mai mult ca perfectul (i.e. pluperfect), subjunctive, gerund and conditional (3 occurrences representing 1% for each form), infinitive and future (3 occurrences representing 0.5% for each form), forming a total of 13%.

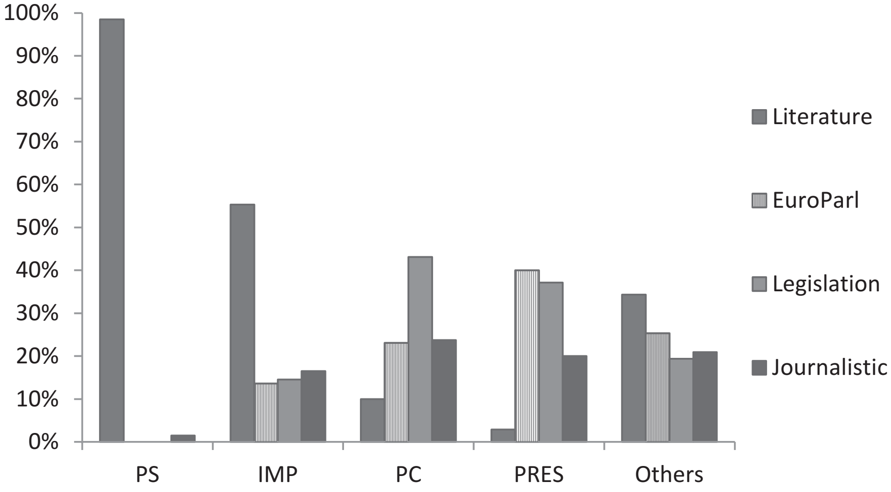


Fig. 3.9 Frequency of French verbal tenses

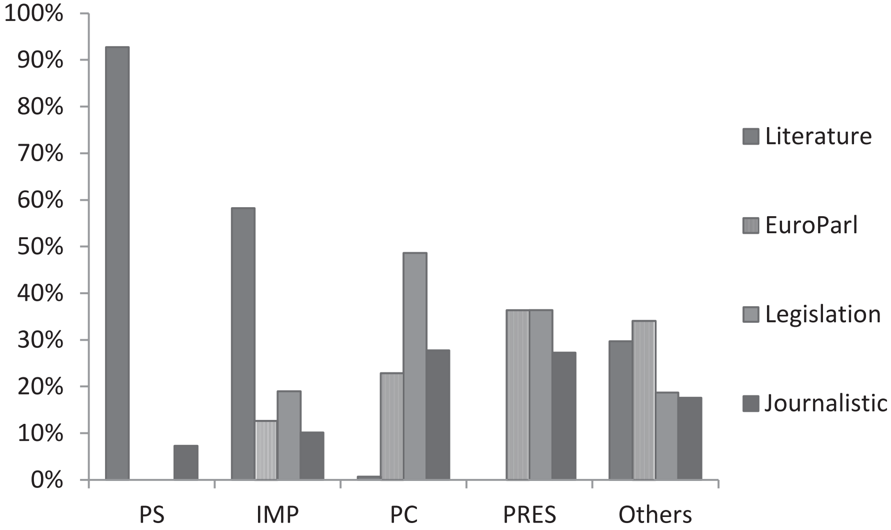


Fig. 3.10 Frequency of Italian verbal tenses

are in the literature register; the highest frequencies, on the other hand, are in EuroParl in French (40%), EuroParl and legislation in Italian (36% for each of the two registers), and legislation in Romanian (57%).

To sum up, the translation divergence of the English Simple Past identified in the English-French bilingual corpus is confirmed by the multilingual corpus. The Simple Past is most frequently translated into French, Italian and Romanian by a

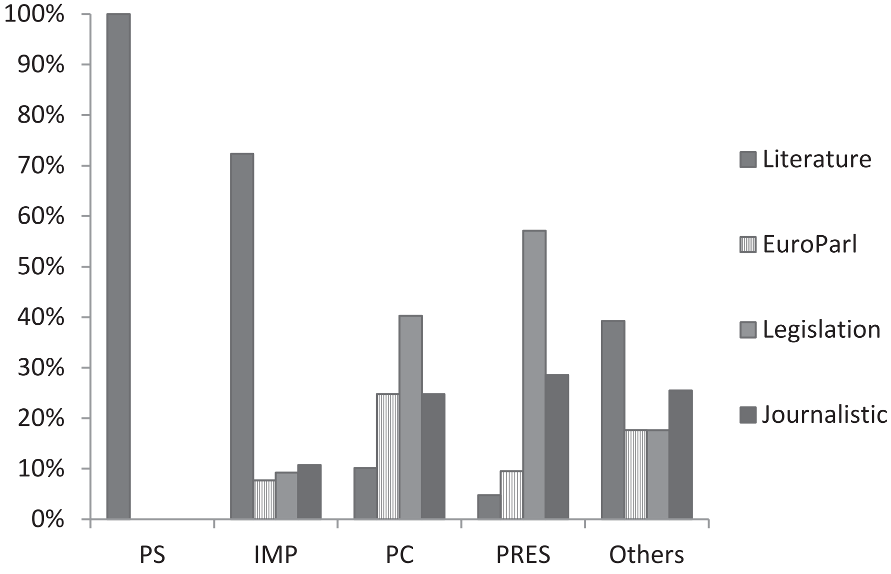


Fig. 3.11 Frequency of Romanian verbal tenses

simple past form in the literature register, and by a compound past form in the other three registers. Similarly, the simple present is used almost exclusively in the EuroParl, legislation and journalistic registers, in all three languages. Finally, the imperfect is used in all four registers to translate a Simple Past.

3.5 Summary

My aim in this chapter was to assess how verbal tenses are used cross-linguistically, by investigating them both in the monolingual side and the translation side of the parallel copora. At the beginning of the chapter, I explained that scholars turned to corpora because of the need for objective, quantifiable and reproducible data. In addition, pragmaticians have also adopted corpus data, to complement or even replace intuitive data, in order to sharpen intuitions, develop and test hypotheses, and avoid basing their research on scant data.

The quantitative analyses of the data indicated that certain verbal tenses are more frequent—and more problematic, with respect to their translation in a target language—than others. Firstly, the analysis of the English-French parallel corpus revealed two main translation divergences: the Present Perfect, and the Simple Past. These two verbal tenses are both frequent in the corpus, and ambiguous: i.e. each of them is systematically translated into a target language by at least two verbal forms. Secondly, the analysis of the French-English parallel corpus revealed two main translation divergences: the Passé Composé, and the Plus-que-parfait. The Passé

Composé is both frequent and ambiguous, whereas the plus-que-parfait is ambiguous but much less frequent. Thirdly, the parallel translations corpus confirmed the Simple Past translation divergence identified in the bilingual corpus. The data on Italian and Romanian provided further evidence justifying the inclusion of the compound past and simple past in one unified category, so reducing the initial four-way divergence to a three-way divergence.

Based on the results, and applying the principle that the most frequent and most ambiguous verbal tense will be the most problematic for machine translation systems, the Simple Past translation divergence was chosen for further experimental investigations, which I will discuss in the next chapter. In this research, the term *disambiguation* does not imply that the Simple Past is polysemous. On the contrary, as argued in Sect. 2.3, Tense is an underdetermined linguistic category which must be worked out contextually. Consequently, a verbal tense does not have several *meanings*, but several contextual *usages*. The notion of the *disambiguation model* therefore refers to disambiguation between the various usages of the Simple Past. The basic idea is that the Simple Past has several usages, and each of these usages may be translated into a certain target language through a different verbal tense.