Part III: Empirical studies

In this part, I describe what constitutes the empirical justification, or motivation, for what has been said in the first two parts of the dissertation.

It should not be forgotten that these studies, even though presented as the last part of my dissertation, were both logically and chronologically prior to the preceding parts. Therefore, I have refrained from changing the original texts too much in the light of my posterior systematization or gain of insight. The chapters making up this part should be read as a pre-theoretic description of the result of several empirical studies that actually shaped my understanding as expressed in the previous parts. In order to integrate them coherently in the text, I have sometimes included a discussion section at the end of the chapter, commenting on the studies in the light of the whole dissertation.

In Chapter 9, I outline the basic (and initial) empirical studies: corpus processing and general quantitative information gathering, which paved the way for the detailed investigations that followed, described in Chapters 10 through 14.

Chapter 10, by far the lengthiest chapter of this part, describes a thorough study of Imperfeito, monolingual and contrastive.

Chapter 11 looks into perception verbs in English and Portuguese and their relationship with aspect.

Finally, Chapters 12 to 14 deal with grammatical aspect. Starting from a clear category in one language, I studied its translation into the other, concluding about its presence and role in the target language:

In Chapter 12, I start from the perfect category in English and study its possible counterparts in Portuguese; in Chapter 13, I look at the apparently similar pluperfects in the two languages; and in Chapter 14, I investigate the Imperfeito/Perfeito opposition in Portuguese by looking at its possible counterparts in English.

This way, I have attempted to cover some of the most general issues of aspect, namely some of those that have been claimed to be universal: the imperfective/perfective opposition and the perfect category.

323

324

Chapter 9: Preliminary studies

This chapter has a double status. It describes the point of departure for all that is treated in this thesis, namely, the preliminary exploratory studies done with the corpus. But it is also the repository of the general quantitative data in a condensed format.

The chapter is divided into four very different sections. Section 9.1 provides a detailed description of the processing involved. Its main goal is to make further similar work easier, freeing researchers from detailed consideration of alternative options, as well as warning them about particular shortcomings diagnosed. Section 9.2 provides a very detailed quantitative description of the tenses occurring in the corpus, in terms of both their translation and distribution. Then, Section 9.3 presents my initial conclusions, which are mainly of historical interest. Nevertheless, they provide motivation for the subsequent sequence of studies performed, as well as illustrate the inadequacy of pure quantitative studies, as is discussed in Section 9.4, which puts the whole chapter into perspective.

9.1 Preparing the data

The preprocessing of the texts included the following tasks: (i) scanning and subsequent proof-reading, as well as general corrections (undoing hyphenation because of line breaks, removing page numbers and special formatting, etc.), (ii) sentence separation, and (iii) sentence alignment.

Information about the basic numbers involved (number of sentences as well as of tensed clauses) is given in Tables 9.1 and 9.2. The exact meaning of "transl(ation) pair" and "tensed transl(ation)" will be elucidated below.

Table 9.1

	Words in English	Words in Portuguese	Sentences (Eng)	Sentences (Port)	Transl. pairs	Tensed transl
EP1	3416	3051	211	224	207	535
EP2	2233	2017	115	127	113	248
EP3	5815	5172	349	410	340	782
EP4	5051	4554	347	393	337	701
EP5	2827	2472	194	235	191	396
EP6	6718	5996	412	472	412	1082
Total	26060	23262	1628	1861	1602	**3744**

325

Table 9.2

	Words in Portuguese	Words in English	Sentences (Port)	Sentences (Eng)	Transl pairs	Tensed transl
PE10	4410	4898	323	324	322	596
PE11	1501	1719	70	71	70	217
PE12	4944	5425	179	181	179	728
PE3	3447	3695	210	211	210	456
PE6	4460	5258	352	355	351	595
PE8	2019	2279	107	107	106	239
PE9	4393	4698	208	210	208	487
Total	25174	27972	1449	1459	1446	**3318**

Let me present first some data on the alignment process, before discussing in detail the annotation and the process of extracting quantitative data.

9.1.1 Alignment

It is well known that not all sentences in an aligned text have exactly one and only one sentence counterpart. The next "translation pair", for instance, aligns 2 sentences with 2 sentences:

> *Only his eyes searched in the darkness, and in the pale light of the moon that crept through the holes in the brush house Kino saw Juana arise silently from beside him. He saw her move toward the fireplace.*
> *Só os olhos vogaram na escuridão. E, à vaga luz da Lua que se infiltrava pelas fendas da cabana, Kino viu, ao seu lado, Joana levantar-se sem ruído e aproximar-se da fornalha.*

Alignment was done manually through the visual display of the texts (one sentence per line, lines numbered) in parallel windows. Whenever there was a situation different from 1-to-1, the texts were (manually) edited, the line breaks in question deleted, and line numbers were (automatically) reassigned.

I should note that the result of alignment was actually a rearranging of either monolingual text (different "sentence" separation, in order to reflect the pairing of the two texts). The texts in each language were kept separate, in order to be processed (searched for, for example) by NLP tools for each language.

A detailed quantitative description of the misalignments can be found in Santos (1994a). Here, I will only state that the most common source for misalignment was the different conventions of direct speech encoding in English and Portuguese. While English as a rule writes direct speech between quotes, separating it from the rest of the text by a comma, Portuguese uses the "travessão" (the long dash, "--"), and a dot to mark the end of the direct speech. So, the Portuguese text in general contained one extra sentence break. In addition, it seems that the placement of the reporting verb itself obeys different rules in the two languages, and hence the

common rearranging of the order of the clauses as well. This situation is illustrated by the following examples:

> *"I go to inform myself," and he closed the gate and slid the bolt home.*
> *-- Vou saber. Fechou a porta e correu o trinco.*
>
> *She put her hand on Coyotito's covered head. "Open it," she said softly.*
> *Pôs a mão na cabeça coberta de Coyotito e disse baixinho: --Abre-a.*

9.1.2 Annotation

9.1.2.1 Introduction

The heaviest corpus task was obviously the (manual) annotation of the tense translations occurring in the corpus. The resulting annotation was also kept in a separate file, later the input for the counting programs. The bridge between the two aligned files and the annotation file is simply the corresponding line numbers.

One of my first interesting findings was the realization that clause "alignment" was actually more complex -- and less parallel -- than sentence alignment. There is, in fact, a far from perfect agreement as far as (finite) clauses are considered. In the pairs below, one can see 1-2, 1-0 and 2-1 clause translations: the first two, in the very same translation pair, were underlined, and put inside square brackets, respectively, to increase perspicuity. Verb forms are in bold face:

> *He turned away from her and **walked** up the beach and through the brush line.*
> *[His senses **were dulled** by his emotion.]*
> *Voltou as costas à mulher, **subiu** pelo areal, **atravessou** as sebes, [com os sentidos **embotados** pela emoção].*
>
> *Her back **was** bent with pain and her head **was** low.*
> ***Levava** as costas curvadas pela dor e a cabeça caída.*

Furthermore, clause order is not necessarily preserved, even if clause number and content remains unchanged. For example, in the next example ("C" stands for clause), C1 C2 C3 in English is rendered as TC1 TC3 TC2 in Portuguese (C2 and TC2 are underlined):

> *His senses were burningly alive, but his mind went back to the deep participation with all things, the gift he had from his people.*
> *Os seus sentidos ardiam, vivos, mas, com aquele dom que os antepassados lhe tinham transmitido, regressava à íntima comunhão com todas as coisas.*

The reader may consult Santos (1994a) for a detailed description and classification of the cases of clause misalignment.

9.1.2.2 In detail

Note that I use "tense" for tense forms which can be simple or complex: for example, present progressive is considered a tense distinct from simple present. Also, I generally omit the word "simple". The annotations were stored in the following format: For each "translation pair" (resulting from sentence alignment), I recorded, first, the "translation pair" number, then, the tense transfer, preceded by the number of times if it appeared consecutively, followed by "x"

(reading "times"). The tense transfer consisted in a mnemonic for the source tense, a dash, and a mnemonic for the target tense.[1] If the mnemonic was shared by the two languages, it was written only once (so **pres**, instead of **pres-pres**). Tense transfers in each translation pair were separated by semicolons. In case the main source verb was *be*, *ser* or *estar* (the two usual translations of *be*), that verb preceded the specification of the tense transfer (e.g. **3 x be PS- I)**.

Since I was specifically interested in the translation of tenses, my annotation only concerned every source tensed clause whose translation gave rise to another clause, tensed or not, in the source text order. So, for example, the last translation pair mentioned was annotated thus:

4.273 be PS - I; PS - I; PS - MQP;

Untensed source clauses were not taken into acount in the annotation, as can be seen in:

> *Há sempre quem suponha, na sua paixão, que destruir Roma, a devassa Roma, a pecadora Roma, é **dar** testemunho dos desígnios de Deus.*
> *There are always those who, in their zeal, believe that by destroying Rome, that debauched Rome, that sinful Rome, they are carrying out the designs of God.*

10.208 pres; PC - pres;

Neither were those tensed clauses which had no clausal equivalent in the target language considered, as shown in:

> *Não era uma tentação que repelia assim; mas era, como bem sabia, um esforço para que o céu se **contentasse** com as relações espirituais de uma oração.*
> *It was not a temptation that she repelled in this way; but it was, as she well knew, an effort to **satisfy** the heavens with the spiritual offering of a prayer.*

11.20 ser I - PS; I - PS; ser I - PS; I - PS;

Tensed clause creation in the target language was not annotated, either, as is shown in:

> *-- Que foi que ele revelou antes de **desfalecer**?*
> *"What did he reveal before he **lost** consciousness?"*

10.110 P-PS;

In cases of a sequence of several verbs in the same tense that get translated into another sequence of another same tense, I only counted the effective number of pairings (i.e., the smallest of the two sequence sizes), and this is why the above pair, featuring two tensed clauses in each language, only gets annotated with one translation.

Finally, it should be noted that aspectualizers were not counted as different "tenses" in the source language, but they were regarded of interest if they were involved only in the translation. I.e., if *He began to do* were translated by *Ele começou a fazer*, only the translation "simple past to Perfeito" (**PS - P**) would be annotated, but if *He ran* were rendered in Portuguese by *Ele começou a correr*, then the translation "past simple to começar a Perfeito" (**PS - começar P**) would be stored. One real example:

> *Foi uma surpresa esquisita que a **percorreu** trémula da cabeça aos pés.*

[1] The mnemonics I display in this section are explained here: **PS** stands for simple past; **I** for Imperfeito; **MQP** for Mais que perfeito; **PC** for Presente do conjuntivo; **P** for Perfeito; **pres** for both simple present and Presente.

*A strange amazement **went quavering** through her from head to toe.*

11.58 P - went ger;

On the other hand, no similar annotation was performed if aspectualizers were absorbed by the target language, i.e., source aspectualizers were not in general given special annotation.

9.1.3 Counting

The counting programs simply processed the files created in the previous step and returned either the number of instances of a pre-defined combination or all possible combinations found.

Since the files where the annotation was stored were written in a simple fixed format, the programs, written in AWK, were not more than a combination of pattern matching and incrementation of counters.

Their output corresponded more or less directly to the numbers that will be presented in tabular form in the next section.

In addition, a list of translation pairs' numbers was computed for each particular phenomenon, in case I wanted to look at the actual cases. Even though this could be transformed easily into a very simple translation browser indexed by tense translations, I never actually implemented it. When I did look into every translation case of a particular phenomenon, for example, I was more often than not interested in a more specific phenomenon than that corresponding to one particular tense translation, involving specific lexical items or the like. So, I used the list of translation pairs only occasionally, while often employing GREP's of the text files, following the general approach of using already existing tools (mainly UNIX tools) as much as possible.

9.2 Quantitative description of the corpus

The counting programs were first made to count instances of translations, i.e., given a particular source tense, how did its translations distribute with respect to the target tenses. (The word "tense" taken in the broad sense described above.) This was actually what I had in mind when annotating the corpus, and the result is displayed in Sections 9.2.1 and 9.2.2.

Later, I made the counting programs count the "inverse" function, i.e., given a particular target language tense, what was the distribution of its origins. The result is displayed in Sections 9.2.3 and 9.2.4, where also some shortcomings of the annotation process are discussed.

9.2.1 The translation of English tenses into Portuguese

Tables 9.3 to 9.7 show the most common translations of each of the major tense forms. Less frequent forms and their translations were not thought worth while including here, even though they will often be discussed in the following chapters.

Table 9.3 lists the translations of 2,305 English simple past occurrences, 332 of which involve the verb *be* (which, as pointed out above, was the sole non-tense factor taken into account in the annotation procedure, namely, the occurrence of *be* as source main verb, or,

conversely, of *ser*, *estar* or *haver* as source main verbs).

330

Table 9.3

	EP1	EP2	EP3	EP4	EP5	EP6	Total
SIMPLE PAST	323	185	499	412	242	644	2305
Imperfeito	125	85	175	179	89	260	913
Perfeito	161	75	266	192	123	318	1135
Infinitivo	5	4	10	9	9	20	57
Gerúndio	5	2	10	3	6	9	35
Imperfeito conjun.	3		3	10		10	26
Mais que perfeito	6	5	11	10	4	3	39
Condicional	1		4	2	4	3	14
Presente		5	5			4	14
Presente conjuntivo		1			1	3	5
Part. pass.	4	3	3	5		3	18
ir + Gerúndio	2	2	4	1	2	1	12
ficou + Part. p./adj			1	1		2	4
ficou a + Infinitivo			2			2	4
pôs-se a + Infinitivo	2	2		1		5	10
estar + Part. pass.	2		1	1	1	1	6

Table 9.4 contains the translations of the 58 past progressive occurrences, Table 9.5 concerns the 50 present perfects and Table 9.6 contains the simple present data, corresponding to 260 occurrences.

Table 9.4

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PAST PROG	8	5	19	4	8	14	58
Imperfeito	3	4	12		5	8	32
Imperfeito prog.	1	1	2	3	2	4	13
Gerúndio			2			1	3
Mais que perfeito	2				1		3
adjectivo						1	1
Presente				1			1
ir + Gerúndio	2		1				3
estar + Part.pass.			1				1
Infinitivo			1				1

331

Table 9.5

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PRESENT PERFECT	3	0	18	14	10	3	48
Perfeito	2		13	13	7	1	26
PPC				1			1
Presente			3		1		4
passiva	1						1
Mais que perfeito			1				1
Futuro conjuntivo			1			1	2
Presente conjuntivo					1	1	2
Infinitivo					1		1

Table 9.6

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PRESENT	18	8	59	98	33	44	260
Presente	15	6	49	88	29	30	217
Futuro do conjuntivo						8	8
Imperfeito	1	2	3			4	10
Perfeito			1	1	1		3
Futuro	1			3		1	5
Presente conjuntivo				5	1	1	7
Imperfeito conjuntivo					1		1
Infinitivo			1	1			2
Condicional	1						1
ficou + adj					1		1

Finally, Table 9.7 displays the case of the translations of the English pluperfect, adding up to 142 cases.

332

Table 9.7

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PLUPERFECT	22	18	32	25	8	37	142
Mais que perfeito	17	17	19	13	3	28	97
Imperfeito	2		5	7	2	1	17
Perfeito	1		3	3		6	13
Imperfeito conj.	2		2	1	1	1	7
Infinitivo			3	1			4
Condicional					2		2
Presente		1				1	2

9.2.2 The translation of Portuguese tenses into English

Conversely, the following numbers were obtained concerning Portuguese as source language: Table 9.8 presents the translation of 1,102 Imperfeitos, Table 9.9 describes 769 Perfeitos, Table 9.10 concerns the 433 cases of Presente, Table 9.11 the 3 of Pretérito Perfeito Composto (PPC), and Table 9.12 the 353 cases of Mais que perfeito.

Table 9.8

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
IMPERFEITO	127	109	214	155	198	99	200	1102
simple past	101	91	147	120	136	64	156	623
past progressive	11	5	14	8	17	17	10	82
gerund	5	3	2	2	5	3	6	26
could	3	2	8	4	4	4	7	32
conditional	1	1	22	8	15	1	4	53
used to	2		9	1		2		13
passive	1		2	1	5	4	4	16
pluperfect			3	1	5	3	4	16
infinitive			1	2				3

333

Table 9.9

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PERFEITO	229	37	132	109	110	79	74	770
simple past	201	35	115	100	100	70	69	690
present perfect	13		8	3	5	2	2	33
present	4		1			1		6
passive	2	1	2	2	1	2		10
gerund	2						1	3
could				1	3		2	6
pluperfect			2	1	3	3		9
conditional	2							2

Table 9.10

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PRESENTE	167	1	99	60	65	29	12	433
present	141	1	76	49	50	23	8	348
present passive	4		5	2	2	1	2	16
future	1		2	6	1	1		11
gerund	5		1	2	2			10
present perfect	3		3					6
present progr.	3		1		2			6
conditional	3		1			1		5
must	2		2				1	5
future progr.	2							2

Table 9.11

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PPC	1	0	2	0	0	0	0	3
present perfect prog.	1							1
present perfect			2					2

334

Table 9.12

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
MAIS QUE PERFEITO	25	24	48	57	102	7	90	353
pluperfect	19	23	43	51	93	7	61	297
simple past	3	1	4	4	5		24	41
passive	2			2	1		2	7
pluperfect prog.					1		1	2
present perfect	1							1
used to			1					1
could							1	1
gerund							1	1
NP					1			1
adjective					1			1

The construction of these tables, however, was not so straightforward as it may appear. A simple example may suffice to show that options must sometimes be taken, which are by no means evident. For instance, Table 9.12 does not include modals in Mais que perfeito, i.e., it corresponds to all Mais que perfeitos but those of Portuguese modal verbs. This is due to the fact that an English modal cannot be in the pluperfect, i.e., *pudera* must translate into *could have*. This latter form (which could also be a literal translation of Portuguese *podia ter*) would be more accurately described by the simple past (of *can*) followed by a perfect infinitive. Still, counting it as pluperfect or as simple past would distort the distribution.

9.2.3 English tenses in translated text according to Portuguese origin

As pointed out above, the other kind of counting (actually only found of interest at a much later stage) was the original of a particular translation, i.e., the inverse function from the translation arrow. For example, given *x* instances of Imperfeito in a translated text, which particular tenses do they translate? Contrary to the previous results, which correspond (roughly[2]) to the total number of tense occurrences in the source text, these numbers only reflect tensed translations, given that no cases of tensed clause creation were annotated. This explains why neither non-finite tenses nor non-verbal (noun, prepositional) phrases appear in Tables 9.13 to 9.24.

Table 9.13 shows which Portuguese original tenses were translated into simple past. Tables 9.14 to 9.18 indicate which ones gave rise to past progressive, simple present, present perfect, pluperfect, and *could*, respectively.

[2] The annotation regarding tense disappearance was not consistent. Especially in cases of clause disappearance in a sequence of same tenses in the source text, no trace of it remained in the annotation.

Table 9.13

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
SIMPLE PAST	317	126	281	234	256	137	284	1635
Perfeito	205	33	118	104	108	70	92	730
Imperfeito	102	91	152	118	136	64	155	818
Mais que perfeito	3	1	3	4	5		24	40
Presente	1		2		1			4
Imperf. Conj.	5	1	5	4	4	2	9	30
Condicional				1			2	3
ir + gerúndio Imp.			1	2			1	4
ir + gerúndio Perf.	1				1			2
vir + gerúndio Imp.				1				1
vir + gerúndio Perf.					1		1	2
progressiva						1		1

I should note that, since modals were annotated separately, they are not included under the label "simple past", but were only counted as *could*, or *must*. Still, no careful treatment of this verb class was done during the annotation, and thus semi-modals such as *be able to* or *have to* are in fact included in these numbers.

Table 9.14

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PAST PROGRESSIVE	13	5	18	11	24	18	13	102
Imperfeito	11	5	14	8	17	17	10	82
Imp. progressivo			2	2	1		2	7
ir + Gerúndio					1			1
vir + Gerúndio					2		1	3
Perfeito	1					1		2
Presente			1		2			3
Mais que perfeito			1		1			2
dever estar Imp.	1							1
estar + Part.pass.				1				1

336

Table 9.15

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
SIMPLE PRESENT	152	1	135	53	59	27	12	439
Presente	141	1	125	50	56	23	8	404
Perfeito	4		1			1		6
Presente Conj.	3		4	1	1	1	1	11
Futuro Conjuntivo	3		2		1		1	7
Imperfeito			1		1		1	3
Imperfeito Conj.	1		1	2				4
Futuro	1						1	2
poder + Inf						1		1
Presente prog.			1			1		2

Table 9.16

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PRESENT PERFECT	23		15	5	5	2	2	52
Perfeito	15		8	3	5	2	2	35
Presente	6		3					9
PPC	1		2					3
haver + Part.pass.				2				2
Imperfeito			1					1
Presente conj.			1					1
Mais que perfeito	1							1

Note also that in Table 9.16 under Perfeito are included the cases of the aspectualizer *acabar de*, which may get translated into present perfect plus *just*. Also, some of the cases of Presente, as well as the only one of Imperfeito, are relative to the verb *haver* in its very specific temporal use.

337

Table 9.17

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
PLUPERFECT	20	24	52	57	106	13	68	340
Mais que perfeito	18	23	43	51	93	6	62	296
Imperfeito		1	3	1	5	3	4	17
Perfeito			2	1	3	3		9
Imperfeito conj.			3	1	2		1	7
MQP conjuntivo			2	2	1		1	6
MQP condicional	2							2
Condicional					1			1
passiva (*ser*)					1			1
MQP progressivo						1		1

Table 9.18

	PE10	PE11	PE12	PE3	PE6	PE8	PE9	Total
COULD	8	4	12	11	12	7	20	72
Imperfeito	3	2	8	5	4	4	7	33
Perfeito	1			2	3		2	8
Condicional	1	1		4	2	1	1	10
poder	1	1	4		2	1	7	16
Imperfeito conj.						1	2	3
Presente conj.	2							2
Mais que perfeito							1	1
Futuro					1			1

9.2.4 Portuguese tenses in translated text according to English origin

Table 9.19 discriminates the origin of translations into Imperfeito, Table 9.20 into Perfeito, and so forth. Progressiva is presented in the form of "*ir* Gerúndio + *estar a* Infinitivo". Neither table includes passives (nor *estar* + Particípio passado constructions).

338

Table 9.19

	EP1	EP2	EP3	EP4	EP5	EP6	Total
IMPERFEITO	145	96	211	194	100	295	1041
simple past	126	85	172	179	88	260	910
progressive	4	4	12		5	8	33
could	6	3	6	7	3	14	39
pluperfect	2		5	7	2	1	17
conditional	1					5	6
present	1	2	7		1	4	14
passive		2	8	9	1	2	22
might						1	1
should, must			1	1			2
subjunctive				1			1
+ PROGRESSIVA	3+2	2+1	1+2	0+3	1+3	1+4	8+15
progressive	2+1	0+1	1+2	0+3	0+2	0+4	3+13
simple past	1+1	2+0			1+0	1+0	5+1
can					0+1		0+1

Table 9.20

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PERFEITO	173	79	287	212	140	330	1221
simple past	164	76	258	192	122	315	1127
present perfect	2		13	11	7	1	34
passive	3		3	5	8	1	20
could	3	2	7		1	7	20
pluperfect	1		3	3		6	13
present			2	1	2		5
might perf			1				1
+ PROGRESSIVA	1+0		4+0	1+0	1+0	1+0	8+0
simple past	1+0		4+0	1+0	1+0	1+0	8+0

339

Table 9.21

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PRESENTE	15	13	73	100	33	44	278
present	14	6	49	85	28	28	210
future	1		6	5	3	6	21
simple past		5	5	1		4	15
pluperfect		1				1	2
present perfect			3		1		4
progressive			2	1			3
can			3	2	1	2	8
conditional			1	1			2
passive			3			1	4
could		1				2	3
must				2			2
might				1			1
should				2			2
subjunctive			1				1

Table 9.22

	EP1	EP2	EP3	EP4	EP5	EP6	Total
MAIS QUE PERFEITO	29	24	38	26	9	33	159
pluperfect	17	17	19	13	3	28	97
simple past	6	5	11	10	4	3	39
passive	4	2	6	3	1	2	18
progressive	2				1		3
present perfect			1				1
could			1				1

Table 9.23

	EP1	EP2	EP3	EP4	EP5	EP6	Total
PPC			1	1			2
present			1				1
present perfect				1			1

9.2.5 The monolingual distributions

For completeness' sake, I display in Tables 9.24 and 9.25 the distribution of the tenses in the four groups of texts, restricted again by the constraint that the translated texts have only been counted in a subset of the cases, namely, those in which they parallel (or translate) a tensed

340

clause. Less frequent forms are not included.

Table 9.24

Tense Name	Original Portuguese (3318 tenses)		Translated from English	
	No.	%	No.	%
Imperfeito	1102	33%	1041	33%
Perfeito	770	23%	1221	39%
Presente	433	13%	278	9.9%
Mais que perfeito	353	11%	159	5.1%
PPC	3	0.09%	2	0.06%

Table 9.25

Tense Name	Original English (3744 tenses)		Translated from Portuguese	
	No.	%	No.	%
Simple past	2305	62%	1635	45%
Past progressive	58	1.6%	102	2.8%
Present	260	6.9%	439	12%
Pluperfect	142	3.8%	340	9.4%
Present perfect	48	1.3%	52	1.4%

The total number of translated tenses was computed by subtracting from the number of tenses in the source text all instances of translation into non-tensed clauses or phrases, amounting to 139 for English coming from Portuguese and to 190 for Portuguese coming from English.

9.3 Initial results

After computing the distribution and occurrence of tense translations (displayed in Sections 9.2.1-2 above), my first study consisted in looking for possible relationships among the tense and aspect markers of the two languages. Somewhat surprisingly, this study led to a few conclusions unrelated to its main aim, to which I shall turn now.

9.3.1 Clause misalignment

One result of the whole process of preparing the corpus concerned the problem of automatic alignment, and clause alignment in particular. Having performed (manually) the alignment, I concluded that the use of tense clues alone would not improve the accuracy of an automatic alignment task, given the several distinct types of misalignment that can occur. Or, to put it more clearly, the number of tensed verbs in the two languages is not a reliable indicator for sentence alignment.

This means that robust clause alignment does not seem to be feasible using tense indicators

in the two languages (in addition to the common sentence length metrics, in words or characters). In any case, this point is only of importance if one expected to perform alignment based on monolingual analyses, that is, having an annotated English and an annotated Portuguese text but no specific contrastive resources.

9.3.2 Mistranslations

I have also realized with dismay that real mistranslations are astonishingly frequent. This was a startling observation, considering that human (and furthermore literary) translation is usually presented as the unattainable model for machine translation.

In addition, in translations from English to Portuguese (my native language), it was frequently the case that I disagreed with the translator's choices.

Even though I did not attempt to quantify these impressions, a careful reading of the several examples displayed in the next five chapters will allow the reader to assess for him/herself this state of affairs.

9.3.3 Automatic detection of translation problems

As far as possible automatic detection of translation problems is concerned, I also concluded that it cannot be based simply on statistical regularities of tense transfers.

On the one hand, it is clear that superficial mismatches do not necessarily mean mistranslations, as can be checked for instance in the examples presented in Section 9.1 above. On the other hand, I have been able to find several cases where the tense translation was common although I believed the translation to be incorrect. Cf. the following translation from simple past into Imperfeito, where Perfeito should have been chosen:

> *His senses were coming back and he **moaned**: "They have taken the pearl..."*
> *Ele estava a voltar a si, **gemia**.--Roubaram a pérola...*
> 'He was coming back to himself, moaning: "They stole the pearl..."'

My conclusion is that it is not possible to build a translation checker program working on unrestricted aligned text in the two languages on the basis of tense translations. If such a program is possible at all, it would at least have to establish correspondences between tense and lexical items.

It might be possible to compile automatically some complex bilingual dictionary entries by looking into the most complex and rare translation pairs and study whether they are due to lexical matters. But this was not, obviously, my concern in this thesis.

9.3.4 Quantitative conclusions

The most relevant conclusions of this study should concern its goal, i.e., the study should allow one to extract some interesting facts from the quantitative assessment obtained. Based on the data in Sections 9.2.1 and 9.2.2, I made the following preliminary remarks (published in Santos (1994a)), concerning the distribution of tenses in narrative discourse in the two languages (items 1 to 3 below), and the distribution of tense translations into the other language (items 4 to

5):

1. The low frequency of the the progressive in English is most surprising, even though it is known that it is more frequent in speech than in writing. This casts doubt on the traditional (Kamp, 1981) DRT analysis of progressive filling the same role as French imparfait (by itself and even more if we assume that Imperfeito is similar, in that respect, to French imparfait).

2. It seems that the distinction between foreground and background in English is not given primarily by tense, since the vast majority of clauses (62%) were in the simple past. This supports Couper-Kuhlen's contention (1987:24) that "for the organization of temporal relations in narration the contribution of the Past tense (as opposed to some other tense) is minimal".

3. Couper-Kuhlen's (1987, 1989) description of relative clauses as "not advancing narrative" seems also to be supported by this study: in all relative clauses with a relative pronoun in EP1 and EP2, 39, only one (2.6 %) was translated by a Perfeito. Strictly speaking, however, this observation does not follow directly from the study described above, because it adds another variable, namely, the use of the tense in a relative clause.

4. Stativity in English seems to strongly favour Imperfeito in the translation, if one considers that the main verb *be* is a clear indication of stativity. On the other hand, *be* (in the simple past) was translated by Imperfeito only in 80% of the cases.

5. The size and the non-balancing of the texts used does not allow the generalization as to specific probabilities of translation between English and Portuguese in general. However, given that the number of source tense forms is fairly small, it seems right to assume that some sort of Zipfian law holds for tense transfers: Given enough data, the probability of the strangest tense translation is greater than zero. In other words, if more and more narrative text were annotated, one would get more and more instances of the translation pairs already found, in roughly the same proportion, while some new (and rare) combinations would still be found.

This scarcity of non-trivial quantitative discoveries entails, in fact, what I consider the most important conclusion of this preliminary study: The data gathered was too coarse to allow significant results.

Finer-grained studies had to be conducted, as is evident if you consider the following example: What is the significance of the observation that the simple past was translated by Imperfeito 40% of the times, if Imperfeito has several distinct semantic values, and so does the simple past? Now, if, for example, the uses of Imperfeito signalling habituality in Portuguese were marked, one could study the translation of such a value into or from English, and possibly even conclude something about habituality in the latter language.

The recognition of this issue motivated a more fine-grained study, focusing on the meaning of Imperfeito, and deepening significantly the annotation (from objective grammatical markers to semantic interpretations), which will be the subject of Chapter 10 below.

9.4 Discussion

While the initial results presented in Section 9.3 are either self-contained or of little

potential for further investigation (the comments there can be seen as the pre-history of this dissertation), this is the place for a brief wrapping up of issues that could be improved if a new study would start now.

9.4.1 On the annotation

As regards the annotation, it is now clear that I lost some information by annotating the texts with a clear source language bias. The most visible defect is the loss of some interesting information on the side of the translated text, i.e., tenses in the translated texts were less thoroughly described. Furthermore, several features of the translation pairing were not described, such as order changes or differences in embedding.

What I missed most was, however, information of an altogether different sort, namely, information on direct speech. During all subsequent work on the texts, it became more and more evident that tense and aspect in direct speech is significantly different from the other contexts.

Conversely, this work demonstrated once more that one cannot know at the outset exactly what are the variables that will be of most importance (or which will be addressed in the course of the study). In other words, there is no such thing as general empirical value outside theory. So, one kind of information whose study was catered for in the initial annotation scheme but which was never put into practice was tense sequencing. I.e., one could study tense switches or tense continuation and their possible implications for translation, because they were annotated differently. This remains material for future work, therefore.

Finally, some good properties of the annotation scheme are also worth emphasizing. For example, keeping the texts in each language separate and the annotation in turn separate from the texts proved to be an invaluable choice. In fact, addition of yet other files with other kinds of information would be possible, in order to study correlations with other factors. (For example, I would be able to include information on direct speech, time adverbials, or kind of subject, in separate files indexed by translation pair (line number).) Besides, I could (and did) process each text separately without worrying about the other language or the annotation.

9.4.2 On quantitative information

Regarding the quantitative assessment, let me note that, in addition to the finer studies focussing on particular grammatical markers/tenses, which will be the subject of the next few chapters, there are obviously other measures of interest, related to the texts in the two languages, some of which I present in Table 9.23.

Table 9. 23

	Original English	Original Portuguese	Translated English	Translated Portuguese
Average clause length (in words)	6.9	7.8	8.7	6.2
Average sentence length (in tensed clauses)	2.63	2.29	-	-
Average sentence length (in words)	16.01	18.06	19.17	12.05

As expected, the number of words per clause and per sentence is higher in Portuguese than in English. Strictly speaking, however, what clause length stands for is the average interval in words between two different occurrences of tensed verbs, since non-finite clauses were not counted. It was surprising, however, to observe that the size of translated Portuguese was so much lower than the English original it translated, especially because not all tensed clauses were counted (i.e., the value 6.2 is larger than reality).

Two further parameters should be relevant, namely the type/token ratio and the number of different verbs. The consideration of the parameter "verbal diversity" in the two languages would allow one to confirm the possible expectation that a language with more grammatical distinctions in the verb morphology would be less rich lexically. To compute it (approximately), however, the texts must be lemmatized and the different lemmata for verbs counted, and I did not get access to an English lemmatizer in time to present the values here.

The main conclusion of the study reported in the present chapter is, however, that general quantitative results had always to be supplemented with a fine analysis of the respective instances, which convinced me that quantitative data on their own are rarely interesting without, again, a general theory which explains particular cases. I thus considered such a detailed investigation of part-of-speech distribution work for a future occasion.

345