'Suggesting rather special facts': a corpus-based study of distinctive lexical distributions in translated texts

Ana Frankenberg-Garcia[1]

Abstract

It is well-known that translated texts read differently from texts that have been written without the constraints imposed by source texts from another language. One of the features that can confer a distinctive feel to translations is the frequency with which certain lexical items are represented in them. Previous research has compared the frequency of specific words in translations and in texts that are not translations, and unveiled substantial differences in their distributions. Most of these studies adopt a bottom-up approach. Their starting point is a given word whose frequency in translated and non-translated texts is then compared. In this study, I adopt an explorative, top-down approach instead. I begin with a Portuguese language corpus of translated and non-translated literary texts, and attempt to identify lemmas which are markedly over- and under-represented in the translations. Our results not only appear to support existing bottom-up intuitions regarding distinctive lexical distributions, but also disclose a number of unexpected contrasts that would not have been discernible without recourse to corpora.

1. Introduction

One of the great advantages of corpus analyses is that they allow us to discover linguistic facts that are not readily visible to the naked eye. As predicted by Baker (1993), over the last decade the use of corpora in translation studies has had a significant impact on the description of the linguistic features of translation. Some characteristics, however, have received much wider attention than others. The phenomenon of explicitation, for example, whereby information that is only implicit in the source text is believed to be made more explicit in the target text (Vinay and Darbelnet, 1958), has been corroborated by a number of quantitative, corpus-based

[1] Instituto Superior de Línguas e Administração, Quinta do Bom Nome, Estrada da Correia 53, 1500-210 Lisboa, Portugal.
 Correspondence to: Ana Frankenberg-Garcia, *e-mail*: ana.frankenberg@gmail.com

analyses (e.g., Øverås, 1998; Olohan and Baker, 2000; Pápai, 2004; and Frankenberg-Garcia, 2009).

A feature which has received much less attention in the literature is the distinctive distribution of lexical items in translated and non-translated texts. In one of the few studies available, Shama'a (1978, cited by Baker, 1993) found that the words *day* and *say* could be twice as frequent in English translated from Arabic than in original English texts, making the English translations read differently and contributing to the identification of those texts as translations. In a more recent corpus-based study, Frankenberg-Garcia (2004) notes that the English adverb *already* was found to be almost twice as frequent in English translated from the Portuguese than in original English texts. A possible explanation for this is that the Portuguese equivalent *já* often has to be used in contexts where *already* is not required, because the use of the English present perfect compensates for the use of *já*. For example, in a sentence like 'Já terminaste?', which translates as 'Have you finished?', it is not necessary to add *already* to convey the perfective aspect meaning of the Portuguese *já*.[2]

But the over-representation of certain lexical items in translations is not necessarily the rule. In another corpus-based study, Tirkkonen-Condit (2004) focusses her analysis on typically Finnish verbs of sufficiency, and she notices that they are markedly *less* frequent in translations than in texts originally written in Finnish.

Both over- and under-represented lexical items can affect the impression a translated text makes on readers – that is, whether or not it reads like a translation. In an earlier study, Tirkkonen-Condit (2002) asked native Finnish speakers to decide whether a selection of text extracts were originals or translations and found that the subjects appeared to base both their correct and incorrect judgments regarding what they thought were originals mostly on the high frequency of certain typically Finnish words. As Baker (1993: 245) suggests, the unusual distribution of certain lexical items in translated texts could be, 'a result of the confrontation of the source and target codes' and a symptom of what is sometimes referred to as 'the third code', although, in poor quality translations, this could also be a sign of the phenomenon of 'translationese' (p. 249).

It is not always very easy to identify which words might be over- or under-represented in translations. One way of doing so is to adopt a bottom-up approach, taking a particular lexical item as a starting point and subsequently comparing its distribution in translated and non-translated texts. To do this, however, we need to make informed decisions at the outset on which lexical items are worthy of such a comparison. In Tirkkonen-Condit's (2004) study, the words tested for under-representation were ones which lacked linguistic counterparts in the source language underlying the Finnish translations. Using this same approach, Frankenberg-Garcia (2007) reported

---

[2] In fact, the addition of *already* would carry the extra meaning that the action in question (having finished) took place earlier than expected.

that the English verb *nod*, with no single-word equivalent in Portuguese, was substantially more frequent in a corpus of original English texts than in English translated from the Portuguese. There are other words, however, that do have straightforward equivalents in the original and translation languages, but whose distribution in translated and non-translated texts is nevertheless distinct. The over-representation of the words *day* and *say* in translated English described in Shama'a's (1978) study is a case in point, similar to Frankenberg-Garcia's (2004) findings with regard to the over-representation of *already* in translated English.

Translators and foreign language teachers, who are continually exposed to the crosslinguistic effects of languages in contact, are sometimes intuitively able to identify lexical items with distinctive frequencies. For example, in a brief, informal discussion carried out prior to this study, a professional Portuguese translator reported that she felt Portuguese adverbs ending in *mente* tended to have an exceptionally high frequency in Portuguese translated from English when compared with texts originally written in Portuguese (Bastos, 2008), and a Brazilian lecturer in Translation Studies commented that the verb *poder* ('can') could be overly frequent in translated Portuguese (Tagnin, 2008). Similarly, according to my own intuitions, the adjectives *diferente* and *possível* ('different' and 'possible') and the adverbs *simplesmente*, *exactamente*, *perfeitamente* and *absolutamente* ('simply', 'exactly', 'perfectly' and 'absolutely') appear to be overly frequent in Portuguese translated from English. However, a bottom-up approach can have limitations, especially if there happen to be words with distinctive distributions that escape our perception.

This corpus-based study is an attempt to investigate distinctive lexical distributions in translated texts from a top-down perspective. Its aim is to confirm existing intuitions about words with distinctive distributions and, at the same time, to find out about exceptionally frequent and infrequent lexis that escape the naked eye.

2. Method

In a bottom-up approach, we would begin by selecting a specific word for analysis and then we would compare its frequency in comparable corpora of translated and non-translated texts. In the explorative top-down approach adopted here, I begin with a corpus of translated and non-translated texts in order to establish groups of words which are markedly over- and under-represented in the translations. The corpus used in the present analysis was COMPARA, a bi-directional, three-billion-words parallel corpus of English and Portuguese literary texts (Frankenberg-Garcia and Santos, 2003).[3] One of the many advantages of bi-directional parallel corpora is that they allow us to compare not only two different languages, but also the translated and

---

[3] COMPARA is available online at: http://www.linguateca.pt/COMPARA/

non-translated subsets of the languages in question. This study took into account thirty-nine different text extracts originally written in Portuguese in the corpus (634,601 words) and thirty-two comparable Portuguese text extracts which had been translated from English (733,282 words).[4] Twenty-two different authors from Portugal, Brazil, Angola and Mozambique, and twenty-five different Portuguese and Brazilian translators are represented in the sample.[5]

The texts had been automatically annotated with the PALAVRAS parser (Bick, 2000) and, at the time of this study, were undergoing manual human revision (Inácio and Santos, 2005). Version 10.0 of COMPARA was used for the present study. Subsequent post-editing corpus improvements may have resulted in slight variations in the total number of words and in the part-of-speech (POS) tags of the corpus, but they should not alter the overall results obtained here in any significant way.

The starting point for the analysis was the distribution of lemmas in COMPARA's sub-corpora of translated and original Portuguese texts (henceforth, 'translated-PT' and 'original-PT'). The lemmas selected for closer examination were all those which had been classified according to the broader part-of-speech categories for nouns (excluding proper nouns), adjectives, verbs and adverbs summarised in Table 1.[6] Grammatical words such as conjunctions and prepositions were not included in the analysis.

After these distributions were obtained, lemmas that failed to reach the threshold of ten occurrences per 100,000 words in at least one of the two sub-corpora under consideration were discarded because I considered them to be insufficiently represented.[7] Lemmas that reached this threshold in one sub-corpus (e.g., original-PT), but not in the other one (e.g., translated-PT) were nevertheless preserved. The sample that passed the pre-established threshold and was thus selected for analysis consisted of 1,003 different lemmas in all, distributed according to the following POS categories:

482 different noun lemmas   113 different adjective lemmas
309 different verb lemmas   99 different adverb lemmas

---

[4] There are three source texts in the corpus that are aligned with two different translations each: EBDL1, EBDL3 and PBJA1. To avoid any distortions caused by counting source texts with multiple alignments twice, only one translation for each was taken into account in this study. In the case of EBDL1 and EBDL3, alignment with Brazilian Portuguese was preferred over alignment with European Portuguese in order to give a better balance in the amount of Brazilian Portuguese represented in the translations. In the case of PBJA1, alignment with an English translation published in 2000 was preferred over alignment with a translation published in 1865, which would have been too different from the other, mostly contemporary translations represented in the corpus.

[5] No distinction was made between the different varieties of Portuguese represented in the corpus; however, as we shall see later, this may have affected some of the results obtained.

[6] See Inácio and Santos (2005) for an in-depth description of these POS categories.

[7] In absolute numbers, this is equivalent to over seventy-three occurrences in translated-PT and over sixty-three occurrences in the slightly smaller original-PT corpus.

POS-category	Types	Tokens	Sub-corpus
ADJ	4,215	33,697	original-PT
	4,118	41,042	translated-PT
V	6,857	130,816	original-PT
	6,256	149,042	translated-PT
N	11,465	137,114	original-PT
	10,517	150,858	translated-PT
ADV	1,025	52,318	original-PT
	1,099	62,136	translated-PT

Table 1: Overall distribution of lemmas per POS category in COMPARA 10.0

Words with alternative spellings were included in the same category because spelling differences were not considered relevant to the study of over- or under-represented words in translation. Thus, spelling differences between Brazilian and European Portuguese such as *direção* and *direcção* ('direction') and other alternative spellings such as *loiro* and *louro* ('blonde') were analysed as one, even though they are treated as separate lemmas in COMPARA. Loan words that are not considered to be part of the Portuguese language, such as the English noun *sir*, were also excluded from the study.[8]

The next step was to calculate the relative frequency per 100,000 words of each of the lemmas above in original-PT and translated-PT in order to compare their differences in frequency, and then determine the amount by which they differed. The lemmas that were at least two times more frequent in translated-PT were regarded as being over-represented in the translations. Conversely, the ones at least two times more frequent in original-PT were considered to be under-represented in the translations. These distinctively over- and under-represented lemmas were then singled out for closer inspection.

As some authors receive more representation than others in COMPARA, a distribution per author was subsequently applied to these lemmas in order to determine whether any of them could have been the product of a distortion caused by a single author. If over one-third of the occurrences of any given lemma could be traced back to just one particular author, then the results for this lemma were considered to be biased and were disregarded. It was not felt necessary to carry out a similar check for translator bias because the Portuguese language translators in COMPARA are fairly evenly distributed. The findings are presented in the next four sections.

---

[8] But see Frankenberg-Garcia (2005) for a detailed study of foreign words in translated and non-translated texts.

3. Distinctive nouns

Of the 482 different noun lemmas that were initially selected for analysis, sixty-nine were found to be over-represented in translated-PT and sixty-eight were under-represented, totalling 137 nouns with distinctive distributions. Of these, forty-six were excluded from further analysis because over one third of their occurrences came from texts by single authors. Table 2a lists the forty-two remaining noun lemmas that were over-represented, and Table 2b lists the forty-nine noun lemmas that were under-represented.[9]

Some interesting trends emerge from the results. Most of the over-represented nouns in Table 2a are abstract nouns. The most conspicuous one is *género/gênero*, which is synonymous with another two lemmas that are markedly more frequent in translated-PT: *espécie* and *tipo*. Several of the nouns in this table also convey the general idea of manner (e.g., *tom*, *modo*, *expressão*, *aspecto* and *atitude*), and quite a few of them are used to classify and group things together (e.g., *membro*, *grupo*, *lista* and *maioria*). By contrast with this, most of the under-represented nouns in Table 2b refer to human beings. Not surprisingly, there are also several nouns in this list that are closely associated with the Portuguese psyche: *lembrança*, *saudade*, *alma* and *tristeza*.

It is also notable that there are a number of near synonyms at opposite ends of the distribution: *rapariga* (over-represented) and *menina* (under-represented), *recordação* (over-represented) and *lembrança* (under-represented), and *escola* (over-represented) and *colégio* (under-represented).[10]

4. Distinctive adjectives

Of the 113 adjective lemmas that were initially selected for analysis, twelve were regarded as over-represented and eleven were regarded as under-represented. Of these, just two were excluded from further analysis because over one-third of their occurrences came from texts by single authors. Table 3a lists the eleven remaining adjective lemmas that were at least two times more frequent in translated-PT, and Table 3b lists the ten remaining adjective lemmas that were at least two times less frequent in translated-PT.

We can see from these results that the most over-represented adjective lemma in translated-PT was *sentado*, and the most under-represented one was *gordo*. Interestingly, most of the adjective lemmas that were at least two times more frequent in translated-PT (with the exception of

---

[9] The English glosses provided in Tables 2 to 10 refer to the most typical translation(s) for each lemma, but do not necessarily correspond to every possible English equivalent found in the parallel alignment.

[10] The distinctive presence of *rapariga* in translated-PT can in part be explained by the fact that the translated-PT corpus contains mostly European Portuguese, where the word is much more common than in Brazilian Portuguese. Not distinguishing between different varieties of Portuguese may have constituted an important intervening variable in the case of lemmas which have very distinct distributions in different varieties of the language.

NOUN lemma	ORIG-PT		TRANS-PT		DIFF
	F	Rel F	F	Rel F	T/O
g(ê|é)nero ('type')	20	3.15	108	14.73	4.7
fa(c)to[b] ('fact')	76	11.98	377	51.41	4.3
plástico ('plastic')	15	2.36	74	10.09	4.3
bocado ('bit')	35	5.52	145	19.77	3.6
membro ('member')	25	3.94	100	13.64	3.5
problema ('problem')	54	8.51	215	29.32	3.4
escola ('school')	44	6.93	168	22.91	3.3
medida ('measure')	31	4.88	118	16.09	3.3
bebida ('drink')	20	3.15	74	10.09	3.2
aspecto ('aspect')	50	7.88	180	24.55	3.1
maioria ('majority')	24	3.78	86	11.73	3.1
início ('beginning')	21	3.31	74	10.09	3.0
altura ('height/stage')	103	16.23	359	48.96	3.0
quadro ('picture')	34	5.36	116	15.82	3.0
emprego ('job')	30	4.73	100	13.64	2.9
possibilidade ('possibility')	30	4.73	95	12.96	2.7
procura ('search')	36	5.67	110	15.00	2.6
recordação ('souvenir')	27	4.25	82	11.18	2.6
casaco ('coat')	29	4.57	87	11.86	2.6
aldeia ('village')	26	4.10	74	10.09	2.5
oportunidade ('opportunity')	35	5.52	99	13.50	2.4
peça ('piece')	52	8.19	145	19.77	2.4
cozinha ('kitchen')	81	12.76	224	30.55	2.4
espécie ('type')	106	16.70	293	39.96	2.4
lista ('list')	28	4.41	77	10.50	2.4
sítio ('place')	57	8.98	148	20.18	2.2
discussão ('discussion')	33	5.20	83	11.32	2.2
semana ('week')	117	18.44	289	39.41	2.1
rapariga ('girl')	122	19.22	301	41.05	2.1
ajuda ('help')	33	5.20	81	11.05	2.1
ombro ('shoulder')	123	19.38	297	40.50	2.1
inglês ('English/Englishman')	39	6.15	94	12.82	2.1
segurança ('security/safety')	34	5.36	81	11.05	2.1
grupo ('group')	84	13.24	200	27.27	2.1
modo ('manner')	159	25.06	376	51.28	2.0
tom ('tone')	88	13.87	207	28.23	2.0
questão ('question')	74	11.66	173	23.59	2.0
tipo ('type')	103	16.23	240	32.73	2.0
dificuldade ('difficulty')	44	6.93	102	13.91	2.0
expressão ('expression')	73	11.50	167	22.77	2.0
atitude ('attitude')	35	5.52	80	10.91	2.0
cortina ('curtain')	37	5.83	84	11.46	2.0

Table 2a: Over-represented noun lemmas in translated-PT[a]

[a]F = raw frequency in the corpus; Rel F = relative frequency/100,000 words; Diff T/O = relative frequency in translated-PT divided by relative frequency in original-PT; and in subsequent tables, Diff O/T= relative frequency in original-PT divided by relative frequency in translated-PT.

[b]Although *fato* can also mean 'suit', only the occurrences in which it means 'fact' are being considered here. It was necessary to distinguish between the two because, unlike in Brazilian Portuguese, in European Portuguese *fato* meaning 'suit' is spelt differently from *facto* meaning 'fact'. Polysemy was not taken into account for the other lemmas analysed in this study.

NOUN lemma	ORIG-PT		TRANS-PT		DIFF
	F	Rel F	F	Rel F	O/T
sobrinho ('nephew')	65	10.24	6	0.82	4.2
lembrança ('souvenir')	84	13.24	10	1.36	4.1
moço ('young man')	79	12.45	8	1.09	4.1
menino ('boy')	206	32.46	34	4.64	3.9
velha ('old woman')	108	17.02	18	2.45	3.7
soldado ('soldier')	110	17.33	24	3.27	3.6
crime ('crime')	154	24.27	36	4.91	3.5
saudade ('nostalgia')	86	13.55	22	3.00	3.3
remédio ('medicine')	67	10.56	16	2.18	3.3
praça ('square')	101	15.92	26	3.55	3.1
prédio ('building')	87	13.71	27	3.68	3.0
fogo ('fire')	124	19.54	37	5.05	3.0
português ('Portuguese')	68	10.72	19	2.59	3.0
cavalo ('horse')	120	18.91	46	6.27	2.9
diabo ('devil')	99	15.60	35	4.77	2.9
menina ('girl')	141	22.22	47	6.41	2.9
o(u|i)ro ('gold')	177	27.89	57	7.77	2.9
prata ('silver')	82	12.92	33	4.50	2.8
arma ('weapon')	108	17.02	42	5.73	2.8
velho ('old man')	278	43.81	107	14.59	2.8
padre ('priest')	201	31.67	82	11.18	2.6
colégio ('school')	67	10.56	27	3.68	2.6
pedra ('stone')	289	45.54	119	16.23	2.5
povo ('people')	93	14.65	41	5.59	2.4
alma ('soul')	273	43.02	114	15.55	2.4
sangue ('blood')	200	31.52	99	13.50	2.3
mistério ('mystery')	89	14.02	44	6.00	2.3
tristeza ('sadness')	67	10.56	33	4.50	2.3
cheiro ('smell')	137	21.59	67	9.14	2.3
senhora ('lady')	320	50.43	154	21.00	2.3
fome ('hunger')	78	12.29	36	4.91	2.3
dono ('owner')	90	14.18	40	5.45	2.3
senhor ('gentleman')	486	76.58	249	33.96	2.2
alto ('top')	71	11.19	36	4.91	2.2
varanda ('veranda')	85	13.39	43	5.86	2.2
unha ('nail')	72	11.35	36	4.91	2.2
sonho ('dream')	182	28.68	98	13.36	2.1
rua ('street')	403	63.50	215	29.32	2.1
graça ('fun/grace')	108	17.02	57	7.77	2.1
primo ('cousin')	65	10.24	34	4.64	2.1
coração ('heart')	289	45.54	149	20.32	2.1
dente ('tooth')	160	25.21	92	12.55	2.0
ordem ('order')	159	25.06	91	12.41	2.0
seda ('silk')	86	13.55	49	6.68	2.0
rei ('king')	67	10.56	38	5.18	2.0
carne ('meat')	106	16.70	59	8.05	2.0
contrário ('opposite')	105	16.55	58	7.91	2.0
boca ('mouth')	296	46.64	162	22.09	2.0
letra ('letter')	87	13.71	47	6.41	2.0

Table 2b: Under-represented noun lemmas in translated-PT

ADJECTIVE lemma	ORIG-PT		TRANS-PT		DIFF
	F	*Rel F*	*F*	*Rel F*	*T/O*
sentado ('seated')	31	4.88	172	23.46	4.8
calmo ('calm')	18	2.84	91	12.41	4.4
maravilhoso ('wonderful')	18	2.84	85	11.59	4.1
evidente ('obvious')	23	3.62	90	12.27	3.4
familiar ('familiar')	22	3.47	83	11.32	3.3
pessoal ('personal')	21	3.31	74	10.09	3.0
especial ('special')	36	5.67	119	16.23	2.9
horrível ('horrible')	26	4.10	82	11.18	2.7
jovem ('young')	54	8.51	147	20.05	2.4
suficiente ('enough')	37	5.83	94	12.82	2.2
principal ('main')	47	7.41	108	14.73	2.0

Table 3a: Over-represented adjective lemmas in translated-PT

ADJECTIVE lemma	ORIG-PT		TRANS-PT		DIFF
	F	*Rel F*	*F*	*Rel F*	*T/O*
gordo ('fat')	73	11.50	26	3.55	3.2
grosso ('thick')	98	15.44	39	5.32	2.9
igual ('equal')	110	17.33	45	6.14	2.8
nu ('naked')	127	20.01	58	7.91	2.5
doce ('sweet')	72	11.35	35	4.77	2.4
raro ('rare')	72	11.35	36	4.91	2.3
triste ('sad')	153	24.11	79	10.77	2.2
rico ('rich')	103	16.23	53	7.23	2.2
alegre ('happy')	71	11.19	42	5.73	2.0
morto ('dead')	71	11.19	41	5.59	2.0

Table 3b: Under-represented adjective lemmas in translated-PT

sentado and *jovem*) are adjectives that reflect personal opinions and feelings more than facts. By contrast, most adjectives that were at least two times *less* frequent seem to focus on an evaluation of reality rather than on personal beliefs.

5. Distinctive verbs

Among the 309 verb lemmas that passed the pre-established frequency threshold, there were thirty-two that were found to be over-represented in translated-PT and nineteen that were considered to be under-represented. No single-author bias was found. However, unlike the nouns and adjective lemmas analysed so far, among the verbs there seems to have been a greater

VERB lemma	ORIG-PT		TRANS-PT		DIFF
	F	Rel F	F	Rel F	T/O
encontrar-se ('find oneself/meet/be')	10	1.58	87	11.86	7.5
acenar ('wave/nod')	18	2.84	101	13.77	4.9
constituir ('constitute')	16	2.52	83	11.32	4.5
inclinar-se ('lean')	19	2.99	92	12.55	4.2
sentir-me ('feel')	32	5.04	135	18.41	3.7
tornar-se ('become')	46	7.25	186	25.37	3.5
replicar ('reply')	24	3.78	95	12.96	3.4
abanar ('shake/rattle')	28	4.41	100	13.64	3.1
sentir-se ('feel')	57	8.98	203	27.68	3.1
comentar ('comment')	37	5.83	128	17.46	3.0
regressar ('return')	55	8.67	168	22.91	2.6
sugerir ('suggest')	35	5.52	104	14.18	2.6
dirigir-se ('turn to')	45	7.09	130	17.73	2.5
preocupar ('worry')	33	5.20	94	12.82	2.5
baixar ('lower')	30	4.73	85	11.59	2.5
virar-se ('turn')	27	4.25	76	10.36	2.4
apanhar ('get/catch/pick/gather')	67	10.56	183	24.96	2.4
compreender ('understand')	115	18.12	311	42.41	2.3
conseguir ('manage/can')	302	47.59	764	104.19	2.2
fazê-lo ('make/do')	82	12.92	207	28.23	2.2
apoiar ('lean')	36	5.67	90	12.27	2.2
manter ('keep')	80	12.61	193	26.32	2.1
lamentar ('regret')	34	5.36	82	11.18	2.1
exclamar ('exclaim')	76	11.98	183	24.96	2.1
provocar ('provoke')	38	5.99	91	12.41	2.1
arranjar ('get/arranje')	77	12.13	183	24.96	2.1
permitir ('allow')	75	11.82	178	24.27	2.1
revelar ('reveal')	49	7.72	116	15.82	2.0
tentar ('try')	225	35.46	525	71.60	2.0
verificar ('check')	34	5.36	78	10.64	2.0
representar ('mean')	48	7.56	110	15.00	2.0
voltar-se ('turn to')	50	7.88	114	15.55	2.0

Table 4a: Over-represented verb lemmas in translated-PT

tendency for over-representation than for under-representation. The verb lemmas were at least two times more frequent, and the ones at least two times less frequent in translated-PT are listed in Tables 4a and 4b respectively.

Tables 4a and 4b show that the morphosyntactic annotation of the COMPARA corpus treats verbs followed by different clitics as separate lemmas. For example, *sentir-se* and *sentir-me* are counted separately. Unlike spelling variations, I decided to preserve the distinction since I found it relevant to this analysis.

Several trends surface in the results above. The most over-represented verb in translated-PT is the link verb *encontrar-se*, and there

VERB lemma	ORIG-PT		TRANS-PT		DIFF
	F	Rel F	F	Rel F	O/T
vencer ('win')	78	12.29	22	3.00	4.1
cuidar ('care for')	134	21.12	44	6.00	3.5
sonhar ('dream')	124	19.54	42	5.73	3.4
morar ('live')	136	21.43	53	7.23	3.0
recolher ('collect/gather')	77	12.13	31	4.23	2.9
fugir ('run away')	230	36.24	91	12.41	2.9
roubar ('steal')	93	14.65	41	5.59	2.6
beijar ('kiss')	69	10.87	33	4.50	2.4
entender ('understand')	262	41.29	131	17.86	2.3
inventar ('invent')	68	10.72	34	4.64	2.3
amar ('love')	168	26.47	89	12.14	2.2
faltar ('miss')	144	22.69	74	10.09	2.2
chorar ('cry')	220	34.67	123	16.77	2.1
quebrar ('break')	68	10.72	38	5.18	2.1
bastar ('suffice')	107	16.86	59	8.05	2.1
conversar ('talk')	171	26.95	93	12.68	2.1
confessar ('confess')	100	15.76	59	8.05	2.0
cantar ('sing')	113	17.81	66	9.00	2.0
cumprir ('meet/deliver')	89	14.02	51	6.96	2.0

Table 4b: Under-represented verb lemmas in translated-PT

are several other link verbs that are at least two times more frequent in the translations: *constituir*, *tornar-se*, *sentir-se*, *sentir-me*, *fazê-lo*, *representar* and *manter*. Two other groups that stand out among the over-represented verb lemmas are the reporting verbs *revelar*, *exclamar*, *lamentar*, *sugerir*, *comentar* and *replicar*, and the verbs used to indicate movement: *inclinar-se*, *regressar*, *dirigir-se*, *baixar*, *virar-se*, *apanhar*, *apoiar*, *voltar-se*, *acenar* and *abanar*. Among the verb lemmas that are at least two times more frequent in translated-PT, we also find verbs that frequently precede other verbs: *tentar*, *conseguir* and *permitir*.

By contrast with the above, most under-represented verb lemmas in translated-PT were highly lexical verbs – often having to do with the dramatic language of literary texts, for example: *vencer*, *fugir*, *beijar*, *cantar*, *quebrar*, *sonhar*, *amar*, *roubar*, *chorar*, *matar*, *morrer* and *nascer*. It was once again possible to find synonymous pairs of lemmas at opposite ends of the distribution, with *compreender* and *apanhar* being over-represented but their respective synonyms *entender* and *recolher* being under-represented in translated-PT.

6. Distinctive adverbs

None of the adverb lemmas analysed in this study had to be excluded from the analysis due to single-author distortions. A total of thirteen out of

ADVERB lemma	Orig-PT		Trans-PT		Diff
	F	Rel F	F	Rel F	T/O
demasiado ('too')	22	3.47	194	26.46	7.6
profundamente ('deeply')	14	2.21	79	10.77	4.9
bastante ('rather/quite')	22	3.47	112	15.27	4.4
claro ('clearly')	73	11.50	320	43.64	3.8
absolutamente ('absolutely')	23	3.62	79	10.77	3.0
completamente ('completely')	48	7.56	161	21.96	2.9
simplesmente ('simply')	36	5.67	120	16.36	2.9
perfeitamente ('prefectly')	41	6.46	127	17.32	2.7
acima ('above')	36	5.67	104	14.18	2.5
imediatamente ('immediately')	54	8.51	148	20.18	2.4
sequer ('not even')	72	11.35	195	26.59	2.3
exa(c)tamente ('exactly')	57	8.98	153	20.87	2.3
através ('through')	91	14.34	230	31.37	2.2

Table 5a: Over-represented adverb lemmas in translated-PT

ADVERB lemma	Orig-PT		Trans-PT		Diff
	F	Rel F	F	Rel F	O/T
enfim ('finally')	169	26.63	35	4.77	5.6
logo ('soon')	567	89.35	215	29.32	3.0
ora ('now/at times')	226	35.61	102	13.91	2.6
ontem ('yesterday')	98	15.44	44	6.00	2.6
jamais ('never')	80	12.61	38	5.18	2.4
amanhã ('tomorrrow')	121	19.07	60	8.18	2.3
porém ('however')	297	46.80	162	22.09	2.1
hoje ('today')	330	52.00	192	26.18	2.0
toda ('entirely')	76	11.98	43	5.86	2.0
todo ('entirely')	101	15.92	57	7.77	2.0

Table 5b: Under-represented adverb lemmas in translated-PT

ninety-nine adverbs were considered to be over-represented and ten out of ninety-nine were regarded as being under-represented in translated-PT. The adverb lemmas that were at least two times more frequent and the ones at that were least two times less frequent in translated-PT are listed in Tables 5a and 5b respectively.

We can see from Table 5a that more than half of the over-represented adverbs end in *mente* and almost all of them are adverbs of manner. Note also that the over-represented adverbs *absolutamente*, *completamente*, *simplesmente*, *perfeitamente*, *imediatamente* and *exa(c)tamente* may be items that are used more or less automatically as dictionary equivalents to phonetically and morphologically similar lexical items in English. By

contrast, none of the under-represented adverb lemmas end in *mente* and most of them are adverbs of time and frequency. Again, there are near synonyms at opposing ends of the distribution, with *todo* and *toda* being under-represented while *completamente* is over-represented.

7. Discussion

Several points of discussion emerge when we look back at the overall results obtained for the four POS categories analysed in this study. To begin with, it is very interesting to note that it was almost exclusively the noun lemmas that were affected by the single-author bias, with forty-six nouns having to be excluded from the analysis. Only two adjective lemmas, and none of the verb and adverb lemmas, were overly influenced by a single author. In fact, the idiosyncratic nouns that authors of literary texts chose to use were not just a question of style and vocabulary preferences. In many cases, biased nouns were dependent on the stories being told. For example, 90 percent of the occurrences the noun *cego* 'blind man' (which had to be excluded from the analysis because it was predominantly used by the Portuguese author, José Saramago), came from a single novel: *Ensaio sobre a Cegueira* ('Blindness').

The noun lemmas also behaved differently from the other POS categories in analysis in that it was the only category for which the number of lemmas considered to be under-represented in translated-PT was greater than the number of lemmas regarded as over-represented. For the adjective and adverb lemmas, the number of over- and under-represented lemmas was very similar. For the verb lemmas, however, the opposite was true: there were substantially more lemmas classified as over-represented than under-represented. These findings seem to be in accordance with the idea that Portuguese is a more nominal language, while English, the source language underlying the translated-PT corpus, is more verbal.

Another interesting finding that emerged was the presence of the near synonyms at opposite ends of the distributions. With the exception of the *rapariga-menina* pair, which, as I explained earlier, is likely to have been a result of an intervening variable, the remaining synonymous pairs could have implications for translator education if we wish to make translators aware that one word is used more typically in non-translated Portuguese literature than the other. Note that the near synonyms discussed earlier are just the ones whose distributions were at least two times more frequent or two times less frequent in translated-PT. If we lower this threshold to less restrictive boundaries, we will find many more near synonyms with contrastive distributions. Table 6 summarises lemmas that were at least 1.5 times more frequent in translated-PT next to near synonyms that were at least 1.5 times more frequent in original-PT. Interestingly, there seems to be a stylistic contrast between some of the near-synonyms listed, in that some of the items that occur in translated-PT sound more formal than their corresponding lemma in original-PT, such as: *recordação*

POS	Lemmas at least 1.5 x more frequent in translated-PT	Synonymous lemmas at least 1.5 x more frequent in original-PT	English gloss
Noun	*rapariga*	*menina*	'girl'
	recordação	*lembrança*	'souvenir'
	escola	*colégio*	'school'
	edifício	*prédio*	'building'
Adjective	*enorme*	*imenso*	'enormous/huge'
Verb	*compreender*	*entender*	'understand'
	recordar	*lembrar*	'remember'
	reparar	*notar*	'notice'
	observar	*examinar*	'observe/examine'
	decidir	*resolver*	'decide'
	obrigar	*mandar*	'force/order'
	manter	*guardar*	'keep'
	apanhar	*recolher*	'pick/gather'
Adverb	*completamente*	*todo/toda*	'completely'
	finalmente	*enfim/afinal*	'finally'

Table 6: Near synonyms with constrastive distributions in original and translated-PT

versus *lembrança*, *edifício* versus *prédio*, *compreender* versus *entender* and *recordar* versus *lembrar*. This could be interpreted as a sign that Portuguese translators attempted to use more formal language than what would be natural in original Portuguese.

The study was also able to shed some light on a number of semantic contrasts. There was a prevalence of over-represented abstract nouns in translated-PT by contrast with under-represented human nouns; there was a preponderance of over-represented adjectives conveying opinions in translated-PT in opposition to under-represented adjectives describing facts; there was a predominance of over-represented adverbs of manner in translated-PT as opposed to under-represented adverbs of time and frequency; and there was a striking number of over-represented link verbs, reporting verbs, movement verbs and verbs that precede other verbs in translated-PT by contrast with highly lexical verbs with dramatic proposi-tional meanings that were found to be under-represented in original-PT.

In addition to the findings above, this study was able to support some previous intuitions regarding words with an exceptionally high frequency in Portuguese translated from English. As perceived by Bastos (2008) and myself, there was a large number of adverbs ending in *mente* among the over-represented adverb lemmas in translated-PT. The adjectives *diferente*

('different') and *possível* ('possible'), which had also been felt to be overly frequent in translated-PT, did not reach the top end of the distribution selected for closer inspection (i.e., lemmas that were at least two times more frequent in translated-PT). However, *differente* was found to be 1.7 times more frequent in translated-PT and *possível* was 1.9 times more frequent.[11] Although the verb *poder* ('can – be able to – allow'), which Tagnin (2008) felt might be over-represented in translations, was not found to be so (in fact, it was only 1.2 times more frequent in translated-PT); its near synonyms *conseguir* ('can – manage') and *permitir* ('allow') were considered to be over-represented in that they were both more than two times more frequent in translated-PT.

Despite the fact that we did not anticipate any particular words that could have been under-represented in translated Portuguese, the top-down approach adopted in this study disclosed quite a large number of lemmas that were at least two times less frequent in translated-PT. While some were completely unforeseen, others, like several lemmas often associated with the Portuguese soul and psyche (*saudade*, *triste*, *tristeza*, *alma*, *lembrança*, *sonho* and *sonhar*) were only to be expected. Likewise, thinking back as a translator and lecturer constantly exposed to Portuguese–English contrasts, the markedly greater presence of the adverbs of time *hoje*, *ontem* and *amanhã* in original-PT makes perfect sense and was not very surprising after all.

8. Conclusions and further research

This study attempted to examine distinctive lexical distributions in translated texts from a top-down perspective. I began the analysis with two comparable corpora of translated and non-translated Portuguese literary texts in order to identify the most over- and under-represented noun, adjective, verb and adverb lemmas in the translations.

The results obtained in this exploratory study not only tended to reinforce existing bottom-up intuitions regarding words with an exceptionally high frequency in Portuguese translated from English, but also disclosed a complex mixture of linguistic and cultural differences between original and translated Portuguese which would not have been possible to detect with the naked eye. This explains the first part of the title of this paper, *Suggesting rather special facts*, which is an allusion to four of the lemmas that were found to be particularly over-represented in translated-PT: *sugerir* was 2.6 times more frequent, *bastante* was 4.4 times more frequent, *especial* was 2.9 times more frequent, and *fa(c)to* was 4.3 times more frequent.

While some of my findings can have an immediate impact on the development of multilingual processing, machine translation, translation aids and translator education, many of the contrasts seen require further research.

---

[11] Indeed, this could mean that the upper and lower thresholds used in this study to decide whether or not to inspect a lemma more closely were too restrictive.

In particular, one should bear in mind that an analysis based on lemmas is very general, and more research is needed in order to learn more about the distributions of different word inflections and of the separate meanings of polysemous lemmas. Also, rather than seen in isolation and devoid of co-text, some of the distinctive lemmas identified would benefit from further collocational analyses. The over-represented verbs *acenar* and *abanar*, for example, are frequent collocates of *cabeça* ('head'). If we examine these verb lemmas in context, it is possible to see that, rather than just the verbs, it is the entire phraseological units *acenar a cabeça* and *abanar a cabeça* ('to nod, to shake one's head') that are over-represented in translated-PT.

Although the analysis of inflected forms, polysemy and distinctive phraseology lay beyond the scope of this work, it is hoped that our methodology and our findings can stimulate such research as well as analogous studies of distinctive lemmas using corpora of different text types and other languages.

Acknowledgement

References

Baker, M. 1993. 'Corpus linguistics and translation studies: implications and applications' in M. Baker, G. Francis and E. Tognini-Bonelli (eds) Text and Technology: in Honour of John Sinclair, pp. 233–50. Amsterdam/Philadelphia: John Benjamins.

Bastos, A. 2008. [Discussion about over-represented words in translated Portuguese] (Personal communication, February 2008).

Bick, E. 2000. The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Århus: Århus University Press.

Frankenberg-Garcia, A. 2004. 'Lost in parallel concordances' in G. Aston, S. Bernardini and D. Stewart (eds) Corpora and Language Learners, pp. 213–29. Amsterdam/Philadelphia: John Benjamins.

Frankenberg-Garcia, A. 2005. 'A corpus-based study of loan words in original and translated texts', Proceedings of the Corpus Linguistics Conference Series, Volume 1, number 1. Available online at: http://www.corpus.bham.ac.uk/PCLC/

Frankenberg-Garcia, A. 2007. Building a parallel corpus for translation research and much more. Presentation at the Postgraduate Seminar in

Translation Studies, Universitat Jaume I. November 2007. Castellón, Spain.

Frankenberg-Garcia, A. 2009. 'Are translations longer than source texts? A corpus-based study of explicitation' in A. Beeby, P. Rodríguez and P. Sánchez-Gijón (eds) Corpus Use and Translating. Amsterdam/ Philadelphia: John Benjamins.

Frankenberg-Garcia, A. and D. Santos. 2003. 'Introducing COMPARA, the Portuguese–English Parallel Corpus' in F. Zanettin, S. Bernardini and D. Stewart (eds) Corpora in Translator Education, pp. 71–87. Manchester: St. Jerome.

Inácio, S. and D. Santos. 2005. Documentação da Anotação Morfossintáctica da Parte Portuguesa do COMPARA. Available online at: http://www. linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf

Olohan, M. and M. Baker. 2000. 'Reporting that in translated English: evidence for subconscious processes of explicitation?', Across Languages and Cultures 1 (2), pp. 141–58.

Øverås, L. 1998. 'In search of the third code: an investigation of norms in literary translation', Meta XLIII (4).

Pápai, V. 2004. 'Explicitation: a universal of translated text?' in A. Mauranen and P. Kujamaki (eds) Translation Universals: Do They Exist? pp. 143–64. Amsterdam/Philadelphia: John Benjamins.

Shama'a, N. 1978. A Linguistic Analysis of Some Problems of Arabic to English Translation. D.Phil thesis, Oxford University.

Tagnin, S. 2008. [Discussion about over-represented words in translated Portuguese] (Personal communication, February 2008).

Tirkkonen-Condit, S. 2004. 'Unique items – over- or under-represented in translated language?' in A. Mauranen and P. Kujamäki (eds) Translation Universals, Do They Exist? pp. 177–84. Amsterdam/Philadelphia: John Benjamins.

Tirkkonen-Condit, S. 2002. 'Translationese, a myth or an empirical fact? A study into the linguistic identifiability of translated language', Target 14 (2), pp. 207–20.

Vinay, J.P. and J. Darbelnet. 1958. Stylistique Comparée du Français et de l'Anglais: Méthode de Traduction. Paris: Didier.