Lexical differences between European and Brazilian Portuguese*

*Anabela Barreiro, Luzia Helena Wittmann, Maria de Jesus Pereira***

Abstract

This paper concerns lexical differences between European and Brazilian Portuguese.

It presents two studies aiming at measuring accurately the degree of difference between these two variants of Portuguese. One study addresses current language, the other looks into technical vocabulary.

In addition, the paper presents a discussion of the kinds of differences that have to be taken into account, and puts forward a proposal for the structure of two contrastive computational dictionaries, describing the work already done on the subject.

I. Introduction

This paper reports the work done during a preliminary phase of a machine translation (MT) project, whose aim is the development of a MT system from English to both European and Brazilian Portuguese. The goal of this phase was to describe the lexical differences between the European (EP) and the Brazilian (BP) variants of Portuguese in a contrastive dictionary[1]. The research covered both current and technical language, originating two contrastive dictionaries, which are separately described here.

The expression *current language* is employed here in opposition to colloquial and literary language and should be equated to standard language. Thus, the term *current language* is used to denote those lexical items that, for their generality, are not specific to small communities or to specialized (technical) subjects, but rather are understood and used by the average speaker. By *technical language*, on the other hand, we mean specialized (as opposed to current) language. It is characterized by a high degree of precision in the definition, low polysemy and a high degree of "semantic mapping" towards other languages.

In the first part, the results of the research in current language are presented. This research was based essentially on already existing European and Brazilian Portuguese computational lexicons.

---

* This paper was written in January 1995 after the work produced during a joint project of INESC and Logos Corporation (USA).

** INESC - R. Alves Redol, 9, 1000 Lisboa. Tel. 3100303 E-mail: anamb@inesc.pt, luzia@inesc.pt, jusp@inesc.pt

1 By contrastive dictionary we mean a dictionary that brings face to face two variants of the same language (see section 2.2).

The second part concerns technical language and was based on a *corpus* of technical words translated from English to both European and Brazilian Portuguese. Looking into the translations allowed us both to observe the differences between the two variants of Portuguese in the areas considered, and to evaluate existing bilingual technical dictionaries.

We provide in this paper both qualitative and quantitative information.

Qualitative information concerns the survey of the different kinds of contrasts between the two variants, identified during the work. The methodology and linguistic criteria used to establish and group the contrasts, as well as the specification of the structure of the dictionaries, are also described. For the study on technical language, we include a paradigm of the translation equivalents we found, and a description of the problems in building both a technical bilingual dictionary English-to-(both variants of) Portuguese and a contrastive technical dictionary between European and Brazilian Portuguese.

Quantitative information regards the number of contrastive pairs found and their detailed distribution and analysis.

II. Current Language Contrastive Dictionary

To build up a dictionary of two variants of the same language that seem very close, but that actually have many and deep differences, either in the construction of phrases and sentences (syntax) or in current vocabulary, forced us to face the complexity of lexical differences. It is important to consider that the two variants are in contact and that some influence exists, at least from Brazilian Portuguese into European Portuguese, probably as a consequence of Brazilian TV programs consumed daily in Portugal.

Trying to handle the complexity foreseen, we distinguished some kinds of contrasts. For instance, there are not only different words for the same referent in each variant, but also different frequency in the use of words common to both variants. Not only the same word is used with different meanings but there are also words related to Brazilian or Portuguese specific realities. Another distinction concerns grammar. Some of the contrastive pairs are orthographic, others can be classified as morphological and, since some expressions and phrases were included, there are still syntactical contrasts to account for.

The next section (2.1) contains a description of several kinds of contrasts. It must be noted that one contrastive pair can belong to more than one kind. In the following sections (2.2, 2.3 and 2.4) there is a description of the dictionary structure, the methodology adopted and the quantitative data.

2.1. Typology of contrasts

The kinds of contrasts discussed here are divided among proper, preferential and optional contrasts, regarding language use. They can also be divided in orthographic, morphological and syntactical as far as linguistic level is concerned. Institutional contrasts, which we address separately, can actually be considered also as a kind of proper contrast.

2.1.1. Proper contrasts

Proper contrasts cover the contrastive pairs and words that are exclusive, i.e., words whose correspondent does not exist or is different in the other variant. Here two types of contrasts were included: (i) different words for the same referent and (ii) words in one variant that have no equivalent in the other, because their referent do not exist, or at least, those words are not common in the culture of one of the countries.

2.1.1.1. Different words for the same referent

Considering the kind of proper contrast where different words for the same referent are used, it is still possible to distinguish some different cases:

a) exclusive words of one variant corresponding to different exclusive words on the other variant, i.e., both elements of the contrastive pair are integrally contrasting words:

EP	BP
autocarro	*ônibus*
hospedeira (de bordo)	*aeromoça*

b) words used in both variants but with different meanings. For example, *banheiro* in European Portuguese means *life-guard*, while in Brazilian Portuguese it means *bathroom*:

EP	BP
banheiro	*salva-vidas*
casa-de-banho	*banheiro*

c) words which have at least one different use in both variants, but that share also, at least, one common meaning. For example, *alcatrão* is the name of a chemical substance with certain properties in both variants, but in Portugal is also used to designate a type of road pavement, while in Brazil the common designation for this is *asfalto* and never *alcatrão*:

EP	BP
(estrada de) alcatrão	*(estrada de) asfalto*
cartão	*papelão*

2.1.1.2. Words without equivalence

There are words that constitute contrasts because they have no equivalent in the other variant. Normally, this happens when the referent is not common to the two cultures where the variants are spoken. The absence of equivalent is marked in our dictionary by "---":

EP	BP
---	*abati*
alcatruz	---

Some other words are specific in the lexicon of one of the variants, because they refer to different realities, as the name of certain plants, fruits or animals. These words are considered contrasts here only if they do not belong to the current language of one of the variants. They probably do not constitute contrasts for specialized scientists in areas as botany, ichthyology, ornithology, etc.:

EP	BP
azinheira	---
---	*sapoti*

2.1.2. Preferential contrasts

Preferential contrastive words refer to words available in both variants, but which are not used with the same frequency. In other words, they exist in both variants with the same meaning, but they become contrasting words from the point of view of their use in current language.

Both words of the contrastive pair can be preferential, as well as only one of them. When only one word of the pair is preferential in one of the variants, this means that its equivalent is only available in this variant.

For instance, in European Portuguese the word *açougue* is available, but *talho* is preferential. The word *talho* is not available in Brazilian Portuguese. In European Portuguese the word *xícara* is available, but *chávena* is preferential. In Brazilian Portuguese the word *chávena* is available, but *xícara* is preferential:

EP	BP
talho	*açougue*
chávena	*xícara*

2.1.3. Optional contrasts

Optional contrastive words are exclusive words of their own variant, but they are also less frequently used synonyms of their equivalent in the other variant. Even if the optional contrastive words are less frequently used than their synonyms that do not constitute contrasts, they can still belong to the current language.

As these contrasts are always optional in relation to their equivalents, only one word in each contrastive pair can be optional.

For instance, in European Portuguese the word *segredo* is optional because *solitária* is also available and preferential. In Brazilian Portuguese the word *sebo* is optional because *alfarrabista* is also available and preferential:

EP	BP
segredo	*solitária*
alfarrabista	*sebo*

2.1.4. Institutional contrasts

Institutional contrastive words are related to different organizational systems in Brazil and Portugal: official institutions, administrative regions, educational and governmental systems, etc. They were distinguished here because they represent a reality that plays the same role, but which is not equally organized in the cultural environment of the two countries:

EP	BP
Ministério dos Negócios Estrangeiros	*Ministério das Relações Exteriores*
distrito	---
Presidente da Junta de Freguesia	*Prefeito*

2.1.5. Morphological and syntactical contrasts

2.1.5.1. Morphological contrasts

A significant number of the contrasts gathered can also be classified as morphological contrasts. These include different derivation (prefixes and sufixes) and different inflexion in the two variants (see Table 4, section 2.4.2.3). The following example shows two contrasts: one regards nominalizations with different derivation suffixes and the other regards two different past participles for the same verb:

EP	BP	
doutora<u>mento</u>	*doutor<u>ado</u>*	(nominalization with different suffixes)
aceit<u>e</u>	*aceit<u>o</u>*	(different inflection)

2.1.5.2. Syntactical contrasts

Despite the fact that our research did not address explicitly syntactical contrasts, we decided to store in the contrastive dictionary those which were found in the course of the study. In fact, some kinds of syntactical contrasts have to be handled using the same methodology which was adopted for lexical contrasts, i.e., they have to be analyzed one by one because it is not possible to define them systematically. This is valid, for instance, for verbal reflexivity, verbal government and prepositional phrases (see section 2.4.2.3):

EP	BP	
reunir	*reunir-se*	(different reflexivity)
participar em	*participar de*	(different government)
a grosso	*em grosso*	(different prepositional phrase)

2.1.6. Orthographic contrasts

As orthographic contrasts we considered pairs of words that differ only in accentuation and/or spelling:

EP	BP	
balancé	*balancê*	(accentuation)
ideia	*idéia*	(accentuation)
linguista	*lingüista*	(accentuation)
amnistia	*anistia*	(spelling)
sumptuoso	*suntuoso*	(spelling)
electrónico	*eletrônico*	(accentuation and spelling)

The orthographic contrasts, even though stored separately, are also classified as proper, preferential or optional. One of the reasons for the detailed identification of the orthographic contrasts is to have them suitably changed in case the proposal of the *Novo Acordo Ortográfico da Língua Portuguesa*[2] is signed in all countries involved.

2.2. Dictionary Structure

The contrastive dictionary was structured in order to make possible and facilitate its use for different purposes. As it is not an explanatory dictionary, no lexical entry is followed by an explanation.

Each entry is organized in four columns: the first column stores the European Portuguese word. The second column stores the equivalent word in Brazilian Portuguese. The third and fourth columns contain remarks about respectively the first (European Portuguese word) and second (Brazilian Portuguese word) columns.

2 The *Novo Acordo Ortográfico da Língua Portuguesa* (New Orthographical Agreement for the Portuguese Language) aims at establishing an orthographical uniformization in all countries where Portuguese is the official language. However, at the time of this study, the agreement had not yet been signed by all parties.

With the European Portuguese words in the first column, the entries are organized alphabetically for European Portuguese. However, note that it takes one simple procedure in order to have the Brazilian Portuguese words in the first column, and one sorting procedure to organize the dictionary alphabetically for Brazilian Portuguese.

The remarks on the third and fourth columns identify the kind of contrast of each lexical entry and, when necessary, its semantic field.

To identify the kind of contrast, the following abbreviations were used in the third and fourth columns:

pref.	= preferential contrasts
opc.	= optional contrasts
SIST.	= institutional contrasts
"." (no abrev.)	= proper contrasts

Orthographic contrasts:

ort.	= orthographic contrasts (spelling)
acent.	= orthographic contrasts (accentuation)
ort./acent.	= orthographic contrasts (spelling and accentuation)

It must be stressed that the orthographic remarks (*ort., acent.* and *ort./acent.*) may be combined with one of *pref., opc.* or *SIST.*. So, when the orthographic remarks appear alone, this means that they are orthographic proper contrasts.

Considering that a word can be contrastive in one of its meanings only, whenever we considered necessary, we noted down in the third and/or fourth columns the area to which the word belongs, in order to clarify in which way there is a contrast. Besides, those contrastive words which relate to specific realities as, for instance, ornithology or botany, were always marked. The encoding was done according to the information available in the paper dictionaries. The marking consists on an (European Portuguese) abbreviation of the name of the area, e.g., *electron*. for *electronics* or *fut.* for *football*, as shown in Table 1. Those contrasts can also be conveyed to the technical dictionary.

The words belonging to phrases or expressions were connected by an underscore to make it easier to automatically process the contrastive lexicon.

Morphological and syntactical contrasting words were not explicitly marked regarding linguistic level.

EP	BP		EP	BP
---	*abaianado*		.	.
Ministério_do_Tesouro	*Ministério_da_Fazenda*		SIST.	SIST.
acupunctura	*acupuntura*		ort.	ort./pref.
actual	*atual*		ort.	ort.
alfarrabista	*sebo*		.	opc.
atómico	*atômico*	acent.	acent.	
baraço	*corda*		.	pref.
canto	*escanteio*		fut.	fut.
chávena	*xícara*		pref.	pref.
coluna	*alto-falante*		electron.	electron.
descolagem	*decolagem*		aer.	aer.
distrito	---		SIST.	.
segredo	*solitária*	opc.	.	
talho	*açougue*	pref.	.	

Table 1 - Extract of the current dictionary

2.3. Methodology

The contrastive dictionary was built up by working with INESC-owned lexical data, divided in different sources, and by adopting different methodologies.

As main source for European Portuguese we used part of the lexicon of PALAVROSO[3], which contains a total of about 57,000 uninflected lexical items. This list will be referred here as the European Portuguese list.

For Brazilian Portuguese we used an on-line list with about 68,000 root forms. This list is also part of INESC-owned lexical data, and will be referred here as the Brazilian Portuguese list.

The contrasts were collected from both lists of words using the same method: from the European Portuguese list, the Brazilian researchers extracted words not common to the two variants and tried to find the cases of contrast with Brazilian Portuguese; the same was done with the Brazilian Portuguese list, where the Portuguese researchers tried to find the corresponding contrasting words in European Portuguese. With this procedure we intended to reach the highest possible coverage of contrasts in both variants. In fact, by using, for instance, only the European Portuguese lexicon it would be impossible to reach specific Brazilian Portuguese words.

This work was done with the help of Portuguese and Brazilian dictionaries and informed by our own knowledge of the two variants. For this task, we used a subset of the European Portuguese list, limited to nouns and adjectives, summing up 34,968 words. The Brazilian Portuguese list was entirely used and includes nouns, adjectives, verbs and adverbs.

A significant number of orthographic contrasts were obtained from the European Portuguese list[4] using a different method. They were extracted automatically following

3 PALAVROSO is a morphological analyser of European Portuguese, developed at INESC. For more information about PALAVROSO see [1], [2], [3] and [4].

the consonant sequences *cc*, *ct*, *pc*, *pç*, *pt*, *mpt*, *bd*, *bt*, *mn*, *mm* and *nn*, for the spelling contrasts described in [5]. Even though a large number of words with these consonant sequences constitute contrasts, this is not systematically valid. For example, the word *pacto* is spelled in the same way in the two variants, but the European Portuguese word *facto* is spelled *fato* in Brazilian Portuguese. Therefore, they were revised one by one. For the accentuation contrasts we analyzed all the words ended in *é* and all the words containing the sequences *ém, én, óm,* and *ón* (e.g. *bebé, académico, biénio, atómico* and *bónus*, whose corresponding words in Brazilian Portuguese are accentuated with circumflex accents). For dieresis checking all the words containing the sequences *gue, gui, que* and *qui* were examined (e.g., *quinquénio* corresponds to *qüinqüênio* in Brazilian Portuguese). These words cover almost all the accentuation differences between the two variants.

An additional list was created, starting from the contrasts already collected and mentioned in [5]. To these, we added data coming from different sources: Portuguese and Brazilian dictionaries, contrastive dictionaries, books, newspapers, magazines and personal knowledge.

Only in the additional list the contrastive pairs that can be considered morphological and syntactical contrasts were discriminated and analyzed.

As regional differences, colloquial words, slang and other words were supposed to be irrelevant for the particular purposes of a machine translation system oriented mainly to technical translations, they were not considered in this survey. However, when contrasts of this register were found, they were stored but not discriminated in a separate file. Therefore, we only mention its number in the following quantitative description.

Some neologisms and foreign words (often not yet included in the available dictionaries) were included whenever we considered that they were frequent enough, as for example *maillot* in European Portuguese and *collant* in European and Brazilian Portuguese.

2.4. Quantitative Description

2.4.1. Total regarding kinds of contrast

The contrasts gathered totalize 4,264, including proper, preferential, institutional, optional and orthographic contrasts. We decided to present in Fig. 1 the total combining the results obtained from the two sources above to display the size of current database, but we should at once note that they cannot be attributed any statistical significance since they merge two lists obtained from different ranges of the lexicon, as was described in the previous section.

4 For this task, the list of words used for European Portuguese included nouns, adjectives and verbs, summming up 48,018 words.

Fig. 1 - General total of contrasts

2.4.2. Total regarding source of contrast

In this section we focus on the several different ways we used to get the above information, analyzing each of them separately.

2.4.2.1. Lists of European and Brazilian Portuguese words

From the lists of European and Brazilian Portuguese words (see section 2.3) we gathered 1,168 pairs of contrasts, by examining each list separately. The two source lists were alphabetically ordered and were handled in the following way:

2.4.2.1.1. European Portuguese list

From the total of 34,968 words (nouns and adjectives) of the European Portuguese list, 3,718 were examined, corresponding to:

from *aba* to *aparvalhado*	2,419
from *bácoro* to *buzinado*	1,299
total	3,718

>From these, 230 were eliminated because they were absent from dictionaries and unknown to the researchers. The 38 contrastive words considered as belonging to the popular or regional register were subtracted and stored in a separate file. Among the remaining 3,550 source list words, 417 were considered contrasting words, which corresponds to 11.74%:

total of words considered	3,718
words eliminated	- 230
popular or regional contrastive words subtracted	- 38
total	3,550
total of contrasts found	417
percentage of contrasts	11.74% (from 3,550)

While analysing the list of the 3,718 words mentioned above, 51 contrasting words that do not belong to the alphabetic order of the examined words were found and added to the source list. Besides, 84 Brazilian Portuguese words without equivalent in European Portuguese were also found and added to the list of contrasts. This produced a

total of 3,601 source list words and the contrasting pairs add up to 552. Thus, the final percentage of contrasts is 15.32%:

words added to the source list	3,550 + 51 = 3,601
new total of contrasts	417 + 51 + 84 = 552
final percentage of contrasts	15.32% (from 3,601)

The total of 552 contrasts is distributed as can be seen in Fig. 2.

Fig. 2 - Total regarding source of contrast: European Portuguese list

2.4.2.1.2. Brazilian Portuguese list

From a total of 68,000 words of the Brazilian Portuguese source list, 6,817 words were examined, corresponding to:

from *aba* to *almofadinha*	3,901	
from *camartelo* to *chinó*	2,054	
from *interinidade* to *janaúba*		862
total	6,817	

>From these, 352 were eliminated because they were absent from the dictionaries and unknown to the researchers. The 72 contrastive words considered as belonging to the popular or regional register were subtracted and stored in a separate file. Among the remaining 6,393 examined words we found 639 contrastive words, which corresponds to 10.07%:

total of words handled	6,817
words eliminated	- 352
popular or regional contrastive words subtracted	- 72
total	6,393
total of contrasts found	639
percentage of contrasts	9.99% (from 6,393)

Only 5 contrasting words that do not belong to the alphabetic order of the examined words were added to this source list. Therefore there are no significant changes in the percentage of the final results of the Brazilian Portuguese list, as it does in European Portuguese list results. Adding these 5 words to the source list and to the list of contrasts we get the total of 6,398 and 644, respectively, corresponding to 10.06%:

words added to the source list	6,393 + 5 = 6,398
new total of contrasts	639 + 5 = 644
percentage of contrasts	10.06% (from 6,398)

The 644 contrasts were distributed as shown in Fig. 3.

Fig. 3 - Total regarding source of contrast: Brazilian Portuguese list

2.4.2.1.3. Comparison of the results based on the European and Brazilian Portuguese lists

Considering that Portuguese and Brazilian researchers adopted different methods of approach to the source list - Brazilian researchers included systematically new contrastive words in the source list whenever they were found while analyzing another word, while Portuguese researchers did not - only the results concerning the original alphabetically ordered words examined in each list were considered for this comparison as shown in Table 2.

	Total number of handled words *	Total number of contrasts **	Percentage of contrasts	Number of pairs common to the two resulting lists
List of EP words	3,550	417	11.74%	28
List of BP words	6,393	639	9.99%	

* This total does not include the words not found in the dictionaries nor the contrastive words considered popular or regional.
** This total includes only current language contrasts. Popular and regional contrasting pairs were not considered.

Table 2 - Comparison of the results in the European and Brazilian Portuguese lists

The remaining percentage discrepancy may be due to the different characteristics of the two source lists.

Only 28 contrasts were found both in the European Portuguese list and in the Brazilian Portuguese list. So, in the total of 1,196 contrasts, 28 are repeated, i.e., 1,168

different contrastive pairs were found. From these 28 pairs, 26 were orthographic contrasts and 2 were proper contrasts. This does not mean that the two lists were completely different from each other. It means that in addition to the difference of the lists characteristics and the difference of the work methods, to use the European or the Brazilian Portuguese list as a starting point does not convey the same results. For example, the word *acanalhado* found in European Portuguese list was not considered a contrast because the word exists in Brazilian Portuguese with the same meaning. But when the word *acafagestado* appeared in the Brazilian Portuguese list, *acanalhado* was attributed as its equivalent, since *acafagestado* does not exist in European Portuguese.

2.4.2.2. Measuring orthographic contrasts

The orthographic contrasts found by examining words with specific consonant and other sequences (see section 2.3) in the entire European Portuguese list "*a*" to "*z*", including nouns, adjectives and verbs, amount to 1,132 and we can see their distribution in Table 3. This means that in a list of 48,019 words, 2.35% are contrastive at the orthographic level.

Even if most of these contrasts are proper, there are also preferential and optional contrasts. Ortographic, optional and preferential contrasts refer to words that can be spelled or accentuated in more than one way in one of the variants, being one of the forms more used. For instance *acupunctura* in European Portuguese can be spelled *acupunctura* or *acupuntura* in Brazilian Portuguese.

Orthographic contrasts	Proper contrasts	Preferential contrasts	Optional contrasts	Total
Accentuation	446	2	2	450
Spelling	495	66	111	672
Acc. and spell.	10	0	0	10
Total	951	68	113	1,132

Table 3 - Orthographic contrasts

2.4.2.3. Other sources of contrasts

The number of contrasting pairs in the additional list, compiled from several different sources (see section 2.3), adds up to 2,033. Adding to this number the 102 contrasting pairs we already had, mentioned in [6], we got 2,135 pairs of contrasts, distributed as Fig. 4 illustrates.

Fig. 4 - Results from several sources

In this list, morphological and syntactical contrasts were counted separately. There are 215 contrastive pairs that can be considered (also) morphological and syntactical contrasts among the above proper, preferential and optional contrasts. We remind that, except for nominalizations and verbalizations, this kind of contrasts was captured occasionally, given that this was not the goal of the research. There are both morphological and syntactical contrasts in Table 4 below, presented in a decreasing order of frequency. Different nominalizations are clearly the most frequent.

Types	Number	Examples	
		EP	BP
nominalizations	131	*doutoramento*	*doutorado*
preposition phrases	16	*de seguida*	*em seguida*
prefixation	12	*fumar*	*defumar*
noun gender	11	*cebolinho*	*cebolinha*
verbal reflexivity	9	*sumir-se*	*sumir*
diminutives	5	*papelinho*	*papelzinho*
past participle	6	*acedido*	*acessado*
noun number	6	*cuecas*	*cueca*
numerals	4	*mil milhões*	*bilhões*
compound words	5	*triplo-salto*	*salto-triplo*
verbalizations	3	*listar*	*fazer uma lista*
adverbs	2	*atempadamente*	*em tempo*
superlative	2	*simpatiquíssimo*	*simpaticíssimo*
verbal government	2	*participar em*	*participar de*
articles	1	*ao serviço de*	*a serviço de*

Table 4 - Morphological and syntactical contrasts

2.4.3. Contrastive pairs found in more than one source

Finally, putting together the pairs of contrasts extracted from the different sources and/or methods in the same file, 205 were eliminated because they were repeated, i.e., they were found in more than one source.

III. Bilingual Technical Dictionary

The goal of this part of the paper is to describe the development of a bilingual technical dictionary which deals with the translations from English to European and

Brazilian Portuguese. We also present several qualitative and quantitative results, namely the description of the dictionary structure and methodology followed by a classification of different kinds of translations and its quantitative distribution.

The data for the translations were collected from English-to-European Portuguese and English-to-Brazilian Portuguese technical dictionaries, glossaries and lists of normalized terms in Portuguese (cf. list of references).

3.1. Dictionary Structure

The technical dictionary is built up of four different columns. In the first column, the technical word in English is placed; the second column contains the translation into European Portuguese; the third column contains the translation into Brazilian Portuguese and, the fourth column stores the information about the terminological area to which the word belongs, as represented in Table 5. Note that here, like in the current language dictionary, the alphabetical order can be suitably changed for any language or variant.

Eng.	EP	BP	Area
active_centre	*centro_activo*	*centro_activo*	Chemistry
adapter	*adaptador*	*adaptador*	Electronics
address	*endereço*	*endereço*	Computers
china_clay	*caulino*	*caulim*	Geophysics
dial	*disco*	*disco*	Telecommunications
engine	*motor*	*motor*	Mechanics
file	*ficheiro*	*arquivo*	Computers

Table 5 - Extract of the technical dictionary

There is no reference to preferential, optional or institutional elements here, as opposed to the dictionary structure described in the first part of this paper. In fact, translations for technical language are supposed to be objective and accurate, while in current language there is no such preciseness.

3.2. Methodology

To build the technical dictionary, we used a corpus containing real texts from a set of technical American English files coming with a commercial MT system. The subjects handled in those texts were: business, transportation, engineering, industry, computers and economy.

>From that corpus we extracted a list of words including only technical nouns, verbs and adjectives. Prepositions, adverbs, articles and other determiners, numbers and punctuation were excluded as well as current language words. Following the selection, we ordered all words alphabetically. To these words we added a significant number of other English words, mainly nouns related due to the same stem or the same semantic

field, or added just because they were very frequent in their field. After this procedure, the number of words in each area was a result of our own intuition about which entries would be the most important, restrained by the characteristics and limitations of the paper dictionaries we were using as source. In fact, some words were reassigned a different area and the whole set of areas significantly extended.

During the task of translating the technical vocabulary, we pulled out all the current language words which were still left behind and some words which appeared in the technical dictionaries although belonging to all terminological areas. Some of them were generally used words and should rather appear in the dictionary of the current language.

Since the fields were not clearly defined in the English source, part of the source list was first translated to European Portuguese, and the other part first to Brazilian Portuguese. Thus, when the starting point was European Portuguese, the fields translated were defined according to the Portuguese technical dictionaries. The opposite happened when the starting point was Brazilian Portuguese. Even if those words appeared in the dictionaries of the other variant, they were not necessarily translated for the same areas, or could be translated without indication of a specific area. In these cases, the entries were not filled in.

We consider that the fact we started from European and Brazilian Portuguese dictionaries with different properties influenced in a general way the results obtained, as displayed in section 3.7.1.

3.3. Typology of translations into Portuguese

Here we survey the kinds of translations for both variants of Portuguese, from a European to a Brazilian Portuguese contrastive point of view.

Whenever there is more than one translation into Portuguese for a particular English term in the technical dictionaries, we decided to list it as many times as necessary (see section 3.3.1 below). Obviously, more information will be needed in a bilingual dictionary to choose the right equivalent in a given context, but this is a matter about which we are not interested now (our aim being the measure of the differences between the two variants).

The kinds of translations which are registered in the technical dictionary include:

3.3.1. Same translation - different area

English words which have the same translation and belong to different terminological areas are repeated according to each of the areas:

Eng.	EP	BP	Area
answer-back	*resposta*	*resposta de retorno*	Computers
answer-back	*resposta*	*resposta*	Telecommunications

3.3.2. Different translation - same area

There are certain cases in which English words have more than one translation in Portuguese in the same terminological area, referring to different objects:

Eng.	EP	BP	Area
bush	*bucha*	*bucha*	Mechanics
bush	*aro*	*aro*	Mechanics
bush	*manga*	*manga*	Mechanics

3.3.3. Different translation - different area

Often, English words have more than one translation in Portuguese according to the area they belong to. In this case, we repeat the word followed with the appropriate translations to the specific area(s) which is(are) marked right opposite in the fourth column:

Eng.	EP	BP	Area
apron	*placa de manobra*	*plataforma*	Aeronautics
apron	*capa protectora*	*cobertura protectora*	Engineering
carrier	*condutor*	*portadora*	Electricity
carrier	*transportador*	*transportador*	Biology

3.3.4. Particular translation - particular area

This section concerns English words which are common to several terminological areas, but that have a particular translation in a given one:

Eng.	EP	BP	Area
average	*avaria*	*avaria*	Navigation

In all other terminological areas, *average* (noun) is translated as *média*.

3.3.5. English word - Portuguese definition

Some English words, instead of single word translations, are translated by complex expressions:

		Area
Eng.	can	Aeronautics
EP	*caixa da câmara de combustão*	Aeronautics
BP	*câmara de combustão de motor turbojato*	Aeronautics

3.3.6. English word - English word in Portuguese

A few English words have no translation into Portuguese, since they are accepted in our language with its original English form:

Eng.	EP	BP	Area
cardan	*cardan*	*cardan*	Mechanics
bit	*bit*	*bit*	Computers

3.3.7. Missing areas for an English word

In several cases, a word can have more than one meaning but, if this is not registered in the dictionaries, we did not store those translations. This is certainly one limitation of the dictionaries.

3.4. Acronyms

Among the technical terms we found also some acronyms. This raised some problems since the dictionaries did not treat these words systematically. So, the criteria used here were the same used in the existing technical dictionaries.

3.4.1. Without equivalent

In some cases, there was not an equivalent acronym in the target language and the full expression which the acronym refers to had to be used in Portuguese:

Eng	EP	BP	Area
LET	*transferência linear de energia*	*transferência linear de energia*	Physics

3.4.2. Without translation

There were also some cases in which the source and the target words were the same, at least in one variant:

Eng.	EP	BP	Area
ie-impact energy	*ie-impact energy*	*ie-impact energy*	Physics
AIDS	*SIDA*	*AIDS*	Medicine

3.5. Contrasts between European and Brazilian Portuguese translations

It is relevant to point out that also in technical language there are often divergences. European and Brazilian Portuguese translations were compared, disclosing an important amount of contrasts. The nature of the contrasts here was mainly orthographic but a significant number of morphosyntactic differences was also annotated. We do not comment on these contrasts since they were already defined and discussed in the first part of the paper.

3.6. Problems

One important question in the present work was to establish criteria to define which words belong to technical language and which words belong to current language. It was not always easy to make this division. In fact, paper dictionaries do not help very much in answering our needs.

3.7. Quantitative Description

3.7.1. Total of translations

>From the set of words which were excluded from our research, there was a large number of current words (including verbs, nouns and adjectives) which always appear in any kind of text, and many words which, although appearing in the technical dictionaries, belong to several technical areas and are reported also in the current dictionaries. For this reason we did not take them into account for the final results of this particular dictionary. They will be subject of future treatment and/or study.

The total presented in Fig. 5 includes 2,435 English source words: 2,409 words were found in the technical dictionaries, and 26 words were found in a current dictionary, but with the specification of the terminological area. In fact, the same way that some current words appear in the technical dictionaries, also some technical words occur in current dictionaries.

In 1,376 cases of translations found we got full pairs, i.e., translation both to European and Brazilian Portuguese. Even though we used the maximum available information, in 47 cases it was impossible to find translation in the technical dictionaries into any of the variants, either into European or into Brazilian Portuguese. This is the reason why only 2,388 words appear in the total results for translations found. In several other cases, it was only possible to find translations in the technical dictionaries into one of the variants, European or Brazilian Portuguese. The total of translations only into European Portuguese is thus 384 and the total of translations only into Brazilian Portuguese is 628. This difference is partly justified by the fact that we used more technical dictionaries for Brazilian than for European Portuguese.

Fig. 5 - General total of translations

It is important to observe that the discrepancy between the translations into European Portuguese and into Brazilian Portuguese is related to the quality and the quantity of the coverage of the dictionaries. In fact, the considerable number of words not translated to one of the variants can be partly justified by the fact that different general technical dictionaries do not cover necessarily the same areas. Even if those words appeared in the dictionaries of the other variant, they were not necessarily translated for the same areas, or were translated without any indication of a specific area. In these cases, the entries were not filled in, but they are an important reference for further research.

We present in appendix a detailed distribution of the technical translations obtained and their frequency. Obviously some terminological areas were much better covered than others. The areas which we covered are due much more to the dictionaries than to the original texts which, in most cases, did not consider these areas.

3.7.2. Results of the search in a specific area

Since all bilingual technical dictionaries consulted covered several areas at the same time, we made a special search in medicine, consulting a specific technical dictionary available for Brazilian Portuguese. Medicine was chosen only as an example, so that we could get an idea of how far one can go using only dictionaries for this sort of investigation.

Before looking into this dictionary, of 193 English terms only 45 were translated, being 148 cases not covered. On consulting it, we could find 80 more translations, ending up with 125 translations for Brazilian Portuguese.

In this dictionary we could also come across with 18 translations to biology, 9 to anatomy, 7 to zoology, 3 to psychology, 2 to immunology and 1 to chemistry, ending up with 26 words unhandled. The results obtained here are already included in the final total presented above. Since it was not possible to do the same for all areas, a significant number of words were not translated.

The outcome of this search was interesting in the sense that it was possible to evaluate different types of dictionaries. In short, we confirmed that specific technical dictionaries offer better information about a given area and cover cases which are not possible to cover with an ordinary technical dictionary.

Finally, we concluded that only a deeper study with the help of specialists in these areas and the consulting of corpora would allow us to translate all occurrences.

3.7.3. Technical Language Contrasts

As we have already seen, from the 1,376 cases where we got translations either to European and Brazilian Portuguese, 928 times we got the same translation and 451 times the translation was different (see Fig. 6). The contrasts corresponded to 32.77%.

109 contrasts are orthographic contrasts: 88 concerning spelling, 16 concerning accentuation and 5 concerning spelling and accentuation; the remaining 342 are proper contrasts.

Fig. 6 - Total for technical language contrasts

Given that 75.83% of the contrasts are proper, if only proper contrasts were considered, the percentage would still be 24.85%, which is a surprisingly high number. However, it must be considerated that this percentage was obtained from contrasts which were not only single words but mostly terms composed of several words, some of them with differences in syntax and style. We also noticed that some of these contrasts seemed to be more a choice of the terminologist than a description of actual use: preciseness in technical dictionaries is sometimes a relative concept.

IV. Conclusion

The research reported in this paper is pioneer in its field. Excepting one contrastive dictionary [18] that presents a collection of words in a bilingual paper dictionary format, we are not aware of any systematic research concerning the lexical differences between European and Brazilian Portuguese.

Our goal in this study concerned the feasibility (and the dificulties involved) in the development of natural language processing tools available for the two variants of Portuguese. We distinguished two kinds of lexical units: current language words and technical words.

The first result of this work was a fine grained discrimination of kinds of contrastive pairs, where frequency of usage of a word was also taken into account. We considered that words with different frequency of use can constitute contrasts as well as, for instance, words that do not exist in one of the variants. In fact, the extent of familiarity of a text for a native speaker depends significantly on the frequency of the words that it contains in the general language.

The second major result was the quantitative extent to which the two variants differ, which was detailed in the sections dealing with the quantitative description. About 11% of the current language lexicons was constituted by contrastive words. Interestingly, this percentage obtained with lists of words in dictionary is only slightly higher than the one observed in our previous corpus-based research [6], where only words in context were considered.

On the other hand, in the technical language lexicons, obtained by translations of English words both into European and Brazilian Portuguese, the contrastive words reached 32%. This percentage was somehow unexpected. In fact, there did not seem to be *a priori* a reason why it should be very different from the current language percentage of contrasts. However, we found that there were almost three times more contrastive words in the technical vocabulary we studied than in the current language lexicon.

Despite the interesting results obtained, including a contrastive dictionary with 4,264 entries, this work should be considered as preliminary research. It highlights important difficulties in the comparison of these two variants of Portuguese and establishes general criteria for a larger study, which must be extensively supplemented by corpora consultation.

AKNOWLEDGEMENTS

To do the work reported here we had the part-time collaboration of Fátima Hoogland, Susana Mendonça and Tânia Regina Pêgo. We would also like to thank Fernando Jesus for technical support.

We are specially grateful to Diana Santos for helpful discussion on methodological issues and general supervision.

This work was only possible due to a joint project with Logos Corporation (USA) which provided not only the funding but also the opportunity to address the subject described in this paper.

GENERAL REFERENCES

[1] MEDEIROS, J.C.: *Morphological Processing and Spelling Correction of Portuguese* (in Portuguese), MSc Thesis, Instituto Superior Técnico, Technical University of Lisbon, February 1995.

[2] MEDEIROS, J.C.: "Corpora Tools" (in Portuguese) in SANTOS, D. (ed.), *Processing of Corpora Text at INESC*, INESC report RT/65-92, December 1992.

[3] MEDEIROS, J.C., MARQUES, R. and SANTOS, D.: "Quantitative Portuguese" (in Portuguese), in *Actas do I Encontro de Processamento da Língua Portuguesa Escrita e Falada, EPLP'93* (Lisbon, February 25-26, 1993), pp. 33-38, 1993.

[4] SANTOS, D.: "Computational Portuguese" (in Portuguese), in *Actas do Congresso Internacional sobre o Português* (Lisbon, April 13-15, 1994), APL, to appear.

[5] WITTMANN, L.H. and PEREIRA, M.J.: "European and Brazilian Portuguese: Survey of Morphological, Syntactical and Orthographic Differences", INESC internal report, Lisbon, July 1994.

[6] WITTMANN, L.H. and PEREIRA, M.J.: "European and Brazilian Portuguese: Quantitative Report of Lexical, Syntactical and Orthographic Differences", INESC internal report, Lisbon, July 1994.

[7] WITTMANN, L.H. and PEREIRA, M.J.: "European Portuguese and Brazilian Portuguese: some contrasts", in *Actas do X Encontro da Associação Portuguesa de Linguística* (Évora, October 6-8, 1994), APL, to appear.

EUROPEAN PORTUGUESE REFERENCES

[8] COSTA, J.A. and MELO, A.S.: *Dicionário da Língua Portuguesa*, Porto Editora, Porto, 6ª ed. corrigida e aumentada, 1992.

[9] FIGUEIREDO, C.: *Grande Dicionário da Língua Portuguesa*, Bertrand Editora, Venda Nova, 23ª ed., 1986.

[10] HOUAISS, A. and CARDIM, I.: *Webster's - Dicionário Inglês-Português*, Círculo de Leitores, Lisboa, 1989.

[11] MORAIS, A.: *Dicionário de Inglês-Português*. Porto Editora, Porto, 1ª ed., 1984.

BRAZILIAN PORTUGUESE REFERENCES

[12] ANTAS, L.M.: *Dicionário de Termos Técnicos Inglês-Português*, Traço Editora, 6ª ed., 1980.

[13] FERREIRA, A.B.H.: *Dicionário Aurélio Electrônico*, Editora Nova Fronteira, Rio de Janeiro, 22ª ed. revista e aumentada, 1986.

[14] FERREIRA, A.B.H.: *Novo Dicionário da Língua Portuguesa*, Editora Nova Fronteira, Rio de Janeiro, 1993.

[15] FÜRSTENAU, E.: *Novo Dicionário de Termos Técnicos Inglês-Português*, Editora Globo, Porto Alegre-Rio de Janeiro, vols I e II, 13ª ed., 1986.

[16] PRATA, M.: *Dicionário de Português*, Schifaizfavoire, Crónicas lusitanas, Editora Globo, São Paulo, 8ª ed., 1993.

[17] STEDMAN, E.: *Dicionário Médico*, Editora Guanabara, Koogan, S.A., Rio de Janeiro, vols I e II, 23ª ed., 1979.

[18] VILLAR, M.: *Dicionário Contrastivo Luso-Brasileiro*, Editora Guanabara, Rio de Janeiro, 1989.

Appendix

Terminological Area	English words	Translations to EP & BP	Translations only to EP	Translations only to BP
Computers	451	228	148	75
Electricity	285	177	25	83
Mechanics	238	95	24	119
Electronics	184	95	10	79
Physics	155	97	24	34
Medicine	140	107	15	18
Chemistry	131	90	14	27
Aeronautics	114	71	14	29
Mathematics	74	48	6	20
Telecommunications	62	44	15	3
Civil Engineering	56	21	26	9
Biology	54	47	3	4
Botany	31	15	16	0
Zoology	25	23	2	0
Psychology	25	18	7	0

Nuclear Physics	24	11	1	12
Navigation	23	14	0	9
Anatomy	23	21	1	1
Geology	22	17	4	1
Mining	20	14	5	1
Military	18	6	0	12
Metereology	18	15	0	3
Engineering	17	10	6	1
Optics	12	4	0	8
Nautics	12	1	1	10
Astronomy	12	6	1	5
Printing	9	7	2	0
Magnetism	9	2	0	7
Geometry	9	3	0	6
Geography	9	3	1	5
Television	8	5	2	1
Space	8	5	3	0
Architecture	8	7	0	1
Mineralogy	7	5	0	2
Ecology	7	4	3	0
Metallurgy	7	0	0	7
Economy	6	5	1	0
Radio	5	4	0	1
Photography	5	1	0	4
Law	5	5	0	0
Atomic Physics	5	1	0	4
Agriculture	5	0	0	5
Radar	4	2	1	1
Immunology	4	4	0	0
Weaving	3	0	0	3
Tipography	3	0	0	3
Textiles	3	1	0	2

Hydraulics	3	1	0	2
Automobile	3	1	0	2
Veterinary	2	0	2	0
Teleprocessing	2	0	0	2
Nuclear Engineering	2	2	0	0
Geophysics	2	2	0	0
Crystallography	2	0	0	2
Topology	1	1	0	0
Telephony	1	0	0	1
Surveying	1	1	0	0
Statistics	1	1	0	0
Sound Equipment	1	0	0	1
Radiology	1	1	0	0
Plastics	1	1	0	0
Physiology	1	1	0	0
Petrology	1	0	0	1
Painting	1	0	0	1
Music	1	1	0	0
Management	1	1	0	0
Industrial Safety	1	0	0	1
Hydrology	1	0	1	0
Graphic Arts	1	1	0	0
Cristallogy	1	1	0	0
Accounting	1	1	0	0

Total	2,388	1,376	384	628