

# PtTenTen: A Corpus for Portuguese Lexicography

Adam Kilgarriff, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha  
and Pete Whitelock

## 1. Introduction

There are a number of ways in which corpus technology can support lexicography, as described in Rundell and Kilgarriff (2011). It can make it more accurate, more consistent and faster. But how might those potential benefits pan out in an actual project? If starting from a blank sheet of paper, how should one proceed?

In this paper we describe such an exercise. Oxford University Press is preparing the *Oxford Portuguese Dictionary*, a new Portuguese–English, English–Portuguese dictionary. It will cover both Brazilian and European Portuguese, with differences of words, spelling and usages noted. Each side will contain around 40,000 headwords and 200,000 meanings. The work here concerns the new analysis of Portuguese for the Portuguese-source side.

The components of the process are:

- Collect the corpus
- Process it with the best available tools for the language
- From parser output to corpus system

First, we present the end point of the process: high quality word sketches for Portuguese within the Sketch Engine corpus query tool. Then in the next three sections we describe the process of getting there. Last, we present an analysis of the contrast between Brazilian and European Portuguese.

## 2. Word sketches and the Sketch Engine

Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. Their value for lexicographic work in English and other languages, as well as the background of the use of corpora in lexicography, have been described elsewhere (Kilgarriff et al., 2004).

First, we introduce corpus query systems and the basic idea of word sketches. Next, we present word sketches for Portuguese.

## 2.1 Corpus Query Systems

A variety of corpus query systems (CQSs) have been used to examine corpus evidence since the rise of the first electronic corpora. Starting with the ground-breaking COBUILD project, lexicographers have been using KWIC (Key Word In Context) concordances as their primary tool for finding out how a word behaves. Later, with the growth of corpora, lexical statistics had to be applied to manage the abundant data and highlight the most salient combinations and collocations. Today, state-of-the-art CQSs allow the lexicographer great flexibility in searching for phrases, collocates, grammatical patterns, sorting concordances according to a wide range of criteria, selecting subcorpora for searching in, say, only spoken text, academic text, or only fiction. Available systems include WordSmith (Scott, 2008), MonoConc (Barlow, 2000), the Stuttgart Corpus Workbench (CWB, Christ and Schulze, 1994) and Davies's SQL architecture (see Davies, this volume).

## 2.2 The Sketch Engine

The Sketch Engine is a corpus query system which gives access to the familiar CQS functions: concordances for several types of queries (simple, lemma, phrase, word form and CQL), with an integrated context-control filter.

The interface includes a variety of viewing and sorting options.

Figure 5.1 Screenshot of concordance interface

Concordance Word List Word Sketch Thesaurus Find X Sketch-Diff ?

Save View options Sort Sample Filter Frequency Collocations ConcDesc ?

Corpus: ptTenTen11  
Hits: 22,793 (7.0 per p  
Page 1 of 1,140

Left  
Right  
Node  
References  
Shuffle

Brasil 180 pontos a do placar pendurada no centro da cúpula ( Fig. 9), além de  
Brasil a produtora ueta para fazer um filmete e pendurar a conta na Secretaria de Comunicação do  
Brasil cabos de aço que por sua vez se encontra pendurada a mastros localizados externamente às instalações  
Brasil site. </p><p> Eu tb quero uns AMRAAM para pendurar no meu ultra-leve. Um em cada asa prá equilibrar  
Brasil <p> Outra questão é quais mísseis poderemos pendurar caso ele vença a concorrência? O Derb e  
Brasil fundindo-se em mim. Na parede leste ela pendurou concessões de terra e retratos do pessoal  
Brasil à beira de um abismo. Para lhe escapar, pendurou-se na raiz de uma árvore junto ao abismo  
Brasil consiste em diversos pinos de ferro que eram pendurados no teto com uma corda, e foram atraídos  
Brasil um pequeno labirinto, uns panos que foram pendurados no teto e que vão até o chão, também formando  
Brasil Bethulia, onde o perplexo povo da cidade pendurou a cabeça do opressor na muralha. </p><p>  
Brasil litros de líquidos, sobretudo água e sucos; pendurar toalhas molhadas nos ambientes das casas  
Brasil ele colocou um fino ramo de ouro no qual pendurou sinos que pareciam botões em flor . Quando  
Brasil retrato é de 1948. Emoldurado, há muito pendurado na parede do velho quarto. Nele um homem  
Brasil aqui, teu amado líder eunuco Ahmadinejad te pendurava pelo pescoço num guindaste em praça publica  
Brasil dele e de Clara Petacci , sua amante, foram pendurados de cabeça para baixo junto com mais cinco  
Brasil passarinhos engaiolados de Seu Carlitos , pendurados no pé de amêndoa esticam as cabeças tentando  
Brasil dos lados da Flórida: Joe Nieuwendyk vai pendurar os patins, encerrando uma carreira que  
Brasil Ela não os mantém quentes. Parece um peixe pendurado no pescoço. Estamos completamente livres  
Brasil devem ir a um jantar com um peixe morto pendurado ao redor do pescoço. Como budistas, vocês  
Brasil uma loucura! Há bandeiras do Reino Unido penduradas nas principais ruas da cidade, e durante

Page 1 of 1,140 Go Next Last

Figure 5.2 Screenshot of concordance viewing and sorting options

However, the features of the Sketch Engine which are of special interest in this paper are not part of standard concordancing programs. These features include word sketches, sketch differences and a thesaurus. They are all fully integrated with standard concordancing.

### 2.3 Word sketches

To identify a word's grammatical and collocational behaviour, the Sketch Engine needs to know how to identify words connected by a grammatical relation. This can be achieved in one of two ways.

The first possibility is to use a 'sketch grammar': the input corpus is loaded into the Sketch Engine, part-of-speech-tagged but not parsed, and the Sketch Engine supports the process of identifying grammatical relation instances through a grammar written as regular expressions over part-of-speech tags.

In the second approach, we parse the input corpus, so that the information about which word-instances stand in which grammatical relations with which other word-instances is embedded in the corpus. Currently, dependency-based syntactically annotated corpora are supported. Phrase-structured trees need heads of phrases to be marked.

For most languages where word sketches have been built, the first method was used. The work described in this paper is the first large-scale use of parser output to create word sketches.

One of the hardest parts of the lexicographer's task is not to miss anything. In Figure 5.3, we see a Word Sketch for the lemma *pulso*, which has a frequency of 22,328 in the corpus. An inspection of the Sketch shows five basic senses. The first sense refers to the joint that connects the hand to the arm (wrist), indicated by collocates such as *relógio* (watch), *fratura* (fracture) and *esquerdo* (left). The second one (pulse) is medical and means the beat that results from the passing of blood through the arteries and veins – it is revealed by collocates such as *femoral* (femoral), *auscultar* (auscultate) and *basal* (basal). The third one (pulse), related to telephony, is revealed by words such *telefônico* (telephone), *tarifação* (pricing) and *tom* (tone). The fourth one is specific to electricity and collocates with *tensão* (tension), *corrente* (current) and *eletromagnético* (electromagnetic). The last sense is figurative and denotes a measure of strength (hand), forming collocations with *firme* (firm), *firmeza* (firmness) and a number of verbs such as *governar* (govern), *comandar* (command), *agir* (act) and *administrar* (manage). An inspection of these collocations will show the pervasiveness of the idiom '*com pulso firme*' (with a firm/strong hand).

pulso		ptTenTen11 freq = 22328 (6.9 per million)		N_de pulso_N 6678 0.1		V_obj pulso_N 5388 0.1		pulso_N mod ADJ 4509 0.1		pulso_N de_N 3176 0.0	
unary rels		oximetria	181 9.61	palpar	20 6.42	eletromagnético	139 7.89	clock	56 7.31		
de pulso	2188 2091.1	oxímetro	172 9.51	fraturar	40 6.1	ultracurto	33 7.85	bpm	12 6.12		
a pulso	498 2091.1	relógio	1849 8.53	apalpar	26 6.01	firme	843 7.72	Inundação	6 5.9		
DET spec_of pulso_N	7332 0.1	palpação	40 6.92	cortar	704 5.43	carotídeo	23 7.18	LH	9 5.81		
teu	42 1.69	oxímetro	17 6.37	lesionar	12 3.98	radial	54 7.15	radiofrequência	9 5.8		
meu	326 1.61	Oxímetro	15 6.19	machucar	28 3.8	pedioso	13 6.55	Inundação	72 5.79		
vosso	11 1.05	cronógrafo	20 6.01	agarrar	58 3.76	femoral	21 6.17	GnRH	6 5.79		
cada	115 1.0	Oximetria	10 5.61	auscultar	6 3.76	paradoxal	33 6.14	radiofrequência	9 5.56		
seu	710 0.5	correia	29 4.81	algemar	6 3.59	ultrassônico	12 6.04	corticosteroide	7 5.55		
ambos	12 0.3	tarifação	8 4.58	checar	17 3.48	esquerdo	252 6.02	Cura	7 5.4		
		largura	88 4.54	adornar	7 3.42	telefônico	130 5.8	laser	70 5.3		
		relojaria	6 4.26	apertar	42 3.02	filiforme	7 5.56	Laser	7 5.07		
		algema	7 3.99	medir	112 2.98	poplíteo	7 5.48	RF	7 4.96		
		fratura	47 3.96	segurar	56 2.95	basal	14 5.39	sincronismo	11 4.93		
		modulação	11 3.94	emitir	128 2.94	transiente	7 5.36	ultrassom	10 4.32		
		batimento	11 3.74	imobilizar	7 2.91	arterial	50 5.35	defunto	8 4.28		
		tensão	12 3.73	morder	13 2.9	aleatório	37 5.16	quartzo	6 4.09		
		medidor	15 3.69	molhar	12 2.76	magmático	6 5.16	injetor	7 4.07		
		franquia	31 3.65	discriminar	14 2.54	braquial	6 5.13	disparo	30 3.87		
		laser	22 3.54	amarrar	17 2.53	excedente	54 5.05	ferro	80 3.74		
		propagação	23 3.46	magoar	7 2.51	atado	6 5.02	cura	49 3.56		
		veu	11 3.44	esfregar	7 2.47	magnético	50 4.99	voltagem	6 3.28		
		amplitude	26 3.38	quebrar	60 2.32	telefônico	7 4.94	artéria	17 3.18		
		torção	6 3.36	rasgar	10 2.18	periférico	46 4.93	radiação	24 3.13		

Figure 5.3 Partial screenshot of Word Sketch for lemma *pulso*

Word Sketches can also remind lexicographers to include less obvious meanings and idioms. The Sketch for the lemma *pendurar* (hang; frequency of 22,793) includes both of these. For instance, the idiom *pendurar o beijo* (literally to drop one's lower lip) is a less common way of saying *fazer bico* (make a grimace). *Pendurar a chuteira* (literally to hang one's soccer spikes) means to end one's career. This same meaning is conveyed by *pendurar o paletó* (literally to hang one's jacket), which has yet another meaning: to pretend to work (*quem não trabalha, pendura o paletó* – literally those who don't work, just hang their jackets). Another sense of *pendurar* is that of doing a seemingly endless task as in *pendurado ao telefone* (hanging on the telephone).

**pendurar** ptTenTen11 freq = 22793 (7.0 per million)

Constructions	pendurar_V_em_N 13566 0.3	pendurar_V_obj N 7096 0.1	V.comp.pendurar_V 5110 0.1	N.mod.pendurar_V[PCP] 4367 0.3
REFL-SUBJ 884 0.3	varal 253 8.66	chuteira 849 10.63	emoldurar 7 3.59	crucifixo 21 5.53
REFL-PASS 664 0.2	cabide 219 8.08	melancia 53 6.5	ficar 939 1.52	estetoscópio 9 5.52
	pescoço 774 8.0	casaco 58 5.33	costumar 24 1.11	estepe 14 5.29
pendurar_V dep PRP 17711 0.2	madeiro 66 7.07	paletó 20 5.31	resolver 49 0.19	melancia 15 4.9
em 12586 2.29	parede 1452 6.94	bandeirinha 16 4.99	andar 55 0.02	crachá 16 4.9
por 1270 1.12	galho 157 6.52	guizo 9 4.98		chocalho 7 4.86
sob 35 0.82	gancho 98 6.25	enfite 29 4.94		argola 12 4.7
sobre 227 0.75	teto 426 6.17	bota 57 4.86		calcinha 14 4.22
durante 71 0.47	prego 81 6.0	luva 55 4.84		paletó 7 4.06
sem 83 0.29	teta 46 5.93	móite 8 4.84		gaiola 16 4.03
há 30 0.24	parapeito 27 5.61	beijo 9 4.82		cabide 8 3.92
até 133 0.15	retrovisor 46 5.59	patim 17 4.8		enfite 12 3.8
	poste 129 5.55	capacete 45 4.69		corda 45 3.51
	corda 193 5.51	bandeirola 7 4.56		morcego 7 3.31
	estendal 21 5.45	plaquinha 8 4.51		lençol 16 3.29
	brocha 18 5.36	crucifixo 12 4.48		póster 7 3.2
	forca 31 5.26	sapatilha 12 4.48		barraco 7 3.2
	lustre 26 5.25	toalha 46 4.48		gravata 8 3.12
	trapézio 21 5.19	cabide 13 4.41		cartaz 44 3.11
	árvore 475 5.13	lençol 36 4.39		casaco 11 3.02
	barbante 19 5.03	saquinho 15 4.35		macaco 12 2.9
	ombro 106 4.97	gaiola 21 4.31		luminária 7 2.86

Figure 5.4 Partial screenshot of Word Sketch for lemma *pendurar*

### 3. Corpus collection

Corpora for lexicography should be large and diverse. If they are, they will provide evidence about anything that should be in the dictionary. If they are not, they will miss things. Our experience with English shows that, in order to get a full account for each of 40,000 words of a language – even the least frequent of them – we need a corpus of at least a billion words.

Where might a corpus of that size, covering a very wide range of text types, be found? The answer is the web. There is now substantial evidence that web corpora, created through the same process of web crawling that the search engines use, offer diverse and very large corpora which compare well with designed collections (Baroni et al., 2009; Sharoff, 2006). Informal and speech-like genres tend to be better represented in web corpora than in many curated corpora, since they contain material from blogs and similar, while curated corpora in the order of a billion words are likely to include high proportions of journalism, the easiest text type to obtain in bulk. While there is no easy answer to the question ‘what text types, and in what proportions, do we get in a web corpus’, we show below that they provide good lexicographic resources. The Portuguese corpus described here is one of the ‘TenTen family’ of corpora (Jakubíček et al., 2013).

#### 3.1 Crawling

The Portuguese corpus was gathered in two parts, the first for European (crawling only in the .pt domain), the second for Brazilian (.br domain). Following Baroni et al., we used the Heritrix crawler (<http://crawler.archive.org/>) and set it up to download only documents of mime type text/html and between 5 and 200KB in size. The rationale of mime type restriction is to avoid technical difficulties with converting non-HTML documents to plain text. The size limit weeds out documents that are too small, which

**Table 5.1** Web crawling stats

	European Portuguese	Brazilian Portuguese
HTML data downloaded	1.10 TB	1.37 TB
Unique URLs	31.5 million	39.1 million
Crawling time	8 days (1–8 Mar 2011)	10 days (1–10 Jun 2011)

typically contain almost no text, and very large documents, which are very likely to be lists of various sorts. Table 5.1 summarizes the sizes of the downloaded data as well as the time required for crawling.

### 3.2 Junk

We do not want our Portuguese corpus to contain material that is not Portuguese text. We do not want it to contain navigation bars, banner advertisements, menus, formatting declarations, JavaScript, html, or material in languages other than Portuguese. It is also important that we represent all texts in a single character encoding (preferably UTF-8) in order to prevent incorrect character display.

Detecting original character encoding of each document is our first step, for which we use the `chared` tool.<sup>1</sup> Once we know what the original encoding is, converting it to UTF-8 is straightforward.

Next, we remove junk (navigation links, advertisements, etc.) with `jusText`.<sup>2</sup> We run it with the inbuilt Portuguese model and with the default settings.

In order to preserve only texts in Portuguese, we apply the Trigram Python class for language detection using character trigrams.<sup>3</sup> We train a Portuguese language model from a 150,000 word text sample taken from Wikipedia and discard all documents for which the similarity score with the language model is below 0.4. This threshold is based on the results of our previous experiments.

The first manual examination of the corpus data revealed a substantial amount of English text despite the applied language filtering. It turned out that there are numerous documents in the corpus which contain half-Portuguese, half-English paragraphs and score slightly above the language filtering threshold. To fix this problem, we applied further anti-English filtering. We compiled a list of the 500 most frequent words of English and removed from the corpus all paragraphs longer than 50 words where the frequent English words accounted for over 10 percent of the words.

### 3.3 Duplicates

Duplicates (and, worse still, many-times-replicated material) are bad both because the lexicographer wastes time passing over concordance lines they have already seen and because they distort and invalidate statistics.

A central question regarding duplication is ‘at what level?’ Do we want to remove all duplicate sentences, or all duplicate documents?

For lexicographic work and other research at the level of lexis and syntax, the sentence is too small a unit, because if we remove all but one copy of a short sentence such as ‘Yes it is’ or ‘Who’s there?’ the remaining text will lose coherence and be hard to interpret. The whole document is too large a unit because we do not want to include long sections of text twice over where one appeared in document X and the other in document Y, and the other parts of document X did not duplicate the other parts of document Y.

The appropriate unit is the paragraph. We identify paragraphs, and then take additional steps to handle short paragraphs (including dialogue turns like ‘Yes it is’), only removing them if their context is also duplicate material.

A naïve approach to de-duplication results in a process that gets slower per million words, the larger the corpus (since there are more already-seen paragraphs to compare a new paragraph with). Our approach increases linearly with the size of the corpus. We de-duplicate after cleaning, since this reduces the bulk of material to de-duplicate. The de-duplication process was applied separately for the European and Brazilian parts. It took 4 hours and 5 hours respectively on a single Intel Xeon 2.13GHz CPU and removed 75 percent and 68 percent of the cleaned material that we had gathered, leaving 804 million tokens of European Portuguese and 3.19 billion of Brazilian.

#### 4. Language technology tools for processing Portuguese

The prospects for getting the computer to help the lexicographer are improved if the text is lemmatized, part-of-speech-tagged and parsed. The lexicographer can then ask queries about lemmas, word classes and grammatical relations (‘what nouns often occur as objects of this verb?’) as well as about word forms and positions (‘what words often come between two and five words after this word?’). We shall be able to provide better reports to the lexicographer.

We investigated past research on the computational processing of Portuguese (e.g., Santos et al., 2008) and established that the leading system was PALAVRAS (Bick, 2000; see Bick’s chapter in this volume). Further investigation revealed that PALAVRAS development has been ongoing for over ten years and did not reveal any newcomers that looked better. We concluded that it was probably, in 2011, the most accurate software for processing Portuguese. We contacted the author and negotiated a license.

Parsing tends to be a slow process. One concern of ours was that parsing a 2 billion word corpus would take months or even years.

We parallelized the processing by splitting the corpus into 12 parts and parsing all of them at the same time on a double 12-core AMD Opteron 800 MHz server. We experienced technical problems with the parser and had to re-start several times with software bug fixes and updates obtained from the developers upon our error reports. Despite good technical support, we were unable to parse the whole data set in a single run without the process dying. In the end, we split the data into many files of around 10MB and ran a fresh instance of PALAVRAS for each file. In the final run, using 12 concurrently running instances of the parser, the processing of the whole data set took 15 days.

The parser crashed on most of the input files. Nevertheless, in most cases it managed to process a significant part of the input first. A substantial part of the corpus data was lost during parsing. The final size of the corpus is 773 million tokens for the European part and 1.2 billion tokens for the Brazilian.

## 5. Into the Sketch Engine

PALAVRAS is a dependency parser. In dependency grammar, the structure of a sentence is identified via a set of labelled dependency links from each word to its governor. For each word in a sentence, PALAVRAS output provides the lemma, the part-of-speech tag, the name of the grammatical relation in which it stands to its governor, and a pointer to its governor.

Although the dependency relations computed by PALAVRAS are eminently suitable for the generation of word sketches, there are many minor ways in which PALAVRAS output is incompatible with or insufficient for the demands of a practical lexicographic tool. Thus an extensive post-processing phase takes place to adjust PALAVRAS output and enrich it in a variety of ways.

In order to explicitly represent a variety of dependencies, PALAVRAS deconstructs items such as preposition–article contractions and verbs with infixed pronominal objects. For instance, the contraction *dos* (of the; plural) becomes two separate words (*de os*) with distinct dependencies, while the verb form *levá-lo-á* (will lead you) becomes two separate words (*levará, o*). It was necessary to reconstruct the surface forms lost by PALAVRAS in order that the lexicographer can extract illustrative examples from the corpus with minimal difficulty.

PALAVRAS also treats a wide variety of multi-word units (e.g., compound nouns such as *direitos humanos*, as well as many others) as single items in the dependency structure. Untreated, this would have the unfortunate effect of omitting the component words from each other's word sketches. A simple parser was developed to establish the internal dependency structure and headedness of such units and the result was plugged back into the larger structure with the correct dependencies.

In providing each word with a single governor, PALAVRAS does not explicitly capture relations of importance for complete word sketches. For instance, in the phrase *é viável sua aplicação* (its application is viable), a subject relation is established between *aplicação* and *ser* (*é*). Post-processing adds in the controlled subject relation between *aplicação* and *viável*, information which may be important in the sketch for these two lemmas. In general, a noun phrase subject will get a subject relation to each verb or adjective in an auxiliary sequence (or an object relation if the verb is passive).

Another type of relation that is added is the trinary relation corresponding to a prepositional phrase and its attachment site. PALAVRAS generates binary relations between the preposition and its governor, and between the preposition and its object. Post-processing adds in the composition of these two, so that each full lexical item will appear on the sketch for the other, in a table headed by the preposition.

A similar treatment is followed for coordination. It is often useful for the lexicographer to see the words with which a headword occurs often, for example *arroz e feijão*

(rice and beans). In dependency grammar, the two conjuncts do not stand in a relation to each other so we post-processed to create a relation between the heads of the two conjuncts, so that once again they appear on each other's sketches.

As well as augmenting the relations correctly computed by PALAVRAS with various others, it is desirable to correct some of the decisions made by the parser. Betraying its lack of statistical processing, PALAVRAS often attaches constituents to remote heads in ways that may be linguistically possible but are much less likely than the more proximate attachments. For instance, in the phrase *dedicam-se aos temas contemporâneos* (is dedicated to contemporary themes), PALAVRAS's choice of *dedicar* (*dedicam-se*) as the governor of *contemporâneo* is jettisoned in favour of the much more plausible *tema*.

Finally, for the purpose of collecting as much data as possible within sketches, spelling variations are neutralized in the lemma chosen for each word, with modern Brazilian spelling being used as the standard.

## 6. Regional variants

There are two main regional variants of Portuguese: Brazilian and European. We had corresponding subcorpora within the corpus as a whole and the Sketch Engine provides a keywords function that can list, in order, all words according to how distinctively Brazilian or European they were. A classification of these words is shown in Tables 5.2 through 5.12.

**Table 5.2** Keywords: Geographical adjectives

Brazilian ptTenTen	European ptTenTen
<i>brasileiro</i> (Brazilian)	<i>européu</i> (European)
<i>carioca</i> (from Rio de Janeiro state)	<i>português</i> (Portuguese)
<i>gaúcho</i> (from Rio Grande do Sul state)	<i>cabó</i> (Cape-(Verde))
<i>paulista</i> (from São Paulo state)	<i>euro</i> (also the currency)

Table 5.2 shows geographical keywords. The Brazilian list includes adjectives pertaining to persons born in particular Brazilian states, like *carioca* (from Rio de Janeiro state), *gaúcho* (from Rio Grande do Sul state) and *paulista* (from São Paulo state), whereas the Portuguese list includes a reference to Europe (*européu*) and Cape Verde (*cabó*).

**Table 5.3** Keywords: Administrative divisions

Brazilian ptTenTen	European ptTenTen
<i>bairro</i> (neighbourhood)	<i>freguesia</i> (neighbourhood)
<i>cidade</i> (city), <i>município</i> (city)	<i>concelho</i> (city)
<i>prefeitura</i> (city council)	<i>junta</i> (city council)
<i>estadual</i> (state, adj.)	<i>aldeia</i> (village)
<i>federal</i>	<i>autarquia</i> (autonomous state organ)

The keywords also reflect administrative and governmental terms that are specific to each country (see Table 5.3). The main national administrative levels for both countries are reflected by words for the neighbourhood (*bairro*, Brazil; *freguesia*, Portugal), the county/city/village (*município* and *cidade*, Brazil; *concelho* and *aldeia*, Portugal), the state/province (*estadual* (adj.), Brazil; *distrito*, Portugal) and the federation (*federal* (adj.), Brazil). Terms for local city governments are *prefeitura* (city hall, Brazil) and *junta* (Portugal).

**Table 5.4** Keywords: Administration and politics

Brazilian ptTenTen	European ptTenTen
<i>prefeito</i> (mayor)	<i>autarca</i> (mayor)
<i>delegacia</i> (police station)	
<i>policial</i> (police officer)	
<i>deputado</i> (deputy)	
<i>governador</i> (governor)	
<i>secretário</i> (secretary)	
<i>senado</i> (senate)	
<i>senador</i> (senator)	
<i>vereador</i> (city council member)	
<i>polícia</i> (police)	
<i>secretaria</i> (secretary)	

The political keywords (see Table 5.4) are full of specific Brazilian terms, based on the presidential system of government (*vereador*, *governador*, *deputado*, *senador*, etc.). The only pair that applies to both variants is the one for mayor: *prefeito* (Brazil) and *autarca* (Portugal).

**Table 5.5** Keywords: Grammatical words

Brazilian ptTenTen	European ptTenTen
<i>diante</i> (in front of, in view of)	<i>perante</i> (in front of, in view of)
<i>você</i> (you, 2nd p. sing.)	<i>teu</i> (yours, 2nd p. sing.)
<i>porém</i> (but)	<i>vosso</i> (yours, 2nd p. pl.)
	<i>vós</i> (you, 2nd p. pl.)
	<i>este</i> (this)
	<i>isto</i> (this)
	<i>quer</i> (whether, want [verb])
	<i>aquando</i> (when)

The grammatical keywords (see Table 5.5) have interesting dialectal choices. Some are known differences between the two varieties, such as *vós* (you) and *vosso* (yours), which are common in Portugal but have largely been replaced with *você* and *seus* in Brazil. It is intriguing to note that second-person pronouns are also keywords of Peninsular versus Latin American Spanish (Kilgarriff and Renaud, 2013); in addition, in informal English, the second-person plural has regionally differentiated forms, with ‘y’all’ in the southern US, ‘you guys’ in the Northeast and Canada and ‘you lot’ in Britain.

*Este* (determiner or demonstrative pronoun) and *isto* (demonstrative pronoun), keywords of European Portuguese, are traditionally used to indicate referents that are close to the speaker, as opposed to *esse* and *isso*, which refer to referents that are near the interlocutor. This distinction is still largely observed in Portugal, but is rapidly disappearing in Brazil, where *esse* and *isso* have taken over.

**Table 5.6** Keywords: Business terms

Brazilian ptTenTen	European ptTenTen
<i>diretoria</i> (director's office)	<i>direcção</i> (director's office)
<i>planejamento</i> (planning)	<i>planeamento</i> (planning)
<i>diretor</i> (director, male)	
<i>diretora</i> (director, female)	
<i>gerente</i> (manager)	
<i>assessoria</i> (secretary)	
<i>atendimento</i> (care)	
<i>capacitação</i> (training)	
<i>demanda</i> (demand)	
<i>etapa</i> (phase)	
<i>pauta</i> (agenda)	
<i>treinamento</i> (training)	
<i>vaga</i> (opening)	
<i>convênio</i> (health insurance; agreement)	

The business keywords (see Table 5.6) are predominantly Brazilian, but the words are known in both countries. A number of these are distinct by spelling: *diretor* (male director), *diretora* (female director) and *diretoria* (the director's office) are spelled with an intervening 'c' in Portugal: *director*, *directora* and *direcção*.

**Table 5.7** Keywords: Technology

Brazilian ptTenTen	European ptTenTen
<i>acessar</i> (access)	<i>aceder</i> (access)
<i>celular</i> (cell phone)	<i>telemóvel</i> (cell phone)
<i>usuário</i> (user)	<i>utente</i> (user)
<i>busca</i> (search)	<i>ecrã</i> (screen)
	<i>ficheiro</i> (file)
	<i>utilizador</i> (user)
	<i>ligação</i> (link)
	<i>utilização</i> (use)
	<i>domínio</i> (domain)
	<i>informático</i> (computational)

As previously mentioned, in technology we find many terms that are unique to each country (see Table 5.7). With the exception of *celular* (cell phone, Brazil) and *telemóvel* (mobile phone, Portugal), they are computing words (*acessar*, to access; *usuário*, user, both in Brazil; *aceder*, to access, *ecrã*, screen; *ficheiro*, file; *utente/utilizador*, user; and

*ligação*, link, in Portugal). Brazilian speakers will be more familiar with *compartilhar* (to share) than with *partilhar*, which is preferred in Portugal. They are both used with the sense of sharing information on the web and it is interesting that each variety has selected a different word to express that same meaning, when online communication might suggest otherwise. The online community in both Portugal and Brazil seems to have a set of vocabulary specific to each country, which is revealed by words like *informático* (informatic), which in Brazil would be *computacional* (computational) or *de computador* (\*of computer), or many of the words in the technology grouping, such as *usuário* (user) in Brazil versus *utente* and *utilizador* in Portugal. Other words in this category predate the web, such as *ecrã* (screen) and *ficheiro* (file) in Portugal, which are *tela* and *arquivo* in Brazil, respectively, and are widely known dialectal markers.

**Table 5.8** Keywords: Sports

Brazilian ptTenTen	European ptTenTen
<i>esporte</i> (sport)	<i>desporto</i> (sports)
<i>esportivo</i> (sports [adj.])	<i>desportivo</i> (sports [adj.])
<i>gol</i> (goal)	<i>golo</i> (goal)
<i>equipe</i> (team), <i>time</i> (team)	<i>equipa</i> (team)
<i>rodada</i> (round)	
<i>copa</i> (cup)	

As with technology, in sports (see Table 5.8) each country has a large set of unique terms, a number of which are often cited to illustrate the vocabulary differences between the two dialects. Some of these did show up on the list, like the word for goal (in soccer or similar sports), which Brazilians call a *gol* and the Portuguese *golo*, or for team, which is *time* or *equipe* in Brazil and *equipa* in Portugal. Some of these words are borrowings from either English or French, which in turn explains some of the differences between the two variants in other areas as well, such as computing (see Table 5.7), where Brazilians tend to either adapt or borrow English terms wholesale (e.g., mouse, drive, backup, *deletar* [to delete]), while the Portuguese tend to follow the French terminology (e.g., *ecrã* and *ficheiro* from the French *écran* and *fichier*).

**Table 5.9** Keywords: Weekdays

Brazilian ptTenTen	European ptTenTen
<i>segunda-feira</i> (Monday)	(none)
<i>terça-feira</i> (Tuesday)	
<i>quarta-feira</i> (Wednesday)	
<i>quinta-feira</i> (Thursday)	
<i>sexta-feira</i> (Friday)	

Weekdays turned up as Brazilian keywords (see Table 5.9), which is puzzling as the same words are used in both countries to name the weekdays. In both variants, weekdays are named in an ordinal manner, in such a way that Monday is called the second day (*Segunda-feira*, from the Latin ‘Feria Secunda,’ the second free day in Easter), Tuesday the third day (*Terça-feira*, from the Latin ‘Feria Tertia’), Wednesday

the fourth day (*Quarta-feira*, ‘Feria Quarta’), Thursday the fifth day (*Quinta-feira*, ‘Feria Quinta’) and Friday the sixth day (*Sexta-feira*, ‘Feria Sexta’). The *feira* (from the Latin *feria*, meaning ‘free day’) is optional, so that one may say for example *terça* to mean *terça-feira* (for Tuesday). All of these forms, with the exception of *terça*, are regular ordinal numbers (in the feminine gender) as well. To find out the source of variation, for each dialect, we pulled out all weekday words from the subcorpus word frequency list, excluding *segunda* (considered to be an outlier, as it alone accounted for more than 20 percent of the combined frequencies, probably because of its use as an ordinal numeral), computed their normed counts (per one million words; pmw) and calculated the mean normed frequencies; we then contrasted the means statistically and found a statistical difference for hyphenated words (e.g., *quarta-feira*) but not the unhyphenated ones (e.g., *quarta*). The mean frequency for hyphenated weekdays is higher for Brazil (29.1 pmw, Brazil, 10.8 pmw, Portugal;  $t = 2.516$ ,  $df = 19$ ,  $p = .021$ ), thereby accounting for their keyword status in Brazilian Portuguese. For the unhyphenated forms, the mean is higher for Portugal (13.0 vs. 11.6, Brazil), but not significantly so ( $t = -.323$ ,  $df = 16$ ,  $p = .751$ ). In Halliday’s (1991) terms, this suggests that the system for weekdays in Portugal is equiprobable, whereas in Brazil, it is heavily skewed in favour of the full form (71 percent).

Tables 5.11, 5.12 and 5.13 present nouns, adjectives, verbs and adverbs that turned up as markers of each dialect but did not fit neatly into the previous categories.

**Table 5.10** Keywords: Adjectives and adverbs

PoS	Brazilian ptTenTen	European ptTenTen
Adjective	<i>ruim</i> (bad) <i>grosso</i> (thick, uncouth)	<i>elevado</i> (high) <i>habitual</i> (habitual) <i>respectivo</i> (respective) <i>secundário</i> (secondary) <i>vasto</i> (vast)
Adverb	<i>demais</i> (too) <i>principalmente</i> (mainly) <i>somente</i> (only) <i>inclusive</i> (including)	<i>demasiado</i> (too) <i>sobretudo</i> (mainly, moreover) <i>designadamente</i> (namely) <i>nomeadamente</i> (namely) <i>igualmente</i> (equally) <i>relativamente</i> (relatively)

Adverb keywords (see Table 5.10) reflect interesting choices. For instance, *designadamente* and *nomeadamente* are used in Portugal to mean roughly ‘namely’ but are very rare in Brazil. Brazilian Portuguese lacks an immediate equivalent and Brazilian speakers typically paraphrase this with expressions such as *isto é* or *a saber*. The words *principalmente* (Brazil) and *sobretudo* (Portugal; both meaning ‘mainly’) are subtle markers of each dialect; the fact that they turned up as keywords demonstrates the strength of both the corpus and the comparative approach.

The noun keywords (see Table 5.11) include many that can be accounted for by minor spelling differences: the Brazilian *controle* (control) is spelled *controlo* in Portugal. *Bilhão*, in Brazil, is spelled *bilião* in Portugal, but the Brazilian word

**Table 5.11** Keywords: Nouns

<b>Brazilian ptTenTen</b>	<b>European ptTenTen</b>
<i>registro</i> (registration)*	<i>registo</i> (registration)*
<i>bilhão</i> (billion)	<i>altura</i> (height)
<i>chance</i> (chance)	<i>aposta</i> (bet)*
<i>controle</i> (control)*	<i>castelo</i> (castle)
<i>disputa</i> (dispute)*	<i>cento</i> (one hundred; (per) cent)
<i>fazenda</i> (farm)	<i>colaboração</i> (collaboration)
<i>foco</i> (focus)*	<i>concerto</i> (concert)
<i>implantação</i> (implementation)	<i>contributo</i> (contribution)
<i>integrante</i> (member)	<i>deslocação</i> (movement)
<i>mato</i> (bushes)*	<i>dimensão</i> (dimension)
<i>mina</i> (mine)*	<i>elemento</i> (element)
<i>mídia</i> (media)	<i>face</i> (face)
<i>palestra</i> (conference)	<i>gama</i> (range)*
<i>programação</i> (program, schedule)	<i>intervenção</i> (intervention)
<i>renda</i> (income)*	<i>nível</i> (level)
<i>rodovia</i> (highway)	<i>percurso</i> (journey)
<i>show</i>	<i>pormenor</i> (particular)
<i>trecho</i> (stretch; leg of a trip)	<i>procura</i> (search)*
<i>investigador</i> (researcher)	<i>recolha</i> (collection)*
<i>edifício</i> (building)	<i>restante</i> (remainder)
<i>zona</i> (zone)	<i>vertente</i> (aspect)
<i>investigação</i> (research)	<i>acção</i> (action)
<i>cotidiano</i> (everyday)	<i>controlo</i> (control)*
<i>morador</i> (dweller)	<i>facto</i> (fact)
<i>pesquisador</i> (researcher)	<i>regresso</i> (return)*
<i>pesquisa</i> (research/search)	
<i>item</i>	

\*Also a verb form

(borrowed from the American system) means 10<sup>9</sup>, which in turn is *mil milhões* (a thousand million) in Portugal, whereas the Portuguese *bilião* (inspired by the French system) means 10<sup>12</sup> and is a *trilhão* in Brazil. The Brazilian *mídia* (media) is spelled *média* in Portugal. Other examples include *planejamento* (planning, Brazil), *planeamento* (Portugal); *registro* (registration/registry, Brazil), *registo* (Portugal); *convênio* (agreement, Brazil), *convénio* (Portugal); *acção* (action, Portugal), *ação* (Brazil); and *facto* (fact, Portugal), *fato* (Brazil). Other keywords are motivated by suffixation: *contributo* (contribution, Portugal) is *contribuição* (Brazil) whereas *deslocação* (movement, Portugal) is *deslocamento* (Brazil). Other nouns are lexical choices, such as *pesquisador* in Brazil but an *investigador* in Portugal (which, in Brazil, would be more readily associated with a police detective) and *fazenda* (farm) in Brazil compared to a *quinta* in Portugal.

The verb keywords (see Table 5.12) reveal lesser-known dialectal choices. One of these has do to with ‘highlight’ words: *ressaltar* (to highlight) is more typical of

**Table 5.12** Keywords: Verbs

Brazilian ptTenTen	European ptTenTen
<i>registrar</i> (to register)	<i>registar</i> (to register)
<i>atender</i> (to take care of)	<i>atribuir</i> (to assign)
<i>atuar</i> (to work)	<i>calhar</i> (to come in handy)
<i>buscar</i> (to search)	<i>constituir</i> (to constitute)
<i>cobrar</i> (to charge; to demand)	<i>contatar</i> (to contact)
<i>conversar</i> (to talk)	<i>efetuar</i> (to carry out)
<i>encaminhar</i> (to send)	<i>equipar</i> (to outfit)
<i>firmar</i> (to sign)	<i>gerir</i> (to manage)
<i>liberar</i> (to free)	<i>permitir</i> (to allow)
<i>ocorrer</i> (to happen)	<i>pretender</i> (to intend)
<i>planejar</i> (to plan)	<i>recordar</i> (to remember)
<i>repassar</i> (to transfer)	<i>referir</i> (to refer)
<i>pegar</i> (pick up)	<i>apanhar</i> (to pick up)
<i>retornar</i> (return)	<i>regressar</i> (to return)
<i>ampliar</i> (to widen; to increase)	<i>pôr</i> (to put)
<i>morar</i> (to live)	<i>meter</i> (put)
<i>compartilhar</i> (to share)	<i>partilhar</i> (to share)
<i>ressaltar</i> (to highlight)	<i>assinalar</i> (to highlight)
	<i>sublinhar</i> (to highlight)
	<i>salientar</i> (to highlight)
	<i>realçar</i> (to highlight)
	<i>alargar</i> (to widen)
	<i>situar</i> (to be at; to place)
	<i>proceder</i> (to proceed)
	<i>arranjar</i> (to get)
	<i>distinguir</i> (to distinguish)
	<i>decorrer</i> (to follow)

Brazilian Portuguese whereas the near synonyms *assinalar*, *sublinhar*, *salientar* and *realçar* (all meaning ‘to highlight’ or ‘to underscore’) are more common in Portugal. *Meter* (to put) is common in European Portuguese, but less so in Brazil, where it can have a rude connotation, generally meaning pushing or forcing something into a place. Verbal equivalencies also include words meaning (a) to pick up: *pegar* (Brazil), *apanhar* (Portugal); (b) to return: *retornar* (Brazil), *regressar* (Portugal); (c) to widen: *ampliar* (Brazil), *alargar* (Portugal); and (d) to register: *registrar* (Brazil), *registar* (Portugal). Keywords motivated by spelling include the Brazilian choices *atuar* (to act; which is *actuar* in Portugal) and *planejar* (to plan, *planear* in Portugal).

One tool that the lexicographer can use to explore the keywords in Sketch Engine is Sketch-Diff, which enables the researcher to compare the collocates sets of different lemmas or to contrast the collocates of a single lemma across different subcorpora. To illustrate, when we compared the lemma *partilhar* (to share) in Brazilian versus European meanings, an inspection of the Sketch-Diff showed that the Brazilian meaning is restricted to the division of property in divorce or in a will, including

collocates such as *judicial*, *amigável* (amicable), *consensual*, *tabelião* (notary) and *cartório* (county clerk); meanwhile, the European meaning appeared far broader, including *informação* (information), *paixão* (passion), *visão* (view) and *ideias* (ideas) in addition to the same legal meaning as in Brazil. With *compartilhar*, the reverse was true. Although both varieties shared a handful of collocates, these tended to be abstract nouns (*alegria*, happiness; *experiência*, experience; *opinião*, opinion; etc.), with concrete nouns being exclusively Brazilian (*toalha*, towel; *seringa*, syringe; *talher*, cutlery; etc.).

## 7. Conclusion

We have presented our experience in ‘setting up for corpus lexicography’ for Portuguese, including building a corpus from the web, cleaning it, removing duplicates, parsing it, loading it into a corpus tool and preparing word sketches from it. We have also presented an account of the two major varieties of the language as represented by the two parts of the corpus that we collected.

## References

- Barlow, M. (2000), *MonoConc Pro*. Houston, TX: Athelstan.
- Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009), ‘The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora.’ *Journal of Language Resources and Evaluation*, 43, (3), 209–26.
- Bick, E. (2000), *The Parsing System PALAVRAS – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD Dissertation, Århus University.
- Christ, O. and Schulze, M. (1994), *The IMS Corpus Workbench: Corpus Query Processor (CQP)*. Stuttgart: University of Stuttgart.
- Halliday, M. A. K. (1991), ‘Corpus studies and probabilistic grammar’, in K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, pp. 30–43.
- Jakubíček, M., Kilgarrieff, A., Kovář, V., Rychlý, P. and Suchomel, V. (2013), ‘The TenTen Corpus Family’, in *Proceedings of the International Corpus Linguistics Conference (Lancaster)*.
- Kilgarrieff, A. and Renau, I. (2013), ‘esTenTen, a vast web corpus of Peninsular and American Spanish’, in *Proceedings of the V International Conference on Corpus Linguistics (Alicante)*, pp. 12–19.
- Kilgarrieff, A., Rychlý, P., Smrz, P. and Tugwell, D. (2004), ‘The Sketch Engine’, in *EURALEX Lorient Proceedings*, pp. 105–15.
- Rundell, M. and Kilgarrieff, A. (2011), ‘Automating the creation of dictionaries: Where will it all end?’, in F. Meunier, S. De Cock, G. Gilquin and M. Paquot (eds), *A Taste for Corpora: In honour of Sylviane Granger*. Amsterdam, Philadelphia, PA: John Benjamins, pp. 257–81.
- Santos, F., Freitas, C., Oliveira, H. and Carvalho, P. (2008), ‘Second HAREM: New challenges and old wisdom’, in A. J. d. S. Teixeira, V. L. Strube de Lima, L. Caldas

- de Oliveira and P. Quaresma (eds), *Proceedings of Computational Processing of the Portuguese Language (PROPOR 2008)*, pp. 212–5.
- Scott, M. (2008), *WordSmith Tools*. Liverpool: Lexical Analysis Software.
- Sharoff, S. (2006), 'Creating general-purpose corpora using automated search engine queries', in M. Baroni and S. Bernardini (eds), *Wacky! Working Papers on Web as Corpus*. Bologna: Gedit, pp. 63–98.

## Notes

- 1 <http://code.google.com/p/chared>
- 2 <http://code.google.com/p/justext>
- 3 <http://code.activestate.com/recipes/326576> language detection using character trigrams

