4

The *Corpus do Português* and the Frequency Dictionary of Portuguese

Mark Davies

1. Introduction

In this chapter, we will focus on two corpus-related tools that we have developed during the past seven to eight years – the *Corpus do Português* (freely available at www.corpusdoportugues.org) and the *Frequency Dictionary of Portuguese* (Davies and Preto-Bay, 2007), which is based on the corpus.

The *Corpus do Português* was created by Mark Davies (Brigham Young University, USA) and Michael Ferreira (Georgetown University, USA) and it uses the same architecture and interface as the other corpora from corpus.byu.edu (see Davies, 2010b). These corpora include the 450 million word Corpus of Contemporary American English (COCA; see Davies, 2009), the 400 million word Corpus of Historical American English (COHA; see Davies, 2012) and the 100 million word *Corpus del Español* (see Davies, 2010a; Davies, 2008; Davies, 2005b; Davies, 2005c).

The *Corpus do Português* is composed of 45 million words from the 1100s to 1999 and the contemporary Portuguese texts are balanced between spoken, fiction, newspapers and academic texts. While there are certainly much larger corpora of Portuguese, we believe that the architecture and interface for the *Corpus do Português* allow researchers to carry out interesting and useful queries, many of which are not possible with other corpora of Portuguese.

The second resource – the *Frequency Dictionary of Portuguese: Core Vocabulary for Learners* (Davies and Preto-Bay, 2007) – was created by Mark Davies and Ana-Maria Raposo Preto-Bay, both from Brigham Young University. The frequency dictionary is based on the 5000 most frequent lemmas in the 1900s portion of the *Corpus do Português* and for each lemma it provides frequency statistics, an English gloss, a sample sentence from the corpus and a translation of that sentence into English. It also contains more than thirty thematically-oriented and frequency-based lists of Portuguese words, including words in specific semantic fields, words relating to grammar issues that are relevant to non-native speakers and lists dealing with genre-based, dialectal and historical differences in the Portuguese lexicon.

In the sections that follow, we first discuss the *Corpus do Português*, paying particular attention to the range of queries that the corpus allows. We then conclude with a somewhat shorter discussion of the *Frequency Dictionary of Portuguese*.

2. *Corpus do Português* – texts

The *Corpus do Português* was designed from the ground up as a corpus that could shed light on several interesting types of variation – historical, dialectal and genre-based variation.

In terms of historical variation, there are approximately 15 million words from the 1100s–1700s, another 10 million words from the 1800s and 20 million words from the 1900s. Table 4.1 shows the sources for the texts from the 1100s–1800s.

Although some of these materials were already in electronic format online, approximately 6 million of the total of 25 million words had to be either keyboarded or scanned. In terms of the scanned materials, we had to compare these texts with

Table 4.1 *Corpus do Português*: texts from the 1100s–1800s

Name	Time period	# texts	# words*
Banco de Dados de História Literária (UFSC)	1800s	348	8,335,407
Scanned books	1100–1899	151	4,180,029
Corpus lexicográfico do Português da Universidade de Aveiro (Telmo Verdelho / João Paulo Silvestre)	1300–1900	42	3,250,147
Corpus Informatizado do Português Medieval (M. Francisca Xavier)	1100–1504	40	2,002,233
Univ. de Aveiro: Textos escaneados (Telmo Verdelho / João Paulo Silvestre)	1300–1899	38	1,922,234
Corpus eletrônico de textos históricos da Georgetown University (Michael J. Ferreira)	1360–1702	21	1,184,941
Corpus Clássicos	1800s	15	997,045
Décadas de Ásia [CD-ROM]	1500s	3	638,834
Corpus electrónico de textos históricos de Braga (Michael J. Ferreira / Brian F. Head)	1489–1570	12	565,372
Corpus electrónico de textos históricos de Coimbra (Evelina Verdéelo)	1301–1724	9	491,239
Biblioteca Nacional Digital	1800s	8	362,512
Tycho Brahe Parsed Corpus of Historical Portuguese	1538–1889	7	324,626
Corpus electrónico de forais de Vila Real (Olinda Santana)	1496–1500	4	164,600
Total 1300s–1800s		690	26,165,252

* Word totals include some punctuation, which create a figure a little higher than the 25 million words from this time period.

Table 4.2 *Corpus do Português*: texts from the 1900s

Content	# texts	# words
Portugal	30,708	12,002,965
Spoken	1,525	1,296,407
Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)	968	428,826
Interviews in *Jornal de Notícias* (Lisboa)	150	248,438
CRPC – *Corpus de Referência do Português Contemporâneo*	139	136,511
Interviews in *Público* (Lisboa)	63	86,784
Instituto Camões: Geografia da Língua Portuguesa	92	53,641
Interviews in *Terras da Beira* (Guarda)	23	2,372
Português Falado – Variedades Geográficas e Sociais	30	40,477
Interviews on the Web	60	259,358
Fiction	155	3,592,059
Scanned novels and short stories	155	3,592,059
Univ. de Aveiro: Scanned texts	3	134,375
Newspapers (1996-97)	10,373	3,648,974
Público (Lisboa)	7088	1,103,964
Expresso (Lisboa)	865	702,841
Terras da Beira (Guarda)	1089	695,068
Jornal de Noticias (Lisboa)	518	646,695
Região de Leiria (Leiria)	813	500,406
Academic	18,655	3,465,525
Enciclopédia Universal	18,655	3,465,525
Brazil	25,141	12,009,402
Spoken	519	1,266,088
Interviews in *O Estado de São Paulo* (São Paulo)	266	356,382
Interviews in *Gazeta do Povo* (Curitiba)	49	54,092
Interviews in *A Tarde* (Bahia)	47	42,139
Interviews in *Jornal de Comércio* (Recife)	31	37,382
Interviews in *Jornal da Cidade* (São Paulo)	22	28,202
Interviews downloaded from the Web	53	265,870
A linguagem falada culta na cidade de São Paulo. São Paulo: T. A. Queiroz	20	181,598
A linguagem falada culta na cidade do Recife. Recife: UFPE	31	300,423
Fiction	87	3,612,996
Scanned novels and short stories	51	2,440,183
Banco de Dados de História Literária (U Fed Santa Catarina)	26	947,637
LacioWeb	10	225,176
Newspapers (1996-97)	19,274	3,765,429
Folha de São Paulo (São Paulo)	12,096	1,141,478
Correio do Povo (Porto Alegre)	1,463	515,043
O Estado de São Paulo (São Paulo)	842	508,609
Gazeta do Povo (Curitiba)	503	505,318
Diário de Pernambuco (Recife)	941	500,574
Agência do Estado (Santa Catarina)	3,275	495,118
A Tarde (Bahia)	154	99,289
Academic	5,261	3,364,889
Enciclopédia da Folha	4,978	1,685,187
LacioWeb	283	1,679,702
Total	55,849	24,012,367

the tens of thousands of pages of printed text in order to correct for scanning errors, which took approximately 200–250 hours.

The corpus from the 1900s was designed so that it represented a balanced collection of texts from Brazil and Portugal (10 million words each) as well as genres (2 million spoken, 6 million fiction, 6 million newspapers and magazines, and 6 million academic). Table 4.2 provides an overview of the texts from the 1900s.

In most cases, the electronic texts were converted from the original format (PDF, HTML, electronic book format, etc.) to plain text files, so that they could be input into the database. However, about six million words of text (primarily novels and short stories) had to be scanned and then corrected, which took approximately 200 hours.

3. *Corpus do Português* – importance of genre balance

We created the contemporary portion of the *Corpus do Português* so that it was balanced between spoken, fiction, newspapers and academic texts. Of course, it would have been much easier to create a much larger corpus composed strictly of newspaper texts, which are easily available on the Web or in other electronic formats. In this way, we could have created a corpus with hundreds of millions of words of text with relative ease. However, had we done so, we would have created a corpus that only represents a small slice of Portuguese.

Let us consider the importance of genre balance from two different angles. First, imagine that we had created a corpus that consisted only of newspapers. Of the top 3000 lemmas in the 20 million words of text from the 1900s in the *Corpus do Português*, 895 of these – nearly 30 per cent – are at least 50 per cent more frequent (per million words) in the newspaper section of the corpus (6,618,000 words) than in the corpus as a whole. In other words, if the corpus were composed strictly of newspapers, we would have the illusion that these words are much more common in Portuguese than they really are, if we had a balanced corpus.

Some examples of these words are:

Noun: *portagem, petista, autarca, meio-campo, montadora, goleiro, israelense, sucursal, damasceno, ex-governador, fim-de-semana, terça-feira, atacante, quarta-feira, relator, seguradora, ex-prefeito, tucano, precatório, autarquia*;

Verb: *empatar, comemorar, sublinhar, homenagear, reafirmar, alertar, afirmar, contactar, divulgar, disputar, frisar, aprovar, disponibilizar, alegar, argumentar, responsabilizar, informar, votar, descartar, apurar*;

Adjective: *governista, distrital, manifestante, autárquico, salarial, gaúcho, atlético, desportivo, municipal, estatal, estadual, policial, parlamentar, mensal, regional* (as well as adjectives derived from place names like *paranaense, portista, palestino, pernambucano, timorense, paulista* and *baiano*);

Adverbs: *anteontem, ontem, designadamente, curiosamente, diariamente, publicamente, recentemente, politicamente, oficialmente, obviamente*.

Another way of looking at the importance of genre balance is to imagine what the corpus would look like *without* a given genre. For example, imagine that – because of the time and effort required to collect fiction texts – we had simply decided to omit fiction entirely, hoping that the other genres would cover for this absence. But we find that in the *Corpus do Português*, of the top 10,000 nouns, verbs, adjectives and adverbs, more than 10 per cent (1093 lemmas) occur at least 80 per cent of the time in fiction (even though fiction is only about 30 per cent of the corpus). In other words, without fiction, researchers would have very little evidence for this 10 per cent of the

Table 4.3 Fiction words in the *Corpus do Português*

PoS	Words
Noun	*brancura* (whiteness) 95/95, *saleta* (room) 208/209, *regaço* (lap) 113/114, *traste* (junk) 105/106, *arranco* (pull) 93/94, *reposteiro* (curtain) 86/87, *espaldar* (back of a chair) 86/87, *soluço* (hickup) 290/296, *negrume* (blackness) 90/92, *sussurro* (whisper) 89/91, *capataz* (foreman) 155/159, *furriel* (soldier) 154/158, *beiço* (lip) 292/300, *pasmo* (awe) 141/145, *gemido* (whine) 237/244, *turba* (mob) 101/104, *tropeiro* (cowboy) 167/172, *mana* (sister) 353/364, *casebre* (shack) 192/198, *sacristão* (sacristan) 286/295
Verb	*resplandecer* (shine) 88/88, *sorver* (sip) 150/151, *empalidecer* (turn pale) 109/110, *entreabrir* (half open) 101/102, *emudecer* (silence) 89/90, *fitar* (stare) 610/618, *volver* (return) 296/300, *segredar* (tell secrets) 145/147, *resmungar* (mutter) 267/271, *altear* (rise) 100/102, *murmurar* (whisper) 829/846, *estacar* (stake) 181/185, *enxotar* (send out) 86/88, *soluçar* (sob) 196/201, *aquietar* (calm down) 111/114, *bocejar* (yawn) 109/112, *espiar* (spy) 290/298, *afagar* (stroke) 177/182, *gaguejar* (stutter) 100/103, *rosnar* (growl) 162/167
Adjective	*doirado* (golden) 174/174, *absorto* (aloof) 103/104, *lívido* (pale) 164/166, *vagaroso* (slow) 108/110, *dorido* (painful) 103/105, *pensativo* (pensive) 181/185, *desdenhoso* (contemptuous) 86/88, *hirto* (stiff) 154/158, *trémulo* (trembling) 317/326, *entreaberto* (ajar) 123/127, *encardido* (soiled) 87/90, *sôfrego* (anxious) 87/90, *cheiroso* (fragrant) 96/100, *abafado* (stuffy/muffled) 208/217, *debruçado* (hunched over) 131/137, *voluptuoso* (voluptuous) 109/114, *fatigado* (weary) 135/142, *lúgubre* (gloomy) 96/101, *estirado* (stretched) 91/96, *bêbedo* (drunk)127/134
Adverb	*vagarosamente* (slowly) 167/173, *nervosamente* (nervously) 82/85, *docemente* (gently) 113/118, *instintivamente* (instinctly) 116/125, *decerto (surely)* 628/677, *vagamente* (vaguely) 228/250, *demoradamente* (at length) 100/111, *suavemente* (gently) 125/142, *bruscamente* (suddenly) 193/221, *devagar* (slowly) 637/731, *repentinamente* (suddenly) 132/152, *adentro* (into) 158/182, *cautelosamente* (carefully) 82/95, *alegremente* (happily) 95/114, *longamente* (thoroughly) 161/195, *deveras* (rather) 117/142, *subitamente* (suddenly) 355/432, *timidamente* (shyly) 92/112, *vivamente* (strongly) 86/105.

vocabulary of Portuguese. Table 4.3 gives some examples of these words and shows what portion of the total tokens come from fiction (e.g., *regaço* (lap): 113 of the 114 tokens are from fiction).

In summary, a corpus of Portuguese – or any other language for that matter – that does not rely on a wide range of genres may provide a very skewed picture of that language. The *Corpus do Português* was designed from the ground up as a corpus that would cover this full range of genres and it thus provides a more accurate view of the language.

4. *Corpus do Português* – annotating the texts

Once the texts were collected and corrected, we then annotated them by lemmatizing the texts and tagging them for part of speech. For contemporary Portuguese, this was relatively easy, since there are many tools and lexicons available, such as Tree-Tagger. For the older stages of the language, however, where spelling variation is much more of an issue, lemmatization and tagging were obviously much more difficult.

For example, in the corpus there are more than 300 distinct forms of the lemma *fazer* (to do/make) that occur from the 1300s to the 1600s, but which are not found in Modern Portuguese. The top fifty such forms (with their frequency in the 1300s–1600s) are *fezesse* 1052, *fazẽ* 1040, *fecto* 824, *fezera* 804, *ffazer* 727, *fezerom* 658, *fezo* 553, *fezerõ* 538, *fezer* 493, *feytos* 464, *façades* 429, *fezessem* 416, *faras* 401, *fectos* 379, *fecta* 377, *fezeram* 302, *fezeron* 291, *façã* 284, *ffez* 250, *feytas* 246, *fazẽdo* 220, *fizerom* 214, *faziã* 208, *fezerem* 186, *fazía* 183, *fezestes* 180, *fezese* 174, *fezeste* 166, *façõo* 150, *fazendo* 142, *fectas* 135, *fezerão* 134, *fazíã* 131, *ffeito* 131, *fezerẽ* 124, *farya* 110, *fizese* 104, *ffaz* 101, *ffecto* 101, *ffeyto* 101, *féz* 100, *fazerẽ* 96, *fezerã* 88, *faredes* 85, *fazedes* 84, *fezessẽ* 84, *fazee* 77, *ffiz* 76, *fezerdes* 75 and *fezesem* 73.

Since none of these forms is found in a lexicon of Modern Portuguese and cannot therefore be automatically tagged or lemmatized, we had to annotate each of these types manually and this process was repeated for tens of thousands of other lemmas besides *fazer*. This was accomplished by displaying up to 50 tokens of each type in a web-based interface along with the surrounding context and then having native speakers of Portuguese identify the 'modern' form of the older form (when available) and in this way the older form would 'acquire' the features of the modern form, such as lemma and part of speech tag. Overall, there were more than 400,000 distinct forms (types) that had to be manually annotated and this process took the six native speakers more than a year of intensive work.

5. *Corpus do Português* – lexical and grammatical queries

Having created what we believe is a well-balanced textual corpus, the next task was to create an architecture and interface that would allow for a wide range of queries. In this section, we will provide brief examples of how researchers can use the *Corpus do*

Português to search by word, phrase, wildcard, lemma, part of speech, or any combination of these. In Part 6, we will also look at semantically-oriented queries, including collocates (both by raw frequency and by Mutual Information score) and using synonyms and customized word-lists. Finally, in Part 7 we will provide some examples of how the interface can be used to look at variation across time, across genres, and in different dialects (Portugal and Brazil).

At the most basic level, users can search for a word or phrase and see its (normalized) frequency across time, in Portugal and Brazil, and in the four main genres of spoken, fiction, newspapers and academic. (Note also that users have a choice of corpus interface – either Portuguese or English.) For example, consider the word [*cujo*] (whose) (the brackets allow us to search for all forms of the word, e.g., *cujo*, *cuja*, *cujos* and *cujas*). As Figure 4.1 shows (note the PER MILlion row), the use of *cujo* has been decreasing over the past 300 to 400 years, it is used slightly less in Brazil than in Portugal and there is a huge difference in its use in different genres of Portuguese – for example, it is used about six times as frequently in academic texts as in spoken Portuguese.

Figure 4.1 shows the overall frequency by century, dialect and genre. However, it is also possible to see the frequency of each individual form – either overall in all selected sections of the corpus, or in the different sections. For example, Figure 4.2 shows the frequency of infinitival forms of verbs starting with *des** and the shades of grey indicate the normalized frequency (darker is more frequent).[1]

Users can click on the numbers in any of the cells to see the Keyword in Context display (KWIC). For example, Table 4.4 shows a few of the entries for *descobrir* (to discover, to find out) in the 1900s. While the entries are centre-aligned in the

SECTION	1300s	1400s	1500s	1600s	1700s	1800s	1900s	PORT	BRAZ	ACAD	NEWS	FICT	ORAL
FREQ	238	845	1610	1637	1040	4434	6457	3394	3063	2712	1609	1890	166
PER MIL	129.45	297.05	371.55	500.28	475.01	455.40	318.64	332.24	304.82	471.28	260.29	318.32	79.68

Figure 4.1 Frequency of *cujo* (whose) by century, dialect and genre

	WORD	TOT	1300	1400	1500	1600	1700	1800	1900	PORT	BRAZ	ACAD	NEWS	FICT	ORAL
1	DESCOBRIR	2022		64	292	120	99	570	856	386	470	179	220	377	80
2	DESCER	1209		1				462	643	358	285		90	462	57
3	DESENVOLVER	991	2		2	5	5		912	507	405	412	317	31	152
4	DESCANSAR	703						322			138	15		174	34
5	DESTRUIR	911			130	100	98	215	350	211	139	113	89	110	38
6	DESPERTAR	590	1	2	8	10	10	255	304	140	164		90	136	
7	DESAPARECER	539			9	7	4	183	336	199				180	
8	DESCREVER	560		6	8			135	368	164	204	243			35
9	DESEJAR	1005		76	212	177	73	230						103	
10	DESIGNAR	419				3		50	366	267		306	34	18	8
11	DESVIAR	468		15	21	56		187				27		96	11
12	DESFAZER	614	58		106	95						20		99	

Figure 4.2 *des** verbs by century, dialect and genre

Table 4.4 Sample Keyword in Context (KWIC) for *descobrir* (to discover, to find out)

Source		Keyword in Context
1	19:Fic:Br	*Queirós: Dora ele tinha medo e raiva, se ajoelhar no terreiro, rasgar a camisa e <descobrir> seus pecados. Senhora, eu sei, só havia de querer abafar tudo.*
2	19N:Br:SP	*e a empregada doméstica que os serve. O marido, artista plástico, tenta <descobrir> o segredo das cores, mas só produz anêmicas naturezas-mortas, encaramujando*
3	19Ac:Pt:Enc	*da criatura viva e não uma substância separável dela. Acreditava que o intelecto consegue <descobrir> o universal, a partir das impressões dos sentidos, e que a alma é*
4	19N:Br:Folha	*o crime, enfileirou os corpos e fugiu. A polícia demorou um mês para <descobrir> os corpos e passou 18 anos sem pistas de o criminoso. Em 1989,*
5	19N:Pt:Beira	*a ETAR, desviando-o, possivelmente, para local que o PSD «não conseguiu <descobrir> por causa do matagal circundante», justifica Carlos Gonçalves.*
6	19Ac:Pt:Enc	*num discurso na praia do Restelo, a expedição que nesse momento partia para <descobrir> o caminho marítimo para a Índia e os motivos que a inspiravam*
7	19N:Pt:Leira	*de Braga, a maioria dos católicos em Portugal «ainda não está preparada para <descobrir> a liturgia como fonte da renovação da vida», pelo que « a catequese de*
8	19:Fic:Pt:Ventura	*Vergonha e a autoconfiança e certeza das suas possibilidades tornavam viável. AMIRO estava então a <descobrir> os limites da sua própria coragem, a desvendar as barreiras*

web interface, because of space limitations in this printed version we show them in paragraph format here. Via the web interface, users can also save KWIC lines to different lists that they have created (for later use) and expand the context to about 200 words.

It is also possible to see a standard Keyword in Context (KWIC) display for any word, phrase, or construction directly and to re-sort by surrounding words. For example, Figure 4.3 shows a handful of entries for *relação* (relation / relationship).[2] In the previous examples, we searched just for individual words or substrings. But because the corpus is lemmatized, we can also search by lemma and, of course, we can search for entire phrases. For example, Table 4.5 shows the first few entries for *[fazer] parte de* (to be part of). Note that in this case we have the combined total for all dialects and genres from the 1900s, but (as with Figure 4.2) we could also show the individual frequencies for these sections of the corpus.

Finally, the corpus is tagged for part of speech, which allows for interesting syntactically-oriented queries. For example, consider the construction [*ter*] *que* [vr*] (have

ue se reconheceu completamente a	relação	entre cracas e outros crustáceos . A
ım objeto para melhor considerar a	relação	entre seus elementos , entre os qua
, não é necessário que se defina a	relação	de dominância (DOM) , pois esta p
Em consequência , desapareceria a	relação	construída aos poucos entre os técni
ıto , com o objectivo de melhorar a	relação	entre sustentação e resistência ao a
Ucanha , não escutara o mocinho a	relação	destes sucessos , e quantas e quant
ırtigo , quando nos contaram que a	relação	desta corte tinha absolvido ao procu
ıdo meio século de estudos sobre a	relação	dos preços com a vida dos povos , h
de que a tecnologia e , portanto , a	relação	entre os insumos , não pode ser afe
ı - e diferentemente dos demais - a	relação	entre texto de partida e texto de che
ındial , pesquisados e checados . A	relação	inclui , além do endereço para acess
marcadores discursivos para cada	relação	retórica , seguem a proposta de Mar
rase . O mesmo pode-se dizer com	relação	à existência da preposição 'a' , deve
e Mendonça os seus projectos com	relação	ao seu futuro casamento : que havia
taram estatísticas semelhantes com	relação	à média aritmética e às medidas de

Figure 4.3 Re-sortable Keyword in Context (KWIC) entries: *relação* (relation/relationship)

to). As with Table 4.5, we can find the frequency of all 2,278 matching strings in the corpus from the 1800s to the 1900s (the most frequent being *tem que ser* (have to be) 337 tokens), *tem que ter* (~ have) 157, *tem que fazer* (~ do/make) 100, *tem que ver* (~ see) 93, *tinha que ser* (had to be) 75, *têm que ser* (have [pl.] to be) 74, *tem que estar* (have [sing.] to be) 60 and *tinha que fazer* (had to do/make) 60). Perhaps the more interesting view, though, is the combined frequency of all forms. Figure 4.4 shows that the construction is clearly increasing over time (e.g., about four times as frequent in the 1900s as in the 1800s), that it is used a bit more in Brazilian than in European Portuguese and that it is used much more in spoken Portuguese than in more formal genres.

Table 4.5 Forms of [*fazer*]+ parte de

	Phrase	Total
1	*Faz parte de*	170
2	*Fazem parte de*	94
3	*Fazer parte de*	63
4	*Fazia parte de*	35
5	*Fez parte de*	25
6	*Fazendo parte de*	18
7	*Faziam parte de*	15

Another example might be the construction [*deixar*] *eu|ele|ela|eles|elas* [vr*] – meaning any form of *deixar* (to let), followed by one of the five pronouns listed (notice the subjective case), followed by an infinitival form of the verb. The most frequent forms of the construction are *deixa eu ver* (let me see) (24 tokens), *deixa eu tomar* (let me take), *deixa eu dormir* (let me sleep), *deixa ele falar* (let him speak), *deixa ele vir*

SECTION	1300s	1400s	1500s	1600s	1700s	1800s	1900s	PORT	BRAZ	ACAD	NEWS	FICT	ORAL
FREQ	2	19	91	107	47	648	4943	2070	2873	447	1503	1079	1914
PER MIL	1.09	6.68	21.00	32.70	21.47	66.55	243.93	202.63	285.91	77.68	231.63	181.73	918.74

Figure 4.4 [*tem*] *que* [vr*] (e.g., *tem que sair*, 'has to leave') by century, dialect and genre

(let him come), *deixa eu fazer* (let me do/make), *deixe eu ir* (let me go), *deixar ele ir* (let him go), *deixa eu entrar* (let me come in) and *deixa eu ir* (let me go). As Figure 4.5 indicates, the construction is increasing over time (4–5 times as frequent in the 1900s as the 1800s), is used much more in Brazil than in Portugal and is used much more in informal genres (especially spoken).

SECTION	1300s	1400s	1500s	1600s	1700s	1800s	1900s	PORT	BRAZ	ACAD	NEWS	FICT	ORAL
FREQ	0	3	3	0	0	9	89	5	84	3	8	35	43
PER MIL	0.00	1.05	0.69	0.00	0.00	0.92	4.39	0.49	8.36	0.52	1.23	5.89	20.64

Figure 4.5 [*deixar*] pron [vr*] (e.g., *deixa eu ver*, 'let me see') by century, dialect and genre

6. *Corpus do Português* – semantically-based queries

In addition to using the corpus to look at the frequency of words, phrases and constructions, we can also use it to examine the meaning of words, primarily through an investigation of collocates, or 'nearby words'. Unlike some other corpus interfaces, where collocates are almost an afterthought, with the *Corpus do Português*, one can find the collocates of words and phrases with just one click.

For example, users could find the collocates of *beber* (to drink; all forms) by simply entering [*beber*] in the Word field and then selecting [Noun] for the collocates field. They would then see that the most frequent collocates (set to four words to the right of *beber*) are: *água* (water) 206 tokens, *vinho* (wine) 139, *copo* (glass) 80, *café* (coffee) 60, *cerveja* (beer) 59, *saúde* (health) 58, *gole* (sip) 55, *leite* (milk) 53, *sangue* (blood) 47, *trago* (*shot*) 40, *aguardente* (firewater) 28, *cálice* (chalice) 28, *goles* (sips) 27, *noite* (night) 23, *chá* (tea) 22, *coisa* (thing) 20, *copos* (glasses) 20, *fonte* (fountain/source) 20, *palavras* (words) 20, *ar* (air) 19, *champanhe* (champagne) 19, *conhaque* (cognac) 19, *cachaça* (rum) 18, *garrafa* (bottle)18, *olhos* (eyes) 18 and *lágrimas* (tears) 17. Likewise, with one click they could find the nominal collocates of *seco* (dry; all forms): *carne* (meat) 53 tokens, *olhos* (eyes) 51, *lábios* (lips) 34, *molhados* (wet) 34, *tempo* (time) 33, *garganta* (throat) 32, *voz* (voice) 30, *frutos* (fruit) 26, *terra* (earth) 26, *boca* (mouth) 25, *árvore* (tree) 24, *rio* (river) 24, *ar* (air) 23, *clima* (climate) 23, *tom* (tone) 23, *água* (water) 22, *folha* (leaf) 22, *galhos* (branches) 22, *estação* (station) 21, *anos* (years) 20, *fruto* (fruit) 20, *tosse* (cough) 19, *leito* (bed) 18 and *pele* (skin) 18.

As with other corpus interfaces, it is also possible to sort the collocates to find which ones are most 'closely bound' to the node word – in other words, given Word 1 (collocate) what is the probability that Word 2 (the node word) will be nearby. With the interface for the *Corpus do Português*, we use the Mutual Information (MI) score to rank the entries. For example, consider the nominal collocates of *fazer* (do/make; all forms). Sorted by the MI score (and with a minimum count of 15 tokens), the top entry is *pazes* (one's peace). A full 108 of the 153 tokens of *pazes* in the corpus are near *fazer*, for an MI score of 5.64. (Perhaps a better way of looking at the connection between the two words is to ask what verb *could* occur with *pazes*, other than *fazer*.) The other entries – in order of decreasing MI score and showing here the ratio of collocate with node word – are *jus* (justice) (53/95 tokens are with *fazer*), *menção* (mention) 139/254, *mesura* (curtsey) 33/62, *cócegas* (tickle) 47/90, *caretas* (faces) 44/87, *negaças* (trick) 20/44, *careta* (face) 87/210, *apologia* (praise) 28/74, *figas* (fig hands) 15/40, *continência* (salute) 22/62, *vénia* (curtsey) 15/44, *alarde* (boast) 16/49, *trejeito* (twitch) 15/50 and *bulha* (racket) 53/197 (still with an MI score of 4.25). Likewise, the MI-ranked collocates of *mão* (hand; all forms) are *calejadas* (calloused; 15/16 tokens are near *mão*, for an MI score of 7.16), *calosas* (calloused; 23/28 tokens), *espalmada* (spread out) 40/54, *papudas* (thick) 16/23, *espalmadas* (spread out (pl)) 44/66, *trémulas* (trembling (pl)) 38/70, *geladas* (cold) 40/86, *enterradas* (buried) 19/59, *amarradas* (tied up) 17/54, *trémula* (trembling) 46/185, *magras* (thin) 30/131 and *apertadas* (tight) 18/94 (and still with an MI score of 4.86).

One of the most useful aspects of collocates is the way that they can help to show differences in meaning and usage between two words. For example, the two words *romper* and *quebrar* are difficult to distinguish for native speakers of English, since they are both translated as 'break'. But there are many collocates that strongly prefer one or the other of the two words in Portuguese. For example, the following collocates occur primarily with *romper* (limited here to four words after *romper*): *aurora* (40 tokens with *romper* and 0 tokens with *quebrar*), *manhã* (morning) 37/0, *seio* (bosom) 23/0, *multidão* (crowd) 16/0, *choro* (cry) 15/0, *marcha* (march) 23/1, *fileira* (row) 10/0, *grito* (scream) 16/1, *nuvem* (cloud) 15/1, *lábio* (lip) 15/1, *custo* (cost) 13/1, *fogo* (fire) 11/1, *mato* (bushes) 10/1, *lado* (side) 10/1, *mata* (woods) 10/1, *relação* (relationship) 32/4, *dia* (day) 69/10, *soluço* (sob) 19/3, *peito* (chest) 12/2, *sol* (sun) 17/5, *contracto* (agreement) 10/4, *coração* (heart) 13/6 and *cerco* (siege) 13/7. Words occurring primarily with *quebrar*, on the other hand, are: *cara* (face; 39 tokens with *quebrar* and 0 tokens with *romper*), *sigilo* (secrecy) 36/0, *galho* (branch/favour) 25/0, *pedra* (rock) 22/0, *cabeça* (head) 41/1, *perna* (leg) 38/1, *recorde* (record) 17/0, *osso* (bone) 16/0, *luneta* (telescope) 12/0, *louça* (dishes) 11/0, *costela* (rib) 11/0, *dente* (tooth) 10/0, *copo* (glass) 10/0, *coisa* (thing) 13/1, *nariz* (nose) 11/1, *monotonia* (monotony) 19/2, *corpo* (body) 13/2, *braço* (arm) 18/3, *rotina* (routine) 11/2, *encanto* (spell) 32/6, *vidro* (glass) 13/3, *gelo* (ice) 12/3 and *encontro* (date) 15/4.

In addition to looking at semantic differences between (near) synonyms, we can also use collocates to compare the collocates of any other set of words, such as *Portugal* and *Brasil* or *homem* (man) and *mulher* (woman). For example, with one simple search we can find verbs that occur more (within four words after the noun) with *homem* than with *mulher*. These include *comandar* (command; 20 tokens with *homem*, 0 with *mulher*, or 25.7 times more frequent with *homem* than *mulher*, taking into account the

overall frequency of these two words in the corpus), *invadir* (invade) 14/0 18.0, *figurar* (appear) 11/0 14.1, *atacar* (attack) 10/0 12.9, *assaltar* (rob) 17/1 10.9, *montar* (ride) 13/1 8.4, *cavar* (dig) 13/1 8.4, *alimentar* (feed) 13/1 8.4, *errar* (miss) 12/1 7.7, *envolver* (involve) 12/1 7.7, *avisar* (warn) 10/1 6.4, *encarregar* (be in charge) 10/1 6.4, *regular* (regulate) 10/1 6.4, *contribuir* (contribute) 10/1 6.4, *confundir* (mix up) 10/1 6.4, *sacrificar* (sacrifice) 10/1 6.4, *construir* (build) 37/4 5.9 and *avaliar* (evaluate) 15/2 4.8.

On the other hand, verbs that occur more frequently with *mulher* are: *prevenir* (prevent; 19 tokens with *mulher* and 0 with *homem*, or 59.1 times as frequent with *mulher* than *homem*, taking into account the overall frequency of these two word), *pintar* (paint) 11/4 4.3, *esfregar* (rub) 11/4 4.3, *convidar* (invite) 15/6 3.9, *lavar* (wash) 21/9 3.6, *benzer* (pray) 11/5 3.4, *estremecer* (shudder) 19/9 3.3, *trair* (deceive) 16/8 3.1, *despertar* (wake) 12/6 3.1, *deter* (stop) 10/5 3.1, *sonhar* (dream) 23/12 3.0, *adorar* (adore) 34/19 2.8, *chorar* (cry) 74/42 2.7, *fiar* (spin) 14/8 2.7, *baixar* (lower) 13/8 2.5, *dançar* (dance) 11/7 2.4 and *agradecer* (thank) 11/7 2.4. It is quite striking how many of the verbs occurring with *homem* refer to violent actions (e.g., *invadir*, *atacar*, *assaltar*), whereas those with *mulher* refer to domestic activities or a supposed 'subservient' status (e.g., *esfregar*, *lavar*, *chorar*, *adorar*).

In addition to finding the most frequent collocates of a given word, with the web interface we can also find all cases of Word 1 'near' a set of Words 2. For example, we could set *olhos* (eyes) as our node word and then look for all words starting with *fech** (the root for 'to close', 'closing' or 'closed') nearby. We would find that the most frequent forms are *fechados* (274 tokens), *fechou* 217, *fechar* 181, *fechando* 64, *fechava* 62, *fecha* 59, *fechei* 27, *fechara* 23, *fecho* 21, *fechasse* 17, *fechado* 16 and *feche* 15. As we will soon see, it is also possible to find all cases of any [Semantic Concept 1] 'near' [Semantic Concept 2], which leads to some fairly powerful searches.

In addition to collocates, it is also possible to use an integrated thesaurus for powerful semantically oriented queries. The thesaurus, which is stored in a linked database table, contains 477,713 entries for 34,352 headwords. For example, the entries in the database for the adjective *duro* are *amargo* (bitter), *árduo* (arduous), *áspero* (rough), *calejado* (calloused), *consistente* (consistent), *corajoso* (courageous), *córneo* (leathery), *cru* (raw), *cruel* (cruel), *custoso* (costly), *desabrido* (unbridled), *desagradável* (unpleasant), *desumano* (inhuman), *difícil* (difficult), *dificultoso* (difficult), *duro* (hard), *enérgico* (energetic), *fastidioso* (fastidious), *fero* (ferocious), *feroz* (ferocious), *férreo* (steely), *firme* (firm), *forte* (strong), *impetuoso* (impetuous), *implacável* (relentless), *inclemente* (unforgiving), *inflexível* (inflexible), *insensível* (insensitive), *intransigente* (uncompromising), *inumano* (ruthless), *obstinado* (obstinate), *penoso* (painful), *resistente* (resistant), *rigoroso* (rigorous), *rijo* (stiff), *ríspido* (harsh), *rude* (rude), *seco* (dry), *severo* (severe), *sólido* (solid), *teimoso* (stubborn), *teso* (stiff), *trabalhoso* (laborious), *valente* (brave) and *vigoroso* (vigorous).

At the most basic level, users can simply enter [= word] in the search form, to see the frequency of synonyms in different sections of the corpus. For example, for the search [= *duro*] the first 12 entries are shown in Figure 4.6.

The real power with the synonyms is the ability to use this information as part of a more complex query. For example, it is possible to search for all adjectives with a synonym of *imagem* (image/picture): (query: [[*imagem*]] [j*] and the double brackets

	TOTAL	1800s	1900s	PORT	BRAZ	ACAD	NEWS	FICT	ORAL
FORTE	8580	2326	4098	1906	2192	945	1187	1505	461
DIFÍCIL	3987	777	3156	1486	1670	439	1203	752	762
FIRME	2631	634	655	293	362	59	140	423	33
VIVO	1239	461	772	314	458	79	268	330	95
CRUEL	1828	836	286	145	141	40	53	181	12
DURO	1582	427	683	322	361	87	161	368	67
SECO	1258	361	640	348	292	134	57	391	58
FEROZ	676	365	241	132	109	38	24	163	16
RUDE	627	366	171	98	73	12	15	140	4
VALENTE	849	347	163	81	82	10	14	131	8
SEVERO	560	278	189	85	104	24	35	123	7
DESAGRADÁVEL	428	191	226	125	101	34	39	129	24

Figure 4.6 Synonyms of *duro* by century, dialect and genre[3]

around *imagem* will find all forms of all synonyms of the word). In this case the most frequent strings would be: *figura humana* (human figure; 58 tokens), *imagens digitais* (digital images) 55, *representação conceitual* (conceptual representation) 38, *representação proporcional* (proportional representation) 38, *gravuras rupestres* (rock engraving) 35, *figuras humanas* (human figures) 34, *pintura portuguesa* (Portuguese painting) 27, *representação nacional* (national representation) 27, *impressões digitais* (fingerprints) 25, *imagens aéreas* (aerial images) 22, *imagens balanceadas* (balanced images) 22, *pinturas murais* (mural paintings) 22, *representações sociais* (social representations) 20, *imagem digital* (digital image) 18, *pintura histórica* (historic/al painting) 18, *figuras importantes* (key figures) 17, *impressão geral* (general impression) 17, *representação gráfica* (graphic representation) 17, *figura principal* (main figure) 16, *figura central* (central figure) 15, *pinturas rupestres* (rock paintings) 15, *figura importante* (key figure) 14, *imagem original* (original image) 14, *imagem viva* (live image) 14, *pintura moderna* (modern painting) 14 and *representação interna* (internal representation) 14. Once we have this list, we can click on any of the phrases to see it in context and then (hopefully) see why some synonyms occur with one adjective while other synonyms occur with other adjectives.

It is also possible to use synonyms as part of a collocates-based search. For example, users could search for synonyms of *agradável* (pleasant) somewhere near a form of *mulher*. Grouping the results by lemma (with the base (masculine) form of the lemma shown here), the most frequent synonyms would be: *belo* (pretty; 210 tokens), *bonito* (beautiful) 191, *bom* (good) 152, *lindo* (beautiful) 36, *encantador* (enchanting) 32, *delicado* (delicate) 26, *alegre* (happy) 20, *doce* (sweet) 14, *sedutor* (seductive) 12, *delicioso* (delicious) 11, *caro* (expensive) 11, *amável* (lovely) 10, *agradável* (pleasant) 9, *suave* (soft) 7, *gentil* (kind) 7, *atraente* (attractive) 6, *atractivo* (attractive) 6, *gracioso* (gracious) 6, *galante* (gallant) 5 and *engraçado* (funny) 5.

Finally, users who want to have maximum control over the content of these semantically oriented queries can create their own customized word lists and then use these as part of the query. The customized word lists can be based on synonym sets from the corpus interface (with words deleted or added), from collocates of a given word or

phrase, or these lists can be created 'from scratch'. All of this can be done quite easily via the web-based interface and the lists can then be used at any point in the future.

For example, users could (via the web-based interface) create a customized list of pieces of clothing and another list of colours and then combine these two lists to find phrases like: *vestido branco* (white dress; 71 tokens), *vestido preto* (black dress) 46, *gravata branca* (white tie) 42, *camisa branca* (white shirt) 29, *vestido azul* (blue dress) 24, *capa preta* (black coat) 20, *chapéu preto* (black hat) 16, *saia branca* (white skirt) 16, *vestido verde* (green dress) 16, *chapéu branco* (white hat) 15, *meias pretas* (black socks) 15, *saia preta* (black sock) 15, *blusa branca* (white blouse) 14, *gravata preta* (black tie) 13, *capa vermelha* (red coat) 13, *saco azul* (blue bag) 13, *capa branca* (white coat) 12, *capa azul* (blue coat) 12, *vestido roxo* (purple dress) 12, *bata branca* (white gown) 10, *botas brancas* (white boots) 10, *capa verde* (green coat) 10, *vestido vermelho* (red dress) 10, *vestidos brancos* (white dresses) 10, *botas pretas* (black boots) 9, *capa amarela* (yellow coat) 9 and *saias brancas* (white skirts) 9.

In summary, via the web-based interface, it is possible to carry out powerful semantically oriented queries of basically any [Concept 1] 'near' any other [Concept 2]. Our sense is that for semantically oriented queries this is much more useful than some other approaches, which are basically limited to looking at strings of words, perhaps with just lemmatization and part-of-speech tagging.

7. *Corpus do Português* – comparing genres, dialects and time periods

Nearly all of the queries discussed so far have looked at all dialects, genres and time periods of Modern Portuguese as one entire block. However, the interface and architecture of the *Corpus do Português* allow users also to limit by any of these three criteria and (even more powerfully) to compare between the three.

Let us first turn to genre-based differences. At the most basic level, users can limit their search to just one of the four main genres – spoken, fiction, newspaper and academic texts – and then retrieve results from just that portion of the corpus. For example, a search for words ending in **mento* in academic texts only will find: *desenvolvimento* (development; 2803 tokens), *movimento* (movement) 2748, *conhecimento* (knowledge) 1796, *comprimento* (length) 1282, *aumento* (rise/raise) 1228, *tratamento* (treatment) 1114, *crescimento* (growth) 1083, *momento* (moment) 1056, *pensamento* (thought) 1002, *comportamento* (behaviour) 970, *elemento* (element) 849, *instrumento* (instrument) 735, *processamento* (processing) 707, *parlamento* (parliament) 663, *reconhecimento* (recognition) 519, *alinhamento* (alignment) 512, *funcionamento* (workings) 511, *pagamento* (payment) 504, *casamento* (marriage) 464 and *estabelecimento* (establishment) 427. Likewise, users can limit any other search – involving collocates, synonyms, customized lists, etc – to just a particular section of the corpus.

The real power, however, comes as we compare one section against another. For example, continuing our focus on genre-based differences, we can easily find those verbs that are more common in academic than in fiction texts (and vice versa). We simply highlight [Academic] for [Part 1] in the search form, [Fiction] for [Part 2] and select the code for verbs in the [Word] field.[4]

The query shows that the following are the most strongly academic verbs: *detectar* (detect; 158 tokens in academic, 3 in fiction), *instituir* (establish) 44/1, *ocasionar* (cause) 34/1, *originar* (originate) 56/2, *estruturar* (structure) 27/1, *especializar* (specialize) 53/2, *possibilitar* (enable) 48/2, *inibir* (inhibit) 24/1, *incrementar* (increment) 23/1, *aprimorar* (improve) 23/1, *leccionar* (teach) 45/2, *ocorrer* (happen) 468/21, *ilustrar* (illustrate) 62/3, *bloquear* (block) 20/1, *acasalar* (mate) 20/1, *seleccionar* (select) 107/6, *designar* (appoint) 329/19, *promover* (promote) 314/19, *ressaltar* (stress) 171/11, *restringir* (restrict) 44/3, *alienar* (alienate) 29/2, *reivindicar* (claim) 28/2, *utilizar* (utilize) 497/36, *gerar* (generate) 273/20 and *produzir* (produce) 658/50. Conversely, the most strongly fiction verbs are: *gritar* (scream; 389 fiction, 2 academic), *almoçar* (eat lunch) 164/1, *sorrir* (smile) 330/3, *jantar* (eat dinner) 1166/11, *calhar* (come in handy) 102/1, *rir* (smile) 706/7, *agradecer* (thank) 94/1, *chorar* (cry) 654/8, *tremer* (shake) 240/3, *acudir* (help) 78/1, *consolar* (comfort) 71/1, *sumir* (disappear) 70/1, *perdoar* (forgive) 124/2, *insultar* (insult) 58/1, *mirar* (look) 56/1, *quietar* (quiet down) 111/2, *vagar* (roam) 104/2, *escurecer* (become dark) 50/1, *ladrar* (bark) 49/1, *fitar* (stare) 97/2, *sacudir* (shake) 96/2, *vomitar* (vomit) 48/1, *arrumar* (clean up) 143/3, *despedir* (say goodbye) 187/4 and *deitar* (lay/lie down) 448/10.

The corpus also allows us to compare the collocates of a given word or phrase in two genres, to investigate differences in meaning between the genres. For example, if we look for collocates of the lemma *duro* in fiction (compared to academic), we find collocates like: *chão* (ground), *olhos* (eyes), *palavras* (words), *seios* (breasts), *colarinho* (collar), *trabalho* (work), *pau* (wood), *coração* (heart), *carnes* (meats), *olhar* (look), *linhas* (lines), *golpe* (blow), *homem* (man), *brilho* (shine), *cabelos* (hair), *pernas* (legs), *pedras* (rocks), *tipo* (type) and *trabalhos* (work). But when we look for collocates of *duro* in academic (that are also uncommon in fiction), we find: *drogas* (drugs), *materiais* (materials), *rochas* (rocks), *fibras* (fibres), *herói* (hero), *cancro* (cancer), *madeiras* (wood), *polimentos* (polish), *plástico* (plastic), *filamentos* (filaments), *partes* (parts), *conectivo* (connective), *cristalinos* (crystals), *cerdas* (bristles), *águas* (water), *ano* (year), *lugar* (place), *massas* (pasta/putty), *moleque* (rascal), *minério* (oar) and *mineral* (mineral).

In addition to comparing features in different genres, we can also compare any feature – e.g., lexical, grammatical, or semantic (via collocates or synonyms) – in the two dialects of European and Brazilian Portuguese. To give one quick example, we can search for those verbs that are more common in one of these two dialects than in the other. We find that the following verbs are used primarily in European Portuguese: *registar* (record), *leccionar* (teach), *recandidatar* (reapply), *colmatar* (fill up), *rentabilizar* (monetize), *poisar* (land), *apetecer* (fancy), *despoletar* (trigger), *estoirar* (blow up), *fintar* (feint), *dignificar* (dignify), *laborar* (labour), *amanhar* (farm), *descodificar* (decode), *sagrar* (become), *acartar* (cause), *rapar* (scrape), *cahir* (fall), *baralhar* (shuffle), *coalhar* (curdle) and *vincar* (crease). The verbs that are found primarily in Brazilian Portuguese, on the other hand, are: *planejar* (plan), *liberar* (release), *esquentar* (heat), *indenizar* (indemnify), *acessar* (access), *checar* (check), *pleitear* (claim), *gerenciar* (manage), *coibir* (curb), *quitar* (pay up), *priorizar*

(prioritize), *conscientizar* (become aware), *coletar* (collect), *cassar* (revoke), *revisar* (revise), *descartar* (dismiss), *decolar* (take off), *despencar* (drop), *mensurar* (measure), *garimpar* (mine), *desvencilhar* (free oneself from) and *descumprir* (fail to comply).[5]

Finally, because the corpus contains texts from the 1100s to 1999, we can also compare any set of features across different centuries, to see how the language is changing. For example, with one simple query we can find **ismo* words that are more common in the 1900s than in the 1800s: *terrorismo* (terrorism), *racismo* (racism), *marxismo* (marxism), *protagonismo* (leadership), *urbanismo* (planning), *nazismo* (nazism), *turismo* (tourism), *fascismo* (fascism), *expressionismo* (expressionism), *optimismo* (optimism), *cubismo* (cubism), *neoclassicismo* (neoclassicism), *surrealismo* (surrealism), *ciclismo* (cycling), *profissionalismo* (professionalism), *pragmatismo* (pragmatism), *neo-realismo* (neo-realism), *neoliberalismo* (neo-liberalism), *sindicalismo* (unionism), *futurismo* (futurism), *colonialismo* (colonialism), *maneirismo* (mannerism), *conservadorismo* (conservatism), *corporativismo* (corporatism), *virtuosismo* (virtuosism) and *existencialismo* (existentialism). Likewise, we can find **ismo* words that have decreased in frequency between the 1800s and the 1900s: *idiotismo* (idiotism), *syllogismo* (syllogism), *christianismo* (christianity), *macavelismo* (machiavellianism), *madamismo* (madamism), *euphemismo* (euphemism), *dialoguismo* (dialogism), *abolicionismo* (abolitionism), *esclavagismo* (slavery), *caiporismo* (bad luck-ism), *decadismo* (decadism), *monarquismo* (monarchism), *humanitismo* (humanitism), *escravismo* (slavery), *exclusivismo* (exclusivism), *galicismo* (gallicism), *indiferentismo* (indifferentism), *nativismo* (nativism), *heroismo* (heroism), *egoismo* (egotism), *servilismo* (servilism), *babismo* (babism) and *coquetismo* (coquetry).

The comparison between different historical periods is not limited to lexical differences. We can also look at differences in usage and meaning via a comparison of collocates. For example, while the word *mulher* (woman) has not really changed meaning from the 1800s to the 1900s, the adjectival collocates that occur with this word certainly have changed, owing to cultural shifts. For example, in the 1800s the collocates often referred to 'moral' characteristics or to the supposed 'weakness' of women: *infame* (infamous; 12 tokens in 1800s, 0 tokens in 1900s), *desmaiada* (fainted) 7/0, *adorada* (beloved) 6/0, *infernal* (devilish) 6/0, *insensível* (insensitive) 6/0, *maldita* (damned) 6/0, *meiga* (sweet) 6/0, *vaidosa* (vain) 6/0, *escravas* (slaves) 5/0, *desonrado* (dishonoured) 5/0, *bárbara* (fantastic) 5/0, *ridículo* (ridiculous) 5/0, *desgraçada* (wretched) 14/1, *indigna* (undignified) 14/1, *divina* (divine) 11/1, *sublime* (sublime) 16/2, *digna* (dignified) 8/1, *indispensável* (indispensable) 8/1, *esplêndida* (splendid) 14/2, *agradável* (pleasant) 7/1, *carinhosa* (tender) 6/1, *abafada* (muffled) 6/1, *fiel* (faithful) 6/1, *fracas* (weak) 6/1, *nobre* (noble) 11/2, *ardente* (fervent) 15/3, *desagradáveis* (unpleasant) 5/1, *infiel* (unfaithful) 5/1, *mísera* (miser) 5/1, *virtuosa* (virtuous) 19/4 and *honestas* (honest) 9/2. In the 1900s, on the other hand, the adjectives are much more prosaic (although they also sometimes refer to characteristics that would have been taboo in the 1800s): *jovens* (young; 32 tokens in the 1900s, 0 in the 1800s), *grávidas* (pregnant) 25/0, *sexual* (sexual) 16/0, *sozinha* (lonely) 15/0, *familiar* (family/iar) 14/0, *principal* (main) 11/0, *sexuais* (sexual)11/0, *mundial* (world) 10/0, *duro* (hard) 9/0, *idosas* (aged) 9/0, *responsável* (responsible) 9/0, *atual* (current) 8/0, *normal* (normal) 8/0, *fértil* (fertile) 7/0, *masculino* (masculine) 7/0, *oficial* (official) 7/0, *português* (Portuguese) 7/0, *estadual* (state) 6/0, *atraente* (attractive) 6/0 and *ausente* (absent) 6/0.

8. The *Frequency Dictionary of Portuguese*

As we can see, the *Corpus do Português* allows researchers to carry out a very wide range of queries. However, the reality is that some people – especially language learners – would probably prefer 'ready-made' resources that are based on the corpus, rather than constantly having to use the corpus directly. Recognizing the value of frequency-based materials for language learners, we have used the frequency data from the corpus to create a frequency dictionary of Portuguese that is based on the *Corpus do Português*. The general approach that we followed in the creation of this dictionary is similar to that used for our *Frequency Dictionary of Spanish* (Davies, 2005a) and the *Frequency Dictionary of American English* (Davies and Gardner, 2010).

There were already a handful of printed frequency dictionaries of Portuguese (Brown, 1951; Duncan, 1972; Kelly, 1970; Bacelar do Nascimento et al., 1987 and Roche, 1975), as well as two or three in electronic format on the web. Nevertheless, none of these was based on a large, balanced corpus of Portuguese (in other words, with texts from a number of different genres). Therefore, the goal was to create a dictionary that would contain the 5000 most frequent lemmas in Portuguese – based on the data from the *Corpus do Português* and with a number of features that were specifically oriented towards the language learner. We began work on the dictionary soon after the *Corpus do Português* was finished in 2006 and the dictionary was published in 2007 (Davies and Preto-Bay, 2007).

The *Frequency Dictionary of Portuguese* is designed to meet the needs of a wide range of language students and teachers. The main index contains the five thousand most common words in Portuguese, starting with such basic words as *o* and *de* and quickly progressing through to more intermediate and advanced words. The dictionary is based on the actual frequency of words in the 20 million words in the texts from the 1900s in the *Corpus do Português*. As we have discussed, these texts are balanced between many different genres of Portuguese (e.g., fiction, newspapers, academic texts and transcripts of actual conversations), and the users can therefore feel comfortable that these are words that one is very likely to encounter subsequently in the 'real world'.

The following information is given for each of the 5000 entries in the dictionary:

> rank frequency (1, 2, 3, ...), headword, part of speech, English equivalent, dialect, sample sentence, translation, range count, raw frequency total, indication of major register variation

As a concrete example, let us look at the entry for *bruxa* (witch):

4522 *bruxa* nf witch
A caça às bruxas é muitas vezes acompanhada de histeria – Witch hunts are often accompanied by hysteria
35 | 235 +fic

This entry shows that word number 4522 in the rank order list is *bruxa*, which is a feminine noun [nf] that can be translated as 'witch' in English. We then see an actual sentence or phrase from the corpus, which shows the word in context, as well as a translation of this sentence into English. The two following numbers show that the word occurs in 35 of the 100 equal-sized blocks (200,000 words each) from the corpus (i.e., the range count) and that this lemma occurs 235 times in the corpus. Finally, the notation [+fic] indicates that the word is much more common in the fiction register than would otherwise be expected. In summary, then, each of the 5000 entries provides the language learners with information about the frequency of the word, its meaning (via the glosses and the sample sentence) and some indication of the distribution of the word in the different genres.

One of the criticisms of frequency dictionaries is that they are just 'lists of words' and that there is no semantic grouping of words. To address this criticism to some degree, we placed throughout the main frequency-based index approximately thirty 'call-out boxes', which serve to display in one list a number of thematically related words. These include lists of words related to the body, food, family, weather, professions, nationalities, colours, emotions, verbs of movement and communication and several other semantic domains.

In addition to vocabulary that is tied to a particular semantic category, however, we also focused on several topics in Portuguese grammar that are often difficult for beginning and intermediate students. For example, there are lists that show the most common diminutives, superlatives and derivational suffixes to form nouns, the most common verbs and adjectives that take the subjunctive, which verbs most often take the 'reflexive marker' *se*, which verbs most often occur almost exclusively in the imperfect and preterit and which adjectives occur almost exclusively with the two copular verbs *ser* (be) and *estar* (be) or the semi-copular *ficar* (get, stay). Finally, there are even more advanced lists that compare the use of nouns, verbs, adjectives and adverbs across registers, and show which words are used primarily in spoken, fiction, newspapers, or academic texts. Related to this is a list showing which are the most frequent words that have entered the language in the past 100 to 200 years.

While there are more than thirty thematic lists in the frequency dictionary, we provide short extracts from just five of them here. First, the following is an example of one of the more than fifteen lists that contain the most common words from a particular semantic class. In this case, we see the top twenty words for members of the family (showing Portuguese, rank order of the word in the dictionary, gender and English gloss):

filho 143 M son, children (pl), *pai* 170 M father, parents (pl), *mãe* 272 F mother, *filha* 468 F daughter, *irmão* 488 M brother, *marido* 737 M husband, *irmã* 1100 F sister, *neto* 1428 M grandson, grandchildren (pl), *parente* 1827 MF relative, extended family member, *esposa* 1885 F wife, *primo* 1968 M cousin, *tio* 2048 M uncle, *avô* 2369 M grandfather, *sobrinho* 2634 M nephew, *viúva* 2683 F widow, *avó* 3551 F grandmother, *cunhado* 3758 M brother-in-law, *tia* 3798 F aunt, *noiva* 3814 F fiancee, bride, *genro* 4043 M son-in-law.

An example of one of the grammatically-oriented lists is the following, which shows the most frequent 'triggers' for the subjunctive, which is always a difficult concept for non-native speakers. The following list shows just the top eight triggers in each of four different categories and it shows frequency in the corpus, the Portuguese word and the English gloss:

Emotion/desire: 1059 *querer* to want; 414 *esperar* to hope, expect; 157 *gostar* to like; 79 *temer* to fear; 67 *desejar* to desire, wish; 61 *bastar* to be enough; 50 *importar* to care, be interested in; 44 *preferir* to prefer

belief/opinion: 203 *pensar* to think; 202 *acreditar* to believe; 92 *crer* to believe in; 80 *admitir* to admit; 77 *imaginar* to imagine; 72 *achar* to find, think; 72 *negar* to deny

force/control: 441 *pedir* to request, ask for; 269 *dizer* to say, tell; 241 *permitir* to permit; 140 *impedir* to stop from doing; 125 *deixar* to leave, allow; 115 *evitar* to avoid; 91 *mandar* to command, order

adjectives: 153 *preciso* necessary, precise; 114 *possível* possible; 85 *bom* good; 77 *natural* natural; 70 *necessário* necessary; 46 *importante* important; 42 *provável* probable; 39 *melhor* better.

An example of one of the genre-based lists is the following list, which shows the nouns, verbs and adjectives that are at least three times as frequent in fiction as would be expected (based on the size of fiction in the 20 million word corpus). Note that these are common, 'everyday' words, but that is precisely the point with fiction texts. They often contain words relating to the 'real world', as the author is creating an imaginary world, which might be less frequent in a corpus that is based primarily on newspapers or academic texts.

Nouns: *olho* 376 M eye; *braço* 959 M arm; *alma* 1053 F soul; *dedo* 1324 M finger; *janela* 1386 F window; *sombra* 1420 F shadow, shade; *silêncio* 1512 M silence; *cabelo* 1516 M hair; *dona* 1630 F madam, owner; *rosto* 1729 M face, cheek; *alegria* 1784 F happiness, joy; *cama* 1810 F bed

Verbs: *sentar* 1366 to sit; *rir* 1848 to laugh; *saltar* 2148 to leap, jump; *apertar* 2218 to press, tighten; *chorar* 2234 to cry; *erguer* 2248 to raise up, support; *escutar* 2306 to listen; *apagar* 2393 to erase, turn off; *admirar* 2476 to admire, be astonished; *calar* 2490 to be quiet, shut up; *descansar* 2608 to rest; *gritar* 2630 to yell, shout

Adjectives: *triste* 2003 sad; *nu* 2116 naked, nude; *gordo* 2206 fat, thick; *íntimo* 2334 intimate, inner; *fresco* 2388 fresh, cool; *cansado* 2532 tired; *sujo* 3014 dirty, soiled; *feio*

3149 ugly; *magro* 3378 thin; *suave* 3439 soft, pleasing, smooth; *mole* 3450 soft, weak; *ridículo* 3464 ridiculous.

A final example of a list from the frequency dictionary shows those words that are much more common (per million words) in the 1900s than in the 1800s. It is quite interesting to see which semantic fields have the most new words (e.g., technology) and it is also important to point out that only a corpus with a diachronic component (such as the *Corpus do Português*) is able to provide such information. The entries below show the Portuguese word, the rank order in the dictionary and the English gloss.

Nouns *televisão* 499 television; *futebol* 1066 soccer; *avião* 1092 aeroplane; *equipa* 1095 team; *investimento* 1469 investment; *desemprego* 2073 unemployment; *liderança* 2085 leadership; *campeonato* 2087 championship; *relacionamento* 2272 relationship; *financiamento* 2292 financing

Verbs *liderar* 2310 to lead; *financiar* 2559 to finance, fund; *candidatar* 3514 to run for office; *incentivar* 3704 to encourage; *solucionar* 4867 to solve; *minimizar* 5329 to minimize; *protagonizar* 5615 to play a leading role; *posicionar* 5694 to position; *disponibilizar* 5787 to make available; *implementar* 5858 to implement

Adjectives *global* 2290 global; *desportivo* 2421 athletic, sporting; *ambiental* 2681 environmental; *comunitário* 2723 community (ADJ); *soviético* 2767 soviet; *tecnológico* 2824 technological; *empresarial* 3001 business (ADJ); *turístico* 3625 tourist (ADJ); *operacional* 3829 operational; *ecológico* 4182 ecological

Adverbs *practicamente* 1749 practically; *basicamente* 3229 basically; *obviamente* 3457 obviously

Aside from the main frequency listing, there are also indexes that sort the entries by alphabetical order and part of speech. The alphabetical index can be of great value to students who for example want to look up a word from a short story or newspaper article and see how common the word is in general. The part of speech indexes could be of benefit to students who want to focus selectively on verbs, nouns, or some other part of speech.

The expectation, then, is that this frequency dictionary will significantly maximize the efforts of a wide range of students and teachers who are involved in the acquisition and teaching of Portuguese vocabulary. This will be in addition to the tens of thousands of researchers (23,000 distinct users in the last three years alone) who have used the *Corpus do Português* as an integral part of their research – especially in terms of looking at variation in Portuguese in terms of historical change, dialectal differences and genre-based variation.

References

Bacelar do Nascimento, M. F., Garcia Marques, M. L. and Segura da Cruz, M. L. (1987), *Português Fundamental. Métodos e Documentos.* Lisbon: INIC, CLUL.

Brown, C. B. (1951), *Brazilian Portuguese Idiom List, Selected on the Basis of Range and Frequency of Occurrence.* Nashville: Vanderbilt University Press.

Davies, M. (2005a), *A Frequency Dictionary of Spanish: Core Vocabulary for Learners.* London: Routledge.

—(2005b), 'The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation'. *International Journal of Corpus Linguistics*, 10, 301–28.

—(2005c), 'Advanced research on syntactic and semantic change with the Corpus del Español', in C. Pusch, et al. (ed.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics.* Tübingen: Guntar Naar, pp. 203–14.

—(2008), 'Spanish and Portuguese Corpus Linguistics'. *Studies in Hispanic and Lusophone Linguistics*, 1, 149–86.

—(2009), 'The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights'. *International Journal of Corpus Linguistics*, 14, 159–90.

—(2010a), 'Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the *Corpus do Português*', in A. Enrique-Arias (ed.), *Diacronía de Las Lenguas Iberorromances: Nuevas Perspectivas desde la Lingüística de Corpus.* Frankfurt/Madrid: Vervuert/Iberoamericana, pp. 137–66.

—(2010b), 'More than a peephole: Using large and diverse online corpora'. *International Journal of Corpus Linguistics*, 15, 405–11.

—(2012), 'Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English'. *Corpora*, 7, 121–57.

Davies, M. and Gardner, D. (2010), *A Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists.* London: Routledge.

Davies, M. and Preto-Bay, A. R. (2007), *A Frequency Dictionary of Portuguese: Core Vocabulary for Learners.* London: Routledge.

Duncan, J. C. (1972), *A Frequency Dictionary of Portuguese Words.* PhD dissertation, Stanford University.

Kelly, J. R. (1970), 'A computational frequency and range list of five hundred Brazilian Portuguese words'. *Luso-Brazilian Review*, 7, 104–13.

Roche, J. (1975), *Sobre o Vocabulário da Poesia Portuguesa.* Paris: Fundação Calouste Gulbenkian.

Notes

1 All of the charts and tables from the *Corpus do Português* in this paper are shown with the English interface, but the corpus is also available with a Portuguese interface. Also note that the web interface allows users to see raw frequency, normalized frequency, or a combination of these.

2 Although it is difficult to see in the printed version, in the online version the surrounding words are colour-coded for part of speech. In addition, in the online

interface, the citations and the KWIC entries themselves are quite a bit longer than what can be shown in this printed version.

3 It is important to point out that not all tokens of a given synonym are in fact used synonymously with *duro*. Such a determination is far too ambiguous and difficult. All we can say is that in some contexts the word is a synonym of *duro*, and then show how many times total that 'sometimes synonym' occurs in the corpus.

4 We will simplify the search by looking just for infinitival forms of the verb, but we could search for all verb forms, and then group by lemma. Also, we limited our search to verbs occurring at least once in each of the two genres.

5 We have limited the search here to infinitival forms, but we could search for all verb forms and then group by lemma. Also, note that one or two authors or texts in a particular dialect may account for the majority of the tokens. Finally, note that at times orthographic variation may be involved, although most of the differences shown are lexical in nature.

PtTenTen: A
l

Adam Kilgarriff, Miloš Jaku
an

1. In

There are a number of ways in which
as described in Rundell and Kilgarriff
consistent and faster. But how might t
project? If starting from a blank sheet of
proceed?

In this paper we describe such an ex
...ford 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.