verbs of movement and prepositions

edited by
António Leal

PORTO
- 2018 -

TÍTULO	Verbs, movement and prepositions
COORDENADOR	António Leal
EDITOR	Centro de Linguística da Universidade do Porto Faculdade de Letras da Universidade do Porto
CONCEÇÃO GRÁFICA	Invulgar - Artes Gráficas, S.A.
ANO DE EDIÇÃO	2018
TIRAGEM	150 exemplares
ISBN	978-989-54104-5-3
DEPÓSITO LEGAL	442271/18

Esta publicação é financiada pela Fundação Calouste Gulbenkian, no âmbito do projeto "Verbos e Preposições em Português Europeu" (referência 139614).

Empirical study of verbs and prepositions in European Portuguese with recourse to Web / Text Mining

João Cordeiro
INESC TEC - Porto and Universidade da Beira Interior, Covilhã

Pavel Brazdil
INESC TEC - Porto

António Leal
Universidade do Porto and CLUP

Abstract

This chapter describes our study of verbs and prepositions for European Portuguese as they are used in current articles in newspapers. The aim is to enrich the information that is available in dictionaries. This particular study focusses on verbs indicating movement. We have analyzed articles in six Portuguese newspapers and extracted more than 200 thousand of potentially relevant verb + preposition/prepositional locution cases. These were processed to identify similar cases and obtain the corresponding frequencies. Furthermore, we have also used a clustering algorithm with the objective of discovering clusters of similar verbs that are associated with similar prepositions/ prepositional locutions. Although this latest set of results is still preliminary, some similarities among verbs were uncovered already. We hope to consolidate these results in the future.

Keywords

Portuguese verbs of movement, verbs and prepositions, automatic extraction from text, clustering of verbs

1 - Introduction

Human language is a dynamic phenomenon, with variations that spread not only geographically, but also temporally. Hence, the existing dictionaries and grammars need frequent revisions. Human studies oriented towards how language is used are costly, particularly in what concerns data collection. Consequently, it makes sense to employ computer-based techniques for this task and even some of the analysis tasks that can be easily automated.

In this work, we present a case study, which is oriented towards the combination of prepositions/locutions with the verbs of movement in European Portuguese. Our goal is to characterize the usage of prepositions with the verbs, therefore creating a repository (database), which can be used for the revision of existing dictionaries, or alternatively, for the creation of new up-to-date dictionaries that can be accessed by computerized processes.

Knowing whether certain prepositions/locutions can be associated with a given verb has several pragmatic goals. First, it can serve to enhance the body of linguistic knowledge. Besides, it can help in language learning for non-native Portuguese speakers. Finally, it can be exploited in automatic spell-checkers and machine translation systems.

The existing dictionaries provide a fair amount of information, concerning verb regency for preposition/locution usage (Luft 1995). Yet, this knowledge is static. Therefore, this work aims at providing a dynamic and more complete resource to enrich the existing dictionaries (Busse 1994; Luft 1995; Borba 1990; ACL 2001, i.a.).

We began our study by focusing on a particular category of verbs – verbs indicating movement, such as: *ir* 'to go', *correr* 'to run', *saltar* 'to jump', *fugir* 'to run away', *voar* 'to fly', etc. In the future, we intend to extend the work to all verbs of the Portuguese language. This allowed us to gather an important repository of cases, like the one shown below, in which verb *fugir* 'run away' and preposition *para* 'to' are marked:

(1) *Depois de roubar um banco com outros camaradas, Rita* [*foge*] [*para*] *o outro lado e, com o beneplácito da Stasi, muda de identidade e instala-se na RDA.*

One of the objectives is to provide an electronic resource to the community enabling to list the most frequent prepositions that combine with a given verb and that distinguish it from other verbs. For example, for the verb *fugir* 'run away' we can

identify a relatively small subset of three classes of prepositions[1]: "*de*+", "*por*+", "*a*+".

We have designed and implemented an automatic system for extracting text relevant to our study, from six online Portuguese newspapers. The system collects and stores valid phrases in which combinations of Verb-Preposition occur. The collected sentences are stored in a relational database for later analysis.

The rest of this document is organized as follows. In Section 2 we provide details about some related work. The technical details on the method used are presented in Sections 3 and 4 together with the results. Section 5 presents the conclusions and discusses also possible future directions.

2 - Related work regarding verbs

2.1 - Dictionaries describing verbs

There are a number of conventional dictionaries of different kinds, which are manually created and periodically updated. Relevant examples for the Portuguese language are Busse (1994), Luft (1995), Borba (1990), and ACL (2001). From these, the dictionary of verbal regency (Luft 1995) is the most relevant to our work, as it shows verbs and the prepositions combining with it.

Figure 1 shows the prepositions that follow the verb *fugir* 'run away', which are *de*, *da*, *para*, *a*, *à*, etc. However, it is evident that these are only the most frequent prepositions following that verb, and others also quite frequent and equally relevant have been omitted. This last set of prepositions can be divided into two groups.

The first one characterizes this verb (or some similar ones) and includes some cases of *de*+ (*dele, dela*) and *por*+ (*pelo*, *pela*). The prepositions of the second group do not characterize this verb, as they may occur with many other verbs. A particular action or event can be executed *before*, *during* or *after* some other one (cases like *durante, depois de, após*). Also, they can occur in a particular location (*em*, *num*, *nas, sobre*) and together with or without some item or person (*com*, *sem*).

[1] The "+" after the preposition token indicates set of prepositions (combination of preposition with different determiners or pronouns). For example, "de+" represents "de", "do", "da", "dos", "das", "dele", "deles", "disso", "dessa", ...

FUGIR 1. TI: *fugir (de...) (para...); fugir por...* Int: *fugir.* Desviar-se ou retirar-se rapidamente (de algo ou alguém) para evitar perigo, tentação, etc.; pôr-se em fuga: *Fugir (da casa) (para a rua). Fugir pelo corredor. Fugir das más companhias. Havia perigo, era preciso fugir. "Filho de rato, foge para o palheiro"* (Prov.). // *Fugir (a, de...).* Afastar-se; ir-se perdendo de vista; distanciar-se: *O barco ia fugindo à vista, aos olhos. Fugiam dos olhos os vultos na distância. A terra ao longe fugia (da vista).* // *Fugir (de...).* Escapar(-se); soltar-se: *O pássaro fugiu (da gaiola). O preso fugiu (da prisão).* // *Fugir (a, de...)* (OBS.); *fugir(-lhe).* Abandonar; deixar; escapar; retirar-se: *Fugir da família, de casa. "... sentia o hálito vital fugir-lhe"* (Gonçalves Dias: Freire). // Evitar (afastando-se); livrar-se: *Fugir do inimigo* ou *ao inimigo* (*Fugir-lhe*). *Fugir do perigo, ao perigo* (*Fugir-lhe*) (OBS.).

Figure 1 - Verbal regencies of verb *fugir* 'run away', from the Luft (1995) dictionary

Our approach is able to identify in particular those prepositions that characterize each verb. One limitation of the dictionary of verbal regency Luft (1995) is its orientation towards Brazilian Portuguese only.

2.2 - Leipzig Corpora Collection (LCC)

The *Leipzig Corpora Collection* (LCC) is a huge collection of texts automatically gathered from many web sources, for more than 350 languages and dialects, along a period of fifteen years (Eckart & Quasthoff 2013). It includes tools for querying the corpus, especially for word co-occurrence statistics. The interface allows visualizing the words that occur frequently in the vicinity (e.g. following) of a given word. For instance, if we use its web portal[2] for Brazilian Portuguese and for the verb *fugir* 'run away', the system returns a sample of sentences in which the verb occurs, together with the respective source from which each sentence was obtained:

(2) *Afirmou que a legislação que impede a candidatura de políticos condenados por órgãos judiciários ou que renunciam ao mandato para* ***fugir da*** *cassação valoriza a moralidade pública.* (www.estadao.com.br, 13.03.2011)

(3) *Na tentativa de* ***fugir do*** *local, o homicida ainda tentou ferir outras pessoas, chegando a riscar um rapaz com a faca que ele usou contra o "Nandinho".* (www.diaadianews.com.br, 05.03.2011)

[2] The LCC portal address: http://corpora.uni-leipzig.de/ (Consulted on February 2018).

The user can also scan for the kind of words appearing before or after the given verb. In this example and if we opt for the word that follows, we get:

da (7,362), do (6,493), . (4,385), , (1,571), dos (1,432), pulando (1,415), das (1,210), correndo (733), levando (673), e (507), com (458), após (458), antes (431), dessa (39), entrando (311), à (309), para (303), ao (295), pelos (283), sem (250), disso (234), de (229), pela (219), novamente (171), pelo (168), desse (160), adentrando (158), quando (156), dela (152), mas (120), escalando (115), dele (103), em (101), usando (92), ...

That is, we obtain the terms that follow most frequently the verb *fugir*. We note that no syntactical categorization is provided, and that terms sorted by frequency are intermixed, irrespective whether they are prepositions, verbs or other syntactic category. In our study, we are especially interested in analyzing the occurrence of prepositions and this interface is not very helpful for this aim. We also note that the LCC resource is oriented towards the use of Brazilian Portuguese, which in terms of *verb + preposition* construction is different in several aspects from European Portuguese.

3 - Methods and techniques involved

Figure 2 presents a conceptual schema of our system. It includes a crawling system for news extraction from web pages, a relational database in which the extracted cases are stored and an interface for users.

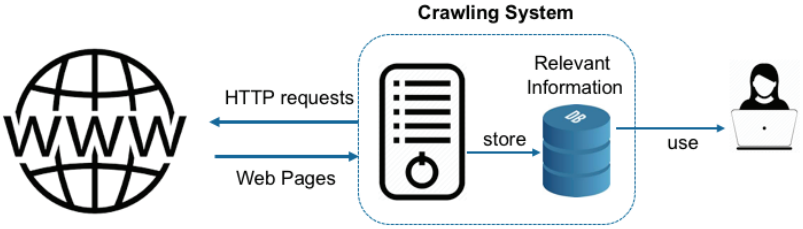


Figure 2 - General conceptual scheme of the implemented system

The central block of the schematic of Figure 2, called the "*Crawling System*", is a process that sequentially scans a set of web sites for the purpose of finding and extracting the relevant information. In our case, this information consists of the well-formed phrases that contain a certain linguistic pattern, namely verbs and prepositions.

3.1 - Web crawling

There are several web crawling strategies and almost all involve deep search in *hyperlink trees*. For example, a search engine like Google uses multiple crawlers to index all found (visible) web pages, on a periodic basis. Starting from a general set of addresses, many provided by organizations, a crawler follows the links found on the pages to get to new pages and so on (Brin 1998).

```
Given a set of Websites W={w_1,...,w_n}
for w_i ∈ W:
   crawlPage(w_i, {})

Procedure crawlPage(url, linksMemo):
   t <-- selectText(url)
   compute(t)
   for s_j ∈ subLinks(url):
      if s_j ∉ linksMemo:
       linksMemo <-- linksMemo U {s_j}
       crawlPage(s_j, linksMemo)
```

Algorithm 1 - Recursive crawling of websites

An HTML page from a website can contain thousands of hyperlinks, either pointing outside or inside the website. Here we use online pages of newspapers, in which new stories, reports and chronicles are typically added every day. Some newspapers keep a full repository of news accessible over a long period of time. These are usually organized according to a thematic hierarchy.

A crawler has to search through the hierarchy, taking care to avoid links pointing to higher levels in it, which would form a closed search loop, causing the crawling process to get stuck. The crawler must also avoid following links to pages that have already been visited in the past. It must memorize the pages that have already been visited. This corresponds to the set represented by the variable linksMemo in Algorithm 1, which captures the search strategy implemented in our system.

Our set of websites corresponds to the set of Portuguese online newspapers: *Expresso, Público, Observador, Jornal de Notícias, Diário de Notícias*, and

UrbiEtOrbi. For each newspaper, the system performs a recursive[3] search, starting with the top page, avoiding links pointing outside the newspaper and links previously visited, during the same run. In our case the depth of the recursive call was limited to a maximum of seven levels.

Figure 3 - Part of the log capturing the web crawling process for the *Observador* newspaper

Figure 3 illustrates a part of the execution of the crawling process for one of the newspapers considered. Each line corresponds to a recursive call. The first column contains the log of runtime since the start of the process. The second column ("branching state") indicates the number of hyperlinks found for each level of the call. For example, in the second column of the third line we have the expression "|361|50|101|", which means that the webpage being processed contains 361 links. Following the first link we find a page with 50 links and following the first one of these we find another page having itself 101 links, and so on. The third column shows the number of relevant cases found on the corresponding visited webpages.

This kind of output provides information about the pages that are processed and the corresponding time. The constraint of having a maximum limit for the depth of recursive calls (7) is merely pragmatic. It restricts the search space. Section 4 includes details about data volume processed and the corresponding runtimes.

In *Algorithm 1*, function "compute(t)" deals with the processing of the selected text (t) from the current webpage. We use a set of linguistic resources to look for the relevant cases in the sentences of t. However, before that, it is necessary to select the text from the webpage, corresponding to the function "selectText(url)", which involves important challenges. The approach adopted is described in the next section.

[3] A call to **crawlPage** within **crawlPage**.

3.2 - Extraction of relevant text

Most web pages, including online newspapers, have spurious text elements, which are unrelated to the main subject of the text and irrelevant. These elements include advertising (advertising slogans, etc.) and text related to the site structure, like hyperlinks and navigational text. One may refer to this kind of text as *accessory text*. This is illustrated in Figure 4.

Figure 4 - A webpage of a Portuguese online newspaper, containing examples of accessory text and news of interest

Given the high variability of the structure and style adopted in web sites, accessory text represents a challenge to any automatic text extraction system.

Our aim is to extract relevant text and discard the accessory one. The work of Pedrosa (2011) addresses this issue, focusing on the automatic elimination of readers' comments to news and advertising. The method is based on the occurrence of certain *keywords*, indicative of the presence of such comments. However, as it was reported by the author, this approach is not easily generalizable to other newspapers or websites, since each one may follow a particular structure and use specific keywords. Thus, this method would require a manual readjustment before being applied to a new site.

The approach that was adopted here analyses the HTML structure of web pages with recourse to existing software packages, such as *jsoup* (Hedley 2017), for

the *Java* programming language, or *rvest* (Wickham 2016) for the *R* language. The tree like structure can be analyzed and the most promising zones for relevant text extraction identified.

Unfortunately, there are websites that do not delimit the text well with appropriate tags. For example, in many cases paragraph tags ("<p>") are absent from true paragraphs in the text. Therefore, we cannot rely exclusively on the HTML structure of a page, as a way to select the relevant text content. It is necessary to employ other techniques as well.

In this work we have applied a heuristic, which despite being rather simple, works quite well, satisfying the practical needs of the project. We have observed that a well-written content sentence tends to be well punctuated, ending with a period, an exclamation or question mark. In contrast, most accessory text is often not well punctuated, having no punctuation at all in almost all cases. We noticed that in the web this is a prevalent pattern. Therefore, for each newspaper webpage we only select well punctuated sentences and with a given minimum number of words (this parameter was set to 3).

In the future we could adopt a more elaborated approach that would require that the text contain informative words, as judged by TFIDF metric (Salton & Buckley 1988; Bruno & Cordeiro 2015), or topics identified using LDA (Blei 2012; Blei, Ng, & Jordan 2003) and other approaches of topic modeling.

3.3 - Localization of verbs and prepositions

The identification of the relevant sentences for extraction requires that we identify certain verbs and prepositions/locutions occurring in them. As our study focuses on movement verbs, we have prepared a list of such verbs beforehand (see Figure 5). Similarly, we have prepared a list of prepositions and locutions (see Figure 6).

abaixar	apressar-se	cambalear	desacelerar	encaminhar-se	formigar
abaixar-se	apressurar	caminhar	desagregar-se	encarreirar	formiguejar
abalar	apropinquar	canoar	desamparar	encruzar	fugir
abandonar	aproximar-se	carambolar	desandar	encruzilhar	fundear
abeirar	arrastar-se	carregar	desaparecer	engolfar	galgar
abeirar-se	arrecuar	cavalgar	descair	enovelar	galopar
aboiar	arredar-se	caçar	descer	enrolar-se	galopear
acalçar	arremessar-se	cercar	descolar	enroscar-se	gandaiar
acelerar	arremeter	chefiar	desembestar	enterrar	gingar
acercar-se	arrimar-se	chegar	desertar	entornar-se	girar
achegar	arrojar	chegar-se	desfilar	entranhar-se	girogirar
achegar-se	arrojar-se	chispar	desgalgar	entrar	gotear
acompanhar	ascender	circuitar	desligar-se	envolver	gotejar
acuar	assomar	circular	deslizar	enxamear	gravitar
adejar	atabular	circundar	deslocar	erguer-se	guiar
adiantar-se	aterrar	circunvagar	deslocar-se	errar	guindar-se
afastar-se	atingir	circunvalar	despedir-se	escalar	imergir
aflorar	atirar-se	claudicar	despegar-se	escapar-se	ingressar
afundar	atravessar	comboiar	despenhar	escapulir-se	intersectar
afundar-se	ausentar-se	competir	despenhar-se	escoar-se	introduzir-se
afundir	avagarar	conduzir	destrepar	escoltar	inverter
afundir-se	avançar	contorcer	desunir-se	escorregar	inverter-se
alar	aviar-se	contorcer-se	desviar-se	esgueirar-se	investir
alcantilar	avizinhar	contornar	devanear	espalhar	ir
alcançar	avizinhar-se	correr	dimanar	espalhar-se	ir-se
aligeirar	bailar	corrupiar	dirigir	esparrar-se	isolar-se
altear	baixar	coxear	dirigir-se	espinotear	içar-se
altear-se	balançar	cruzar	distanciar-se	esquiar	jornadear
alvorar	bamboar-se	cursar	divagar	estatelar-se	lançar-se
alçar-se	bambolear-se	dançar	elevar	estender-se	larear
amarinhar?	bandurrar	deambular	elevar-se	estugar	largar
andar	bandurrear	debandar	emanar	esvoaçar	laurear
andarilhar	boiar	declinar	emergir	evadir-se	levantar-se
andejar	bordejar	deixar	encabeçar	exceder	levar
antecipar-se	cabriolar	derramar-se	encaixar-se	flanar	liderar
apartar-se	cair	derrapar	encalçar	fluir	locomover -se
apressar	calcorrear	desabar	encaminhar	flutuar	manar

Figure 5 - Partial list from the 382 movement verbs

a	ao	da	debaixo de	doutras	exceto
a alguns passos	ao lado de	dacolá	defronte de	doutrem	face a
de	ao longo de	dalgo	dela	doutro	feito
a bel-prazer de	ao pé de	dalgum	delas	doutros	fora
a braços com	ao redor de	dalguma	dele	dum	fora de
a caminho de	aonde	dalgumas	deles	duma	frente a
a conselho de	aos	dalguns	dentro de	dumas	graças a
a despeito de	apesar de	dalgures	dentro em	duns	junto a
a dois passos de	após	dalguém	depois de	durante	junto de
a fim de	após de	dali	desde	dês	longe de
a julgar por	aquém de	dalém	dessoutra	em	mais
a mais de	através de	daquela	dessoutras	em baixo de	mediante
a meio caminho	atrás de	daquelas	dessoutro	em caso de	menos
de	até	daquele	dessoutros	em cima de	mercê de
a meio de	até a	daqueles	desta	em favor de	na
a menos de	cerca de	daqueloutra	destas	em frente a	na conta de
a par com	co	daqueloutras	deste	em frente de	nalgum
a par de	coa	daqueloutro	destes	em lugar de	nalguma
a partir de	coas	daqueloutros	destoutra	em prol de	nalgumas
a respeito de	com	daqui	destoutras	em razão de	nalguns
a seguir a	com base em	daquilo	destoutro	em redor de	naquela
a um passo de	como	daquém	destoutros	em torno de	naquelas
abaixo de	conforme	das	detrás de	em troco de	naquele
acerca de	consoante	daí	diante de	em vez de	naqueles
acima de	contra	de	disso	em via de	naqueloutra
adiante de	cos	de acordo com	disto	em vias de	naqueloutras
afora	cum	de caras com	do	embaixo de	naqueloutro
além de	cuma	de cima de	donde	enquanto a	naqueloutros
ante	cumas	de conformidade	dos	entre	naquilo
antes de	cuns	com	doutra	excepto	nas

Figure 6 - A partial list of 269 prepositions and prepositional locutions considered

For each sentence, it is necessary to verify whether a particular movement verb occurs in that sentence together with a preposition or prepositional locution on the right-hand side.

Another possibility would be to use here a part-of-speech tagger, for example the *LX-Parser* (Silva, Branco, Castro & Reis 2010), that provides an automatic analysis of the constituents of each sentence and could provide a tagging of verb-preposition combinations. However, as our focus was restricted to only certain types of verbs (i.e. movement verbs), which could easily be represented in the form of a list, we did not resort to such tools.

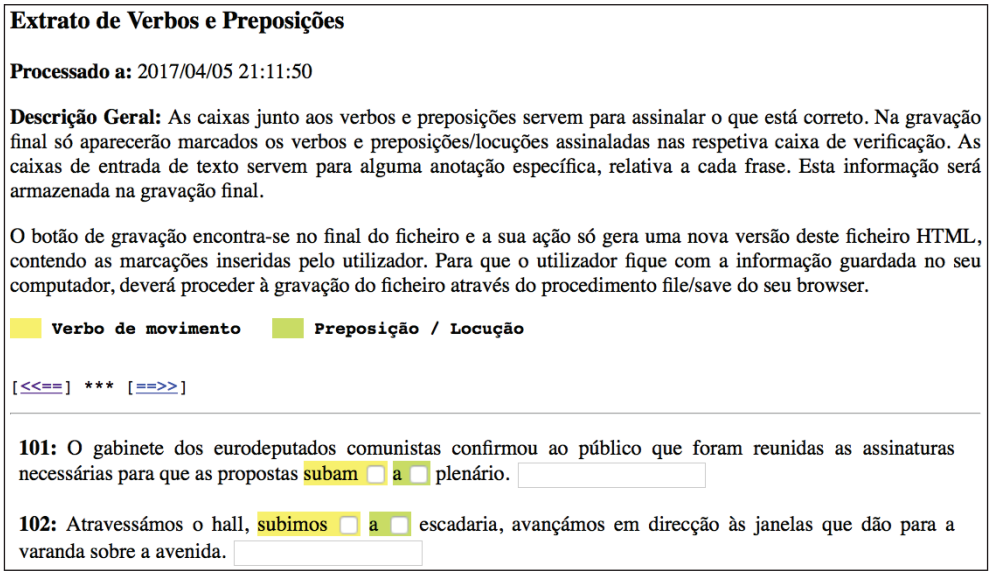
Furthermore, our aim was to analyze not only prepositions, but also a rich set of prepositional locutions (consisting of more than one token), for which the PoS tagger might not provide an adequate solution. This reinforced our belief that using an explicit list represents a good choice here.

We note that the list of verbs used (Figure 5) includes verbs in the infinitive form, contrary to what occurs mostly in the text, in which verbs are usually conjugated, regarding person, number, tense and mood. To overcome the difficulty of matching two forms that are slightly different (ex. *fugir* versus *fugiu*), we have employed an automatic lemmatizer for Portuguese, namely the ***Unitex*** lemmatizer (Muniz 2004).

3.4 - Visualization and interactive correction of cases

The process of identifying the combinations of prepositions or verbs in text is not entirely reliable. In order to obtain a better-quality result, the combinations extracted need to be analyzed by specialists. The aim is twofold. One aim is to identify the badly marked cases (prepositions or verbs). The other one is to obtain a cleaned-up dataset that can be used in further analysis.

Our system includes thus an interface that allows selecting, viewing and marking different cases. The interface is shown in Figure 7.


Extrato de Verbos e Preposições

Processado a: 2017/04/05 21:11:50

Descrição Geral: As caixas junto aos verbos e preposições servem para assinalar o que está correto. Na gravação final só aparecerão marcados os verbos e preposições/locuções assinaladas nas respetiva caixa de verificação. As caixas de entrada de texto servem para alguma anotação específica, relativa a cada frase. Esta informação será armazenada na gravação final.

O botão de gravação encontra-se no final do ficheiro e a sua ação só gera uma nova versão deste ficheiro HTML, contendo as marcações inseridas pelo utilizador. Para que o utilizador fique com a informação guardada no seu computador, deverá proceder à gravação do ficheiro através do procedimento file/save do seu browser.

Verbo de movimento **Preposição / Locução**

[<<==] *** [==>>]

101: O gabinete dos eurodeputados comunistas confirmou ao público que foram reunidas as assinaturas necessárias para que as propostas subam a plenário.

102: Atravessámos o hall, subimos a escadaria, avançámos em direcção às janelas que dão para a varanda sobre a avenida.

Figure 7 - A part of a web page showing some retrieved cases to be marked

The top section of each page contains a descriptive header to help the user. Each page includes a maximum of 50 cases. There are also navigation links allowing going to the previous or subsequent page. This is done using arrows at the end of the header. At the end of each case/sentence there is a text box, enabling the specialist to confirm that the case is correct and hence should be kept. Each page has a button for saving the marked cases.

The erroneous cases identified could be used in future to enhance the functioning of our system. The knowledge of these cases can be synthesized by a human expert, by indicating rules and exceptions to be considered. Alternatively, the erroneous cases could be supplied as input to a machine learning system that can be trained to identify such cases in future.

4 - Implemented system and results

This work has already allowed us to collect a quite extensive number of examples and of case studies, showing the effective use of prepositions/prepositional locutions with verbs of movement for the European Portuguese, as illustrated in some examples below:

(4) 1. *Os problemas de dependência ao jogo* [*surgem*] [*em*] *cinco anos, no máximo numa década.*
2. *Naquele momento, Eder, o herói,* [*desceu*] [*ao*] *nível de toda uma nação.*
3. *O gabinete dos eurodeputados comunistas confirmou ao Publico que foram reunidas as assinaturas necessárias para que as propostas* [*subam*] *[a] plenário.*
4. *O despiste de um autocarro na madrugada de domingo na Estrada Nacional 79, na direcção Mâcon-Moulins, França, provocou quatro mortos, três feridos graves e 25 ligeiros, que* [*seguiam*] *no veículo* [*a caminho de*] *Genebra, Suíça.*

The verbs are marked in yellow and the corresponding prepositional locutions/ prepositions in green.

4.1 - Implemented system

The relevant cases are stored in a *relational database*, in which we store the extracted sentences, the web pages to which they belong and the combination of "*verb + preposition/prepositional locution*" occurring in the sentence. The Entity-Relationship (ER) diagram (Elmasri & Shamkant 2010) of our database is shown in Figure 8.

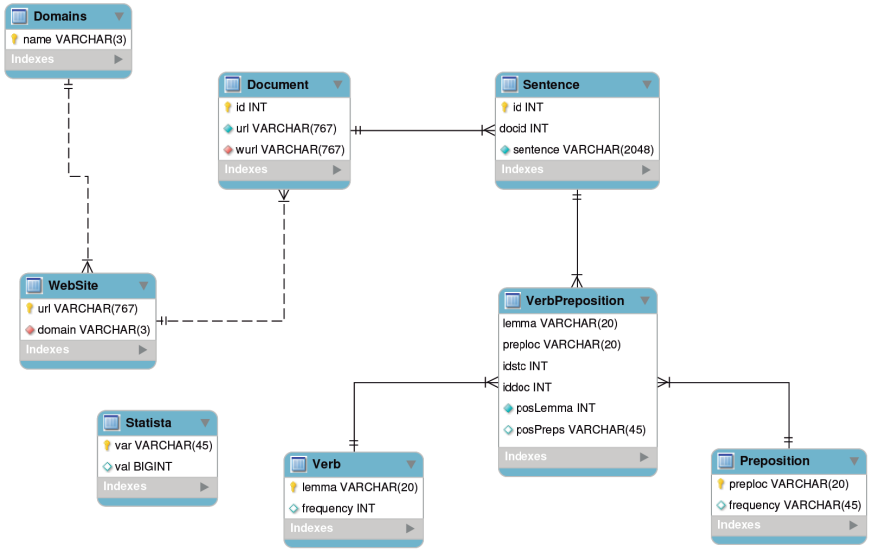


Figure 8 - The Entity-Relationship diagram (E-R) of the VPLPBD database.

The central table of the database is *VerbPreposition* that associates three fundamental entities: the verb, the preposition and the sentence, also represented by three tables (*Verb*, *Preposition* and *Sentence*) with the same names.

The *Sentence* table stores the relevant cases extracted from the news, sentences in which combinations of verbs occur with the prepositions and the prepositional locutions contained in the list.

The *Document* table represents a news story from an online newspaper whose website reference is stored in the *WebSite* table.

The *Domains* table stores the online newspaper domains, which so far are only from the ".pt" domain. Other domains may be entered in the future, such as ".org".

The *Verb* and *Preposition* tables store the lists of verbs and prepositions we are working with (parts of them were shown in Figures 5 and 6).

Our study used data gathered for six Portuguese newspapers - *iOnline*, *Publico*, *Expresso*, *DN*, *Observador*, *UrbiEtOrbi* - in the period between the 28th and 30th of March 2017. Some summary statistics regarding this process are shown in Table 1.

Execution features	**outcomes**
Number of tokens involved	15 206
Number of sentences processed	1 599 423
Number of documents processed	52 032
Number of relevant cases extracted	226 337
Total execution time	21h49

Table 1 - Descriptive statistics related to processing six Portuguese newspapers

The number of relevant cases identified represents 14.15% of the total sentences processed. Table 2 shows the number of pages analyzed and extracted cases for each of the six newspapers considered.

Online Newspaper	Analyzed Pages	Proc. Time	Extracted Cases	%
www.ionline.pt	2 725	1:12:43	7 914	3,50%
www.publico.pt	19 952	8:16:30	96 862	42,80%
expresso.sapo.pt	4 872	2:32:54	16 684	7,37%
www.dn.pt	3 648	1:35:31	10 615	4,69%
observador.pt	15 206	5:34:54	74 465	32,90%
www.urbi.ubi.pt	5 629	2:37:02	19 797	8,75%
Totals	52 032	21:49:34	226 337	100%

Table 2 - Number of pages analyzed and extracted cases per each newspaper

We note here that the majority of the extracted cases (75%) came from newspapers *Público* and *Observador* and consequently a great proportion of extracted cases come from these two sources.

4.2 - Sequentially ordered results

The data collected from the web are stored in our relational database, allowing us to obtain various relevant statistics, as well as to select specific combinations of verb and prepositions, which are important for further study.

lemma	preploc	posLemma	substr(sentence,1,120)
chegar	à altura de	0	Chegara à altura de responder aos ataques da coligação e por fim à polémic
deixar	a braços com	9	Para o conseguir, separavam as duas funções, deixando os bombeiros sobretu
sair	a braços com	3	Com portugal recentemente saído de um resgate e ainda a braços com uma ape
seguir	a caminho de	49	O despiste de um autocarro na madrugada de hoje na estrada nacional 79, na
seguir	a caminho de	49	O despiste de um autocarro esta madrugada na estrada nacional 79, na direçã
seguir	a caminho de	51	O despiste de um autocarro na madrugada de domingo na estrada nacional 79,
seguir	a caminho de	49	O despiste de um autocarro na madrugada de domingo na estrada nacional 79,
viajar	a caminho de	5	Há indicações de que ele viajou via holanda a caminho de lyon, disse wim d
ir	a conselho de	18	A versão final do decreto-lei que transforma a adse num instituto público
voltar	a conselho de	44	Numa altura em que muitos temiam que o museu de évora passasse a ser gerid
vir	à custa de	17	A questão que se tem colocado é se a capacidade de se concentrar em várias
passar	à direita de	9	Aí, e na queda a pique, ou passa os comandos à direita de sempre ou se des
abandonar	a fim de	3	O reino unido abandona, assim, definitivamente o mercado único, a fim de r
conduzir	a fim de	38	Em cima do acontecimento, o vespertino "a capital" contou: "o batata, que
investir	a fim de	7	O exército dos estados unidos anda a investir em força na impressão 3d, a
passar	a fim de	81	A unita, maior partido da oposição angolana, pediu, esta segunda-feira, um
seguir	a fim de	43	Face às palavras de hirose " que não era presidente da tepco aquando do ac
chegar	à frente de	33	Da américa profunda às grandes cidades, as caravanas -- onde se destacam o

Figure 9 - Some data stored in the VPLPBD *VerbPreposition* table

The table shown in Figure 9 presents a small portion of the data stored in the *VerbPreposition* database table. For each line, the first two columns show the extracted case. The third column shows the position (in number of words) in which the verb occurs in the sentence. The last (4th) column shows the initial part of the sentence. For example, in the line referring to the verb *abandonar*, we see that it appears conjugated (*abandona*) in position 3 of the sentence (the count begins at zero).

When analyzing the occurrence of prepositions and verbs in the extracted cases, we note that we are dealing with long tail distributions (Zipf's law, or power law), similar to what happens with the occurrence of words in corpora. Figure 10 shows the distribution for the subset of the 30 most frequent verbs and prepositions/ prepositional locutions ordered by frequency.

Verbos

rank	lemma	frequency	Percent. (%)
1	ir	50689	22.69111
2	passar	20677	9.25613
3	chegar	16164	7.23587
4	vir	13811	6.18254
5	deixar	10917	4.88703
6	levar	10065	4.50563
7	voltar	9094	4.07096
8	sair	7840	3.50960
9	entrar	7200	3.22311
10	surgir	6434	2.88020
11	seguir	6308	2.82380
12	andar	6210	2.77993
13	tornar	5574	2.49522
14	avançar	4277	1.91461
15	partir	3044	1.36266
16	subir	2701	1.20911
17	cair	2582	1.15584
18	regressar	2438	1.09138
19	mudar	2266	1.01438
20	correr	2038	0.91232
21	envolver	1982	0.88725
22	atingir	1962	0.87830
23	acompanhar	1475	0.66029
24	liderar	1285	0.57523
25	fugir	1171	0.52420
26	viajar	1140	0.51033
27	virar	1091	0.48839
28	dirigir	1055	0.47227
29	descer	1026	0.45929
30	conduzir	945	0.42303

Preposições

rank	preploc	frequency	Percent. (%)
1	a	63500	28.42601
2	de	27705	12.40224
3	para	14577	6.52545
4	em	11843	5.30156
5	no	11706	5.24023
6	da	9343	4.18243
7	ao	9343	4.18243
8	com	8704	3.89638
9	na	7835	3.50737
10	por	7640	3.42007
11	do	7552	3.38068
12	mais	6163	2.75889
13	pela	3171	1.41951
14	como	2913	1.30402
15	aos	2421	1.08377
16	dos	2375	1.06318
17	pelo	2005	0.89755
18	às	1953	0.87427
19	até	1878	0.84069
20	nos	1848	0.82726
21	das	1711	0.76594
22	nas	1515	0.67820
23	num	1355	0.60657
24	entre	1301	0.58240
25	numa	1207	0.54032
26	sobre	1141	0.51077
27	sem	950	0.42527
28	desde	742	0.33216
29	depois de	699	0.31291
30	durante	640	0.28650

Figure 10 - The distribution of some of the verbs and prepositions in our dataset

We observe that the first six verbs occur in about 53% of the cases. In case of prepositions, the mass concentration at the top is still higher, with only the first four occurring in about 52% of the cases.

Not all verbs listed in Figure 5 occurred in our dataset, some are rare and do not occur, even after processing near 1.5 million sentences. From our list of 398 verbs, 198 (49.75%) occurred at least once in an extracted case. The other ones

were not encountered at all. The situation for prepositions/ prepositional locutions is somewhat similar. From our list of 269 prepositions, 171 (63.57%) occurred at least once in the extractions. Here, the percentage is a bit higher.

4.3 - Results for individual verbs

The results presented in the previous section were reorganized. The aim was to join all cases relative to each verb. So, for instance, for verb *fugir* we obtain the following prepositions and their frequencies:

Verb	Preposition	Frequency	Rel. Frequency
fugir	a+	231	33.82%
	com+	42	3.59%
	de+	586	50.04%
	no+	41	3.50%
	por+	205	17.51%
	Others	66	5.64%
	Total	**1171**	**100.00%**

Table 3 - Frequency of preposition sets for the verb *fugir* 'run away'. Each set contains a number of related prepositions, for instance, "a+" represents "a", "ao", "à", "às", etc.

For a great proportion of the 382 movement verbs used by our extractor, various prepositions were found co-occurring with those verbs. From our list of 269 prepositions, 171 were found co-occurring with the verb *fugir*, although their frequencies of occurrence were quite different. Table 3 shows several sets of such prepositions co-occurring with the verb *fugir*. Each set class represents a subset of prepositions of a similar type. This provides valuable information to linguists for further analysis and comparisons with existing information. For instance, it may be possible to characterize a verb based on its distribution over prepositional sets.

4.4 - Verb clustering

The collected data was used to conduct an experiment to discover clusters of verbs based on their association with prepositions. This was done with recourse to a method of automatic clustering. The training data was arranged in a form of a table, in which each line contains the lemmatized verb, which is followed by frequencies of prepositions that co-occur with that verb. This is exemplified in Table 4.

Lema	às	até	atrás de	cerca de	com	sem	senão	sob	sobre	trás	visto
ir	466	593	5	123	2805	142	4	24	402	0	7
passar	29	43	0	30	388	98	3	6	73	0	1
chegar	665	122	1	21	366	48	1	3	15	0	0
vir	55	93	7	7	523	26	0	3	74	0	1
deixar	35	30	7	15	196	128	0	3	77	1	0
levar	71	143	0	23	136	23	0	0	15	0	0
voltar	131	3	0	4	134	7	0	1	8	0	1
sair	57	13	0	12	233	45	0	5	30	0	0
entrar	12	5	0	2	140	27	0	3	0	0	0
surgir	11	8	2	21	328	30	0	11	80	0	0

Table 4 - Example of training data used for clustering
(only a small portion of verbs and prepositions are presented)

Each line is considered as a training instance in which individual prepositions represent the attributes. The complete dataset contains 198 instances (verbs) and 171 attributes.

EM[4] clustering algorithm (Dempster, Nan & Donald 1977) was used to determine the best number of clusters. This system recommended the 3 clusters to be used. A subsequent application of k-Means (with K=3) generated clusters with centroid words *ir*, *levar*, and *passar*. The dataset was preprocessed, and all values became normalized. The clustering metric employed was the *cosine similarity*. The verbs more related to the centroid words were:

ir ==> *circular*, *cair*, *desfilar*, *dançar*, *marchar*, *afundar*, *voar*, ...
levar ==> *emergir*, *propalar*, *progredir*, *surgir*, *flutuar*, ...
passar ==> *passear*, *espalhar*, *errar*, ...

[4] The Expectation Maximization clustering algorithm.

Some affinities are evident (e.g. [*ir*, *cair*], as both verbs select complements that are prepositional phrases headed by *a*) and others not so much. We believe that a dataset that includes a carefully selected set of features should lead to better results. We hope to achieve a better understanding regarding different classes of verbs.

5 - Conclusions and future work

This work was oriented towards the study of verbs and prepositions in European Portuguese as they are used in current articles in newspapers. We have analyzed articles in six Portuguese newspapers and extracted about 226 thousand of potential relevant *verb + preposition* cases.

For each of the 382 verbs related to movement, we have provided a list of all possible prepositions that were encountered together with their frequencies. This provides valuable information for further analysis.

The extracted cases were processed by a clustering algorithm in order to discover clusters of similar verbs that are associated with similar prepositions. Although this was quite a preliminary study, some similarities between verbs were uncovered by this process, like for example [*ir*, *cair*], [*passar*, *passear*]. A larger dataset will help us to consolidate these results, as well as discover new ones. We can then compare these results to what is currently known and described in grammars of Portuguese language.

Future work

A small percentage of extracted cases were false positives. They represent cases that were identified erroneously by the system as *verb + preposition* group. A sample of the extracted cases were manually analyzed and tagged by human experts. The outcome is useful in two ways. First, the cleaned dataset can be used in further processing and/or analysis. Second, the two categories of examples (true positive and false positive cases) could be used to train a classifier for a more advanced extraction system. This would lead to a more advanced version of the current system. The learning process could continue, as long as new-tagged examples are provided, thereby improving continuously its accuracy.

So far, the system targeted only a few Portuguese mainstream newspapers, but there are a wide range of other possibilities, from regional newspapers to online

magazines and blogs. Expanding the text input sources, combined with accurate extraction methods, will certainly yield better results, both in quantitative or qualitative terms.

A more user-friendly front-end for interacting with the continuously stored information is also being designed and implemented. It will allow users to search for usage of certain patterns and count specific *verb + preposition* combinations, in a more user-friendly fashion than what current dictionaries provide nowadays.