INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

# Validation of Lexical-Syntactical Matrices

**Fernando Miguel Filipe Gomes**

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática e de Computadores

## Júri

| | |
|---|---|
| Presidente: | Doutora Ana Maria Severino de Almeida e Paiva |
| Orientador: | Doutor Nuno João Neves Mamede |
| Co-orientador: | Doutor Jorge Manuel Evangelista Baptista |
| Vogal: | Doutora Maria Luisa Torres Ribeiro Marques da Silva Coheur |

**Novembro de 2009**

# Agradecimentos

Gostaria de agradecer o tempo e dedicação dos meus orientadores Prof. Nuno Mamede e Prof. Jorge Baptista, que em muito me ajudaram.

Gostaria também de agradecer a toda a equipa do $L^2F$, especialmente ao Tiago Luís pela sua ajuda no que diz respeito às ferramentas Hadoop e Condor.

Por fim também agradeço ao meu colega de trabalho Ricardo Portela, com quem partilhei reuniões e discuti ideias e soluções ao longo deste trajecto.

Lisboa, 8 de Novembro de 2009

Fernando Gomes

Aos meus Pais

# Resumo

Esta tese focou-se na validação de matrizes léxico-sintácticas, de forma a verificar se a informação nelas contida é correcta e pode ser usada posteriormente. A validação é baseada em comparações estatísticas entre resultados obtidos através de testes em grandes quantidades de dados (corpus), e da informação contida nas matrizes. Esta informação consiste em propriedades lexicais.

As propriedades validadas são de natureza morfológica, distribucional e transformacional. Cada uma delas tem de ser verificada individualmente e por processos distintos.

A comparação estatística foi efectuada com o auxílio de computação em GRID e de *software* de calendarização e programação paralela. Finalmente, foi feita uma avaliação do trabalho efectuado para validar os resultados.

# Abstract

This thesis focuses on the validation of lexical-syntactical matrices, in order to verify if the information they contain is correct and can therefore be used in future endeavors. The validation is based on a statistical comparison between results obtained from a large *corpus* and the information contained in the matrices. This information consists on properties of lexical items.

The properties that were validated are of morphological, distributional and transformational nature, and each of them must be verified individually through distinct processes.

The statistical comparison is done with the aid of GRID computing, as well as scheduling and parallel programming software. Finally an evaluation of the work is performed to check the findings.

# Palavras-Chave
# Keywords


## *Palavras-Chave*

Matrizes léxico-sintácticas

Comparação estatística

Corpus

Propriedades transformacionais

Propriedades distribucionais

## *Keywords*

Lexical-syntactical matrices

Statistical comparison

Corpus

Transformational properties

Distributional properties

# Índice

# Lista de Figuras

# Lista de Tabelas

# 1 Introduction

## 1.1 Motivation

Information is the most important commodity in today's world, not only to researchers in various fields but also to corporations, private and public institutions, among others. In this context we are beginning to realize that information retrieval is one of the most important fields in computer engineering today.

Many linguistic studies have produced linguistic data and some of them may be converted in a tabular format, which is called lexical-syntactical matrices. However, as some of this data results from acceptability judgments and are not directly derived from corpus evidence, it may lack linguistic adequacy, or require empirical validation.

Available linguistic data can be found, for example, for verbal (Oliveira, 1984), adjectival (Freire, 1994; Casteleiro, 1981) and nominal (Ranchhod, 1990; Baptista, 2005) constructions. A large segment of the Portuguese language has therefore received careful and detailed description. To our knowledge, the relation between this highly theoretical descriptions and linguistic corpus-based evidence has not been systematically undertaken. Tools that enable such confrontation are required, and this is one of the objectives of this dissertation.

By definition we characterize a matrix as "a rectangular display of features characterizing a set of linguistic items, especially phonemes, usually presented as a set of columns of plus or minus signs specifying the presence or absence of each feature for each item"[1]. The linguistic data here studied consists of numerous properties, such as the presence of *Nhum* (human nouns) in a given syntactic slot (e.g. subject). Properties here considered include morphological, distributional and transformational attributes of lexical items.

One important motivator in this work is the possibility, in the future, of integrating in the $L^2F$/INESC-ID[2] chain (Mamede, 2007) information from the matrices, allowing for more accurate processing results.

Finally, another strong motivation behind the choice of this thesis is that any results we find can

[1] http://dictionary.reference.com/browse/matrix

[2] Laboratório de Sistemas de Língua Falada do Instituto de Engenharia de Sistemas e Computadores - Investigacão e Desenvolvimento.

help future research, not only on the computational field, but also on NLP and linguistics.

## 1.2 Objectives

The main objective of this thesis is the validation of the lexical-syntactical matrices, the evaluation of the methods here used and the processing of the corpus needed to perform that validation. Other objectives to be pursued are the manual validation of the data on the matrices and experimentation of different methods to validate the matrices, in order to achieve maximum efficiency.

The validation step is divided into two main stages, the first one identifies the set of properties found in the matrices with the dependencies and nodes extracted by the $L^2F$ chain. The second stage consists in searching the dependencies/nodes of the tabled words.

In spite of the importance accorded to the validation task, both the corpus processing and the evaluation of the entire process are essential, since it would be impossible to validate anything without a processed corpus, and a scientific endeavor should not take place without an evaluation.

## 1.3 Strategy

The first step is the processing of a *corpus* (CETEMPúblico (Santos & Rocha, 2001)) using the processing chain of $L^2F$, featuring XIP (Xerox Incremental Parser. (R.C.E. Xerox, 2003)), on a GRID computing framework (basically a cluster of networked computers, acting in concert). Essentially the use of the GRID consists in dividing the processing of the input by several computers, allowing the combined computational effort of that cluster to obtain faster results. Next, a relation is established between the properties of the matrix and the dependencies/nodes extracted from the *corpus*. Afterwards, a search is conducted over the processed text for the dependencies/nodes related to the words present in the matrices (as well as numerical results, such as number of occurrences). Finally, an evaluation of the matrices is made, via a comparison between the results from the test corpus and the information in the matrices.

## 1.4 Tools Used

The tools used in this process are the processing chain of $L^2F$, the $L^2F$ GRID and the Condor and Hadoop systems that work on the AFS[3]. The $L^2F$ chain is used to process the corpus into an XML file (showing the properties of each node and the dependencies between them), in order to show the

[3] Andrew File System.

input text with lexical, morphological and syntactical information needed to perform its analysis. The GRID is used to accelerate the processing of the corpus by using several computers at the same time, and the file system is used to create a stable way to run the corpus in the $L^2F$ chain. Condor manages the scheduling of the GRID use during the corpus processing, by queuing and prioritizing processes, while Hadoop is used to perform jobs on large data (like the output of the processed corpus), while maintaining a high-throughput data access.

### 1.4.1 Processing Chain of $L^2F$

The natural language processing chain available in $L^2F$ (Mamede, 2007)(Figure 1.1) is basically a set of different tools that do specific tasks, like tokenization or morpho-syntactical tagging, on an input text and transform it into relevant lexical and syntactical information (be it in tree or XML form).

The various tools present in the chain are:

- Tokenizer: Divides the text into individual parts (tokens) and also identifies certain textual elements, like numbers and punctuation;
- Palavroso (Medeiros, 1995): Labels the tokens with POS (parts-of-speech), like noun or adverb;
- Sentence Divider: Divides the input text into sentences;
- RuDriCo (Pardal, 2007): A rule-based morpho-syntactical disambiguator, which groups segments, reconstitutes contracted words, changes lemmas and applies other morpho-syntactical rules;
- Marv (Ribeiro et al., 2003): A morpho-syntactical statistical disambiguator based on the Viterbi algorithm, which chooses one tag for each word based on its category or subcategory and its context;
- XIP (R.C.E. Xerox, 2003): The last step in the chain is the syntactical parser XIP, which introduces lexical information, applies local grammars, segments the text into *chunks* and finds the dependencies between them.

The process for a given input *O Rui perguntou que horas são no Japão. No Japão são 10 horas, respondeu o António.* 'Rui asked what is the time in Japan. In Japan it is 10 o'clock, answered António.' begins with the Tokenizer tool dividing the text into individual elements (*O, Rui, perguntou*, etc.) and identifies "." as punctuation and "10" as a number. Next, Palavroso takes over and labels the words with POS and their respective fields (for instance *Rui* is labeled as a proper noun, with the values singular and masculine for the fields number and gender). After this, the Sentence Divider separates the input text into two sentences using the punctuation element "." identified by the Tokenizer: *O Rui perguntou que horas são no Japão.* and *No Japão são 10 horas, respondeu o António.*.

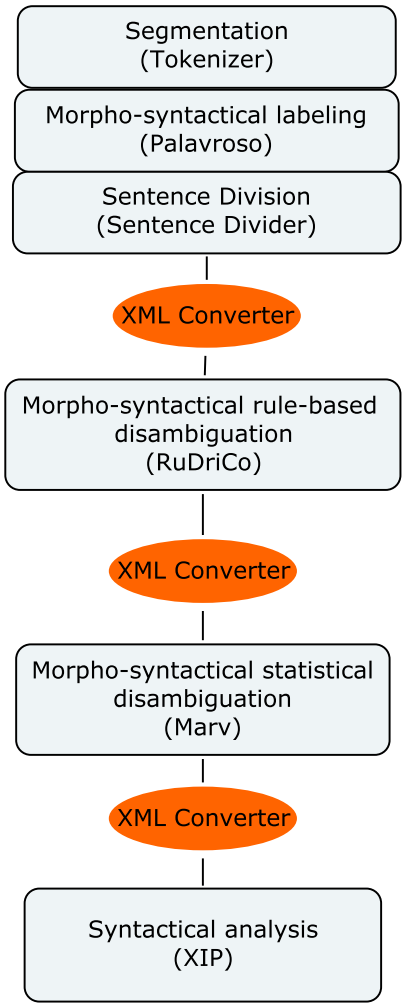



Figura 1.1: $L^2F$ Lexical Processing Chain

Afterwards, the results from the Sentence Divider will be transformed into XML by an XML Converter in order to use RuDriCo, the morpho-syntactical disambiguator. Here a certain number of contracted words are split into individual lexical items (e.g. *No* becomes *Em* + *o*) and other morpho-syntactical rules are applied.

After being again modified by an XML converter, the output resulting from RuDriCo is fed into Marv, beginning the statistical phase of the morpho-syntactic disambiguation process. In this step a single label is chosen for each individual element, based on the Viterbi algorithm. For instance, the word *são* can be both a conjugated form of the verb *ser* as well as an adjective. In this case, Marv decides that given the context *são* is a verb.

After another conversion, this time from the Marv to the XIP format, the syntactical parser XIP begins its processing of the data by applying local grammars, adding lexical information and by segmenting the strings of words into word chunks, i.e., by grouping words into blocks; for example *O Rui* is grouped into a chunk and labeled as a noun phrase (NP).

XIP also specifies dependencies between words and chunks in the sentences, like the DETD (definite determiner) relation between *O* and *Rui*, that establishes that *O* is a definite determinant for the noun *Rui*.

Figure 1.2 shows the syntactical trees on a chain output for the example sentences, and Figure 1.3 shows the dependencies.

### 1.4.2 Condor

Condor (Condor Team, 2008) is a scheduling application from the Condor Research Project, at the University of Wisconsin-Madison, that works on many different file systems (among them the AFS and the HDFS[4]). It is mostly used on tasks that require large amounts of processing, since it queues all the processes, attributes priorities to them and manages the resources until the task is completed. Afterwards, the user is informed that the task has been completed and all computational resources are released.

Thanks to this application, it is possible to use the $L^2F$ GRID to process the CETEMPúblico[5] corpus in a manageable time, since its "flocking" technology allows the various Condor applications in the GRID's computers to work as one. Besides this, Condor possesses many more advantages, such as:

- optimization of computational resources by allowing many processes to be run at the same time;

[4]Hadoop Distributed File System.

[5]corpus de aproximadamente 180 milhões de palavras em português europeu, criado pelo projecto Processamento Computacional do Português

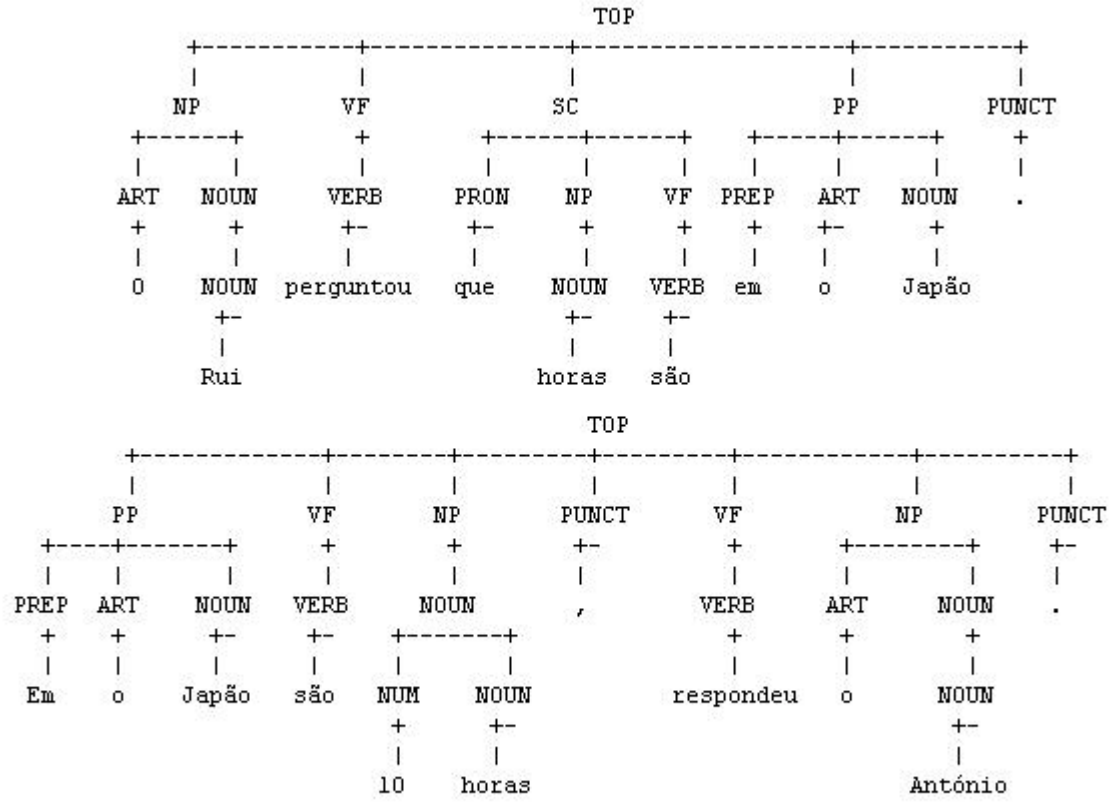



Figura 1.2: Syntactical trees for the sentences *O Rui perguntou que horas são no Japão* and *No Japão são 10 horas, respondeu o António.*

```
MAIN(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>)
DETD(<NOUN:Rui|Rui|2|4|Np***sm***|hmn|>,<ART:O|o|0|0|Td***sm***|hmn|>)
DETD(<NOUN:Japão|Japão|33|37|Np***sm***|hmn|>,<ART:o|o|30|31|Td***sm***|hmn|tokenend|>)
PREPD(<NOUN:Japão|Japão|33|37|Np***sm***|hmn|>,<PREP:em|em|30|31|S****__**s|hmn|tokenstart|>)
VDOMAIN(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>)
VDOMAIN(<VERB:são|ser|26|28|V*ip3p_***|hmn|>,<VERB:são|ser|26|28|V*ip3p_***|hmn|>)
MOD_POST(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<NOUN:Japão|Japão|33|37|Np***sm***|hmn|>)
MOD_POST(<VERB:são|ser|26|28|V*ip3p_***|hmn|>,<NOUN:Japão|Japão|33|37|Np***sm***|hmn|>)
SUBJ_PRE(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<NOUN:Rui|Rui|2|4|Np***sm***|hmn|>)
SUBJ_PRE(<VERB:são|ser|26|28|V*ip3p_***|hmn|>,<NOUN:horas|hora|20|24|Nc***pf***|hmn|>)
CDIR_SENTENCIAL_POST(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<VERB:são|ser|26|28|V*ip3p_***|hmn|>)
SUBORD_COMPLETIV(<PRON:que|que|16|18|R******p**|>,<VERB:são|ser|26|28|V*ip3p_***|hmn|>)
EMBED(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<VERB:são|ser|26|28|V*ip3p_***|hmn|>)
INTROD_COMPLETIV(<VERB:perguntou|perguntar|6|14|V*is3s_***|hmn|>,<PRON:que|que|16|18|R******p**|>)
NE_INDIVIDUAL_PEOPLE(<NOUN:Rui|Rui|2|4|Np***sm***|hmn|>)
NE_LOCAL_COUNTRY_ADMIN_AREA(<NOUN:Japão|Japão|33|37|Np***sm***|hmn|>)
NE TEMPO GENERICO(<NP:horas|hora|20|24|Nc***pf***|hmn|>)

MAIN(<VERB:respondeu|responder|63|71|V*is3s_***|hmn|>)
DETD(<NOUN:Japão|Japão|43|47|Np***sm***|hmn|>,<ART:o|o|40|41|Td***sm***|hmn|tokenend|>)
DETD(<NOUN:António|António|75|81|Np***sm***|hmn|>,<ART:o|o|73|73|Td***sm***|hmn|>)
PREPD(<NOUN:Japão|Japão|43|47|Np***sm***|hmn|>,<PREP:Em|em|40|41|S****__**s|hmn|tokenstart|>)
PREDSUBJ(<VERB:são|ser|49|51|V*ip3p_***|hmn|>,<NOUN:10|10|53|54|Nc***__***|hmn|,horas|hora|56|60|Nc***pf***|hmn|>)
VDOMAIN(<VERB:são|ser|49|51|V*ip3p_***|hmn|>,<VERB:são|ser|49|51|V*ip3p_***|hmn|>)
VDOMAIN(<VERB:respondeu|responder|63|71|V*is3s_***|hmn|>,<VERB:respondeu|responder|63|71|V*is3s_***|hmn|>)
SUBJ_POST(<VERB:respondeu|responder|63|71|V*is3s_***|hmn|>,<NOUN:António|António|75|81|Np***sm***|hmn|>)
NE_INDIVIDUAL_PEOPLE(<NOUN:António|António|75|81|Np***sm***|hmn|>)
NE_LOCAL_COUNTRY_ADMIN_AREA(<NOUN:Japão|Japão|43|47|Np***sm***|hmn|>)
NE_TEMPO_HORA(<NOUN:10|10|53|54|Nc***__***|hmn|,horas|hora|56|60|Nc***pf***|hmn|>)
```

Figura 1.3: Dependencies for the sentences *O Rui perguntou que horas são no Japão* and *No Japão são 10 horas, respondeu o António.*

- advanced configurations are possible, due to the open source status of the tool, such as the management of *fairness* and optimization of space usage (with backfilling and node setting);
- creation of a High-Throughput Computing Environment (where large computational power is used in a short timeframe);
- identification of the work status of machines (idle machines can be assigned to new queued processes);
- independence of file systems (computers in a GRID may run different file systems and still run the same specific job).

Finally, its easy integration with the file system here used and with the Hadoop tool makes it an ideal application to the job of scheduling the processing of the corpus. The processing rate of the corpus was of 21 minutes for each 60Mb (200 files of 300kb each). Since the corpus size is approximately 1,2Gb, we can conclude that the time needed to fully process it is about 7 hours.

### 1.4.3 Hadoop

Hadoop (Hadoop Team, 2008) is a software platform used to run applications that process large quantities of information, by implementing the MapReduce programming model (Dean, 2006). It also pos-

sesses a distributed file system named HDFS that was designed to store large files across several machines in clusters, while maintaining a high-throughput data access. In our case, the data processed in the GRID with the aid of the Condor scheduling application will be stored in HDFS and will be accessed by our search queries in a distributed fashion.

Both Hadoop's MapReduce framework and file system follow a master/slave architecture, with a master server (called *Namenode* in the file system and *jobtracker* in the framework) that serves as a liaison to each slave node (*Datanodes* in the file system, *tasktrackers* in the framework). In the framework it queues user requests and assigns the reduce and map tasks to the slaves (that execute the map and reduce tasks), while in the file system regulates file access and determines the block mapping to each slave (that reads and writes the requested jobs).

The MapReduce model Hadoop follows has two phases, the Map and the Reduce. The Map phase converts the input into a number of fragments and distributes them, each of these fragments is called a map task and possesses a key/value pair (K,V). After this phase the maps are grouped by key producing tuples (K',$V'^*$), and the tuples are partitioned into a number of fragments. The Reduce phase consists of assigning the tuple fragments to reduce tasks that assign them an output key/value pair (K,V) and distributes them. This allows Hadoop to divide applications into small blocks that can be allocated to HDFS (along with its backups) and placed on computers in the cluster, and process them as if they where a whole application in a single computer.

Hadoop's main advantages are:

- Scalability: can store and process large quantities of data (on the order of petabytes);
- Economy: distributes data by several computers in a cluster according to availability;
- Efficiency: processes in parallel, augmenting speed;
- Reliability: maintains data backups and redeploys failed computer tasks.

These advantages are crucial to maintain (and access) large data such as the output of the processed CETEMPúblico corpus, as well as to perform the series of data searches we require.

The programs that were devised to run on the corpus are used in this platform, and take on average 30 minutes. This is due to the size of the corpus, since running a program on a smaller part (60Mb, for instance) presents results in less than one minute.

### 1.4.4 GRID

The GRID concept was developed by Ian Foster and Carl Kesselman (Kesselman & Foster, 1998) and it is not a tool *per se*, but rather a method of processing large quantities of information in a limited time

frame, by allowing the concurrent use of a cluster of computers to the specified processes. The concept has been evolving even since, being preferred in favor of the supercomputer, as a method of processing large quantities of data.

One of the GRID's most interesting properties is the CPU scavenging, that allows computers on a network to "borrow" its unused instruction cycles to a given task, basically enhancing the GRID with additional disk space, RAM and CPU power for as long as the computer stays unused (a typical "client" of this property is the SETI project[6]).

Its use optimizes the use of the abovementioned Hadoop tool and its file system, by allowing the division of the tasks (and output files) among the GRID computers (the Condor application mentioned earlier was also important in the scheduling of the data processing).

## 1.5 Roadmap

In section 2 we identify the lexical and syntactical properties present in the matrices, and describe them in some detail. We also describe a system that present useful ideas.

In section 3 we present the strategies used and the steps taken to implement the solutions to the problems identified earlier, describing their functions and presenting examples for each program.

In section 4 we discuss the evaluation process for the programs, and present the more relevant aspects of the obtained results.

Finally, in section 5 we present the conclusions reached and present some ideas for future work.

[6] http://setiathome.ssl.berkeley.edu/

# 2 State of the Art

## 2.1 Introduction

In this thesis, we describe the properties of lexical-syntactical matrices as well as the different items and their respective properties. We also look into some works on NLP like the Semantic space models of Padó and Lapata (Padó & Lapata, 2007), which is an innovative framework used to represent lexical meaning. This framework and its various parameterizations show an alternative method of retrieving lexical information from *corpora*.

In this section, we also look carefully at the different properties of the lexical elements present in the matrices and describe their properties (mainly distributional and transformational).

We characterize distribution in linguistics as[1]:

- A group of contexts and situations in which a unit or morpheme may occur.

So, basically, distributional properties are those that indicate the positions in a sentence in which a word may occur. An example of this would be the presence of a human noun (*Nhum*) in the *N0* (subject) or *N1* (direct object) position. Further discussion of these properties is found on the distributional properties subsection.

Distributional properties are here taken in a as much the same way as selectional/semantic constraints on argument positions. Things are slightly different regarding choice of auxiliaries.

Another set of properties we studied were the transformational, which are described as[2]:

- a rule that systematically converts one syntactic form or form of a sentence into another;
- a construction or sentence derived by such a rule; a transform.

[1] http://www.portaldalinguaportuguesa.org/index.php?action=terminologyact=viewid=1823)
[2] http://www.answers.com/topic/transformation

Transformational properties transform a phrase into another, by reordering, inserting and removing elements, like the form *N0 V N1 ⇒ N1 ser Vpp N0*. This is ilustrated by the passive transformation in the example *O João leu o livro* 'Pedro read a book.', which can be transformed in *O livro foi lido pelo João* 'A book was read by Pedro.'. This and other properties will be shown in the transformational properties subsection.

In the following subsections we describe some of the matrices used, the properties according to their presence on the psychological verbs matrices and the Semantic Space Models and PMI-IR methods.

## 2.2 Introduction to the matrices available in Portuguese

### 2.2.1 Notation

This short section serves to describe some notations used in the property lists of the matrices described in the following sections.

- NP = Noun Phrase
- N0 = Subject NP
- N1 = 1st Object NP
- N2 = 2nd Object NP
- Nnr = Non-Restricted Noun
- Nhum = Human Noun
- N-hum = Non-Human Noun
- V-inf = Verb in the Infinitive form
- V-a = Adjective derived from a verb
- V-n = Noun derived from a verb
- Qu F = Completive sentence
- Vpp = Verb in the Past Participle form
- V-se = Verb with an associated clitic (se)
- Prep = Preposition
- V concret = Verb that has more than one meaning, namely a concrete construction

- Vcop = Copulative verb
- Vaux = Auxiliary verb
- Adj = Adjective

### 2.2.2 Psychological Verbs Matrices:

A psychological verb is a verb that expresses emotional states, be it by an Experiencing Object (the verb *preocupar* 'to worry') like in the example *A situação preocupa-me* 'The situation worries me', or by an Experiencing Subject (the verb *amar* 'to love') like in *O Pedro ama-a* 'Pedro loves her'.

According to Levin (Levin, 1993), there are four different classes of psychological verbs:

- the transitive verbs with the experiencer in the subject (like *adorar* 'to adore');
- the transitive verbs with the experiencer in the object (like *alegrar* 'to cheer');
- the intransitive verbs with the experiencer in the subject (like *admirar-se perante** 'to marvel at');
- the intransitive verbs with the experiencer in the object (like *agradar* 'to please').

Levin's work is based on english verbs, but in translating these classification for portuguese verbs we find that the third class (intransitive with experiencer in the subject) does not seem to exist. An example of such constructions could be the use of the verb *desesperar* 'to despair' in sentences such as:

- *O Pedro desesperou de obter isso.* 'Pedro despaired to obtain it.'

However, this verb also presents the transitive construction:

- *Isso desespera o Pedro.* 'It despairs Pedro.'

The relation (if any) between the two structures is unclear.

Portuguese psychological verbs of the second type where described by (Oliveira, 1984) in matrix form. The various properties presented in these matrices are the following:

- The presence of human nouns in the subject of a *N-V-Nhum* pattern. This pattern has an obligatory human noun on the direct object and the verb has to belong to the list of psychological verbs collected in the matrix;
- The *se*-passive property, that has a human noun following a verb that contains a reflexive pronoun (clitic) attached to it (*me*, *te*, *nos*, *vos* and *se*);

- Another passive construction is also referred, the *N-ser-Vpp* pattern, that consists of a human noun followed by the auxiliary verb *ser* 'to be' and a psychological verb in the past participle form;
- The equivalence to adjectival constructions, which do not change the basic meaning of the verbs and where the adjectives are derived by adding a suffix to them (*-do*, *-dor*, *-nte*, *-eiro*, *ivo* and *-tório*).
- The "V concret" property, which indicates that a psychological verb also has a non-psychological meaning;
- The equivalence to a deverbal noun construction, which may present specified patterns;
- And the Ni=QuF property, that is the presence of subordinate clauses ("*completivas*") on the subject and direct object positions.

## 2.3 Properties

### 2.3.1 Distributional Properties

In this section we look at the various distributional properties represented in the lexical matrix of psychological verbs, and we illustrate them. The *Nhum* (human noun) property indicates that in a given argument position (like N0, subject) a "human" noun can appear, the typical case of a *Nhum* is a proper name such as *Pedro*. For example, we consider that the verb *gostar* 'to like' selects a N0=Nhum:

- *O Pedro gosta de discotecas.* 'Pedro likes nightclubs.'

While the verb *divertir* 'to amuse' determines a human direct object:

- *A notícia divertiu o Rei.* 'The news amused the king.'

The non-human noun (*N-hum*) is not exactly a distributional property, since it represents all other types of nouns that are not classified as *Nhum*.

The Ni = QuF property indicates that in a given argument position (Ni; say N0=subject or N1=first complement) the verb selects a subordinate clause ("*completiva*" in Portuguese terminology). Subordinate clauses can appear on subject or complement positions (in brackets, in the following examples):

- *(Que o Pedro faça isso) irrita o João.* 'That Pedro does it, annoyes João.'
- *O Pedro gosta de (que a Ana lhe faça isso).* 'Pedro likes that Ana does it to him.'

The notation QuF usually indicates finite (subjunctive/indicative) subordinate clauses (the subordinate clauses are in the subjunctive mode in the examples above). The matrix author does not distinguish these two situations, probably due to the fact that most psychological verbs show a QuF in the subjunctive mode. Other types of QuF are explicitly encoded:

- (i) Infinitives "Vinf W":
  - *Aborrece-me (ter de lavar a loiça).* 'Having to wash the dishes bores me.'
  - *Adoro (lavar a loiça).* 'I love washing the dishes.'
- (ii) Factives "O facto de Vinf W":
  - *Agrada-me (o facto de ter sido promovido)* 'The fact of having been promoted pleases me.'

It should be noted that the matrices (Oliveira, 1984) seem to accept factives with the main verb in the subjunctive mode.

### 2.3.2 Transformational Properties

In this section we look at the transformational properties present in the psychological verbs matrix. The first properties examined are those consisting in the equivalence between the psychological verb construction and a deverbal noun (noun derived from a verb) sentence, i.e. a transformational relation usually called a nominalization (Baptista, 2005).

The nominalizations here studied involve a small set of standard structures. The various sentence forms used to identify this property are:

- *N0 dá Det V-n a N1*: Shows the deverbal noun (V-n) preceded by an N0, the verb *dar* 'to give' and a determinant, and followed by the preposition *a* and an N1. An example of this pattern is:
  - *O Pedro dá tranquilidade à Maria=O Pedro tranquiliza a Maria*
    'Pedro gives confort to Maria'='Pedro conforts Maria'
- *N0 causa Det V-n a N1*: Shows the deverbal noun preceded by an N0, the verb *causar* 'to cause' and a determinant, and followed by the preposition *a* and an N1. An example of this pattern is:
  - *O Pedro causou uma ferida à Maria=O Pedro feriu a Maria*
    'Pedro caused an injury to Maria'='Pedro injured Maria'
- *N0 faz Det V-n a N1*: Shows the deverbal noun preceded by an N0, the verb *fazer* 'to do' and a determinant, and followed by the preposition *a* and an N1. An example of this pattern is:

Tabela 2.1: Suffix Presence

| $Suffix$ | $Present$ | $NotPresent$ |
|---|---|---|
| *-do* | *Aborrecer - Aborrecido* | *Insidiar - *Insidiado* |
| *-dor* | *Animar - Animador* | *Abrandar - *Abrandador* |
| *-nte* | *Revoltar - Revoltante* | *Incendiar - *Incendiante* |
| *-eiro* | *Justiçar - Justiceiro* | *Traumatizar - *Traumatizeiro* |
| *-ivo* | *Afligir - Aflitivo* | *Agastar - *Agastativo* |
| *-tório* | *Vexar - Vexatório* | *Trespassar - *Trespassatório* |
| *-oso* | *Viciar - Vicioso* | *Acobardar - *Acobardoso* |

  - *O Pedro faz um ultraje à Maria=O Pedro ultraja a Maria*
    'Pedro makes an outrage at Maria'='Pedro outrages Maria'

- *N1 está em Det V-n perante N0*: Shows the deverbal noun preceded by an N1, the verb *estar* 'to be' and the preposition *em*, and followed by the preposition *perante* and an N0. An example of this pattern is:

  - *A Maria está numa alegria enorme perante esta notícia=A notícia alegra a Maria*
    'Maria is very satisfied about the news'='The news satisfy Maria'

- *N1 fica em Det V-n perante N0*: Shows the deverbal noun preceded by an N1, the verb *ficar* 'to stay' and the preposition *em*, and followed by the preposition *perante* and an N0. An example of this pattern is:

  - *A Maria fica num grande encantamento perante as histórias do Pedro=As histórias do Pedro encantam a Maria*
    'Maria stays in an amazement regarding Pedro's stories'='Pedro's stories amaze Maria'

- *N1 (está + fica) em Det V-n contra N0*: Shows the deverbal noun preceded by an N1, the verb *estar* or *ficar* and the preposition *em*, and followed by the preposition *contra* and an N0. An example of this pattern is:

  - *A Maria fica numa raiva enorme contra o Pedro=O Pedro enraivece a Maria*
    'Maria is in a rage against Pedro'='Pedro enrages Maria'

The V-a properties found on the psychological verbs matrix correspond to the equivalence to adjectival constructions. The various suffixes described the matrix are *-do*, *-dor*, *-nte*, *-eiro*, *-ivo*, *-tório* and *-oso*. Table (2.1) shows example of verbs that can, and cannot use them.

The "V concret" property indicates which verbs also have a non-psychological meaning, like *agitar* 'to shake / to agitate' for instance, that has a psychological meaning in *Esta notícia agita a população* 'The news agitated the crowd', and a non-psychological meaning in *A Joana agita a garrafa* 'Joana shook the bottle'.

The two remaining properties are: N1 *é* Vpp *por* N0, which corresponds to the passive transformation with the auxiliary verb *ser* 'to be'; and N1 V-se por N0, which stands for the *se*-passive transformation; in this later sentence form (or transform (Harris, 1964), or alternation (Levin, 1993)), the verb is kept and a reflex pronoun (clitic) is added to the passive constructions, the object NP becomes the subject, while the subject NP appears as a prepositional phrase PP; usually this PP is omitted, especially in the se-passive construction.

Examples of these properties are shown on the following sentences:

- *A Maria foi embaraçada pelas declarações do Pedro.=As declarações do Pedro embaraçam a Maria.*
  'Maria was embarrassed by Pedro's statements'='Pedro's statements embarrass Maria.'
- *A Maria angustia-se com o estado de saúde do Pedro.=O estado de saúde do Pedro angustia a Maria.*
  'Maria anguishes over Pedro's health.'='Pedro's health anguishes Maria.'

## 2.4 Semantic Space Models

Several frameworks exist on NLP to represent lexical meanings; the Dependency-Based Construction of Padó and Lapata (Padó & Lapata, 2007), focus on vector space models. These models use word co-occurrence counts to recover lexical information from large corpora, by adding each semantical property to a vector.

This framework is based on existing semantic word space models of two different types, word-based co-occurrences and syntax-based models. The word-based models, like Lowe's (Lowe, 2001), do not consider any syntactical relationship between words, simply counting the number of occurrences of certain words in the vicinity[3] of the target word; the cosine similarity formula (Salton, 1989) is used for measuring the distances (2.1). Table 2.2 shows the model at work on the sentence *O Jorge pode comprar muitos jornais* 'Jorge can buy many newspapers' for the word space (*Jorge*, *comprar*, *muitos*, *livros*), where each word in the space will be recorded if it is presented in the vicinity of each target word.

$$sim_{cos}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2.1}$$

Syntax-based models are significantly different from the co-occurrence model, since they capture syntactical relations between words. The basic elements of these models, like Grefenstette's (Grefenstette, 1994), that differentiates itself from the word co-occurrence method by using tuples (r,w), were w

[3] 2 words to the left or to the right.

Tabela 2.2: Word-based semantic space

| | $O$ | $Jorge$ | $pode$ | $comprar$ | $muitos$ | $jornais$ |
|---|---|---|---|---|---|---|
| *Jorge* | 0 | 0 | 1 | 1 | 0 | 0 |
| *comprar* | 0 | 1 | 1 | 0 | 1 | 1 |
| *muitos* | 0 | 0 | 1 | 1 | 0 | 1 |
| *livros* | 0 | 0 | 0 | 0 | 0 | 0 |

Tabela 2.3: Grefenstette's semantic space

| | $(subj, Jorge)$ | $(aux, pode)$ | $(mod, muitos)$ | $(obj, jornais)$ |
|---|---|---|---|---|
| *Jorge* | 0 | 0 | 0 | 0 |
| *comprar* | 1 | 1 | 0 | 1 |
| *muitos* | 0 | 0 | 0 | 0 |
| *livros* | 0 | 0 | 0 | 0 |

is a word in a type r relation with a target word (2.2), the measure of association here used was based on Jaccard's coefficient (Jaccard, 1901). An example of this model is shown in table 2.3 were we can see, for instance, that Jorge has a subject relation to *comprar* in the sample phrase *O Jorge pode comprar muitos jornais*.

$$sim_{jacc}(t_1, t_2) = \frac{Attr(t_1) \cap Attr(t_2)}{Attr(t_1) \cup Attr(t_2)} \tag{2.2}$$

Semantical Space Model Framework moved away from words as basic units and focus more on syntactical information, yet maintaining the simplicity that allows its use in multiple languages. The algorithm used to create the semantic space model uses paths that are no more than sequences of dependencies extracted from the parsing of a sentence, they can be length 1 paths[4] or others, such as the path between *Jorge* and *muitos* ($\langle Jorge, comprar, jornais, muitos \rangle$). We can see a lexical tree showing this on figure 2.1.

Three important functions were also considered, related with the paths in the parsing tree:

- the Context Selection Function, that determine the paths in the graph that will be used to represent a target word (like only using length 1 paths or paths with a length above 2, for instance);
- the Path Value Function, that assigns weights to paths (allowing differentiation of words by their syntactical roles, like giving a bigger weight to subjects than determiners);
- and the Basis Mapping Function that creates the dimensions of the semantic space (by defining the dimension independently of path construction and by collapsing equivalent paths).

The first step in the construction of semantic spaces following this framework is the building of the context; contexts are defined as anchored paths (paths in dependency graphs that start at a target

[4]Individual words.

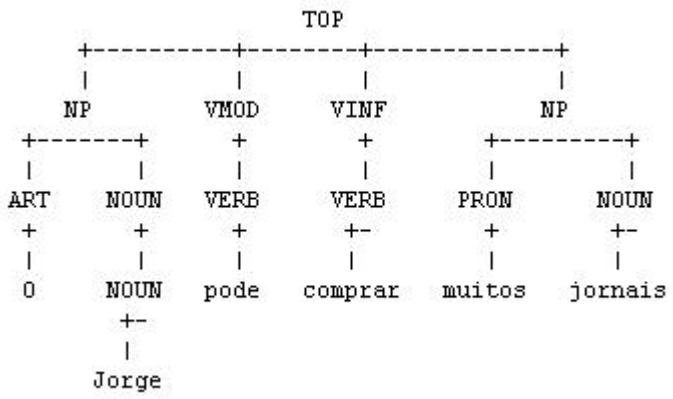



Figura 2.1: Syntactical tree for the sample paragraph

word t), which are joined in sets for each anchored word and from were the Context Selection Function chooses one to use. This choice will determine the data that will be collected in the semantic space (this can be dependency paths of length 1 (Grefenstette, 1994) or only paths that possess a [V, subj, N] structure).

The second step consists of mapping all the paths onto basic elements, allowing different paths with a similar structure to be mapped to the same basic elements (labeling them). Two different models exist to do this: Lowe's word-based (Lowe, 2001), which basically considers context words as basic elements (maps all paths that end on a word w onto the basic element w), and Grefenstette's (Grefenstette, 1994) that only consider length 1 paths.

The third and last step of the semantic model definition is the specification of the relative value of each different path. This can be done in two ways, either by letting the path value function v assign a number to it, or by the traditional method of giving all paths an equal weight (normally 1). This allows a new level of differentiation since paths mapped onto the same basic element can have different weights.

Following the three step plan, a flexible framework is achieved, which can create semantic spaces over words, POS and syntactical relations. These semantic spaces are sparser than the word-based model because their context selection function is more selective. Table 2.4 is an example of this space model at work on the phrase *O Jorge pode comprar muitos jornais*.

This framework was tested on three experiments (Single-Word Priming, Detection of Synonymy and Sense Ranking), and to this end a setup was needed to parse the test corpus (British National Corpus, consisting of 100 million words) and to select the parameters for the tests, like dimensions of the test data, size and type of context , etc. This is not normally done in the test data, but instead on a development data (in this case the benchmarked set collect by Rubenstein and Goodenough (Rubenstein

Tabela 2.4: Dependency-based semantic space

| | $O$ | $Jorge$ | $pode$ | $comprar$ | $muitos$ | $jornais$ |
|---|---|---|---|---|---|---|
| *Jorge* | 0 | 0 | 0 | 1 | 0 | 0.5 |
| *pode* | 0 | 0 | 0 | 0 | 0 | 0 |
| *comprar* | 0 | 1 | 0 | 0 | 0 | 1 |
| *muitos* | 0 | 0 | 0 | 0 | 0 | 0 |
| *livros* | 0 | 0 | 0 | 0 | 0 | 0 |

& Goodenough, 1965)) to prevent over fitting. The most important parameters found in this way were the following three context selections and three path value functions:

Context Selections:

- minimum contexts that contain length 1 paths and 27 templates[5];
- medium contexts that contain length $\leq 3$ paths and 59 templates;
- maximum contexts that contain all templates (123) and lengths $\leq 4$ paths.

Path Value Functions:

- Plain, that assigns the same value "1" to all paths;
- Length, that assigns a value based on a length-based weighting scheme (2.3) (with $\pi$ being the distance between words);
- Gram-rel, that uses hierarchy to rank paths according to grammatical relations (2.4) (with subj being the subject, obj the object, obl an oblique[6] and gen a genitive).

$$v_{length}(\pi) = \frac{1}{\|\pi\|} \tag{2.3}$$

$$v_{gram-rel}(\pi) = \begin{cases} 5 & \text{if subj } \in l(\pi) \\ 4 & \text{if obj } \in l(\pi) \\ 3 & \text{if obl } \in l(\pi) \\ 2 & \text{if gen } \in l(\pi) \\ 1 & \text{else} \end{cases} \tag{2.4}$$

[5]Considers only direct relations.
[6]Prepositional phrase.

Tabela 2.5: Performance Table

| $Context/Path$ | $plain$ | $length$ | $gram-rel$ |
|---|---|---|---|
| minimum | 0.58 | 0.58 | 0.58 |
| medium | 0.60 | 0.62 | 0.59 |
| maximum | 0.56 | 0.59 | 0.55 |

Tabela 2.6: Mean distance values for Related and Unrelated prime-target pairs and Prime Effect size for the dependency and the ICE models

| $LexicalRelation$ | $N$ | $Related$ | $Unrelated$ | $Effect(dependency)$ | $Effect(ICE)$ |
|---|---|---|---|---|---|
| Synonymy | 23 | 0.267 | 0.102 | 0.165 | 0.063 |
| Superordination | 21 | 0.227 | 0.121 | 0.106 | 0.067 |
| Category coordination | 23 | 0.256 | 0.119 | 0.137 | 0.074 |
| Antonymy | 24 | 0.292 | 0.127 | 0.165 | 0.097 |
| Conceptual association | 23 | 0.204 | 0.121 | 0.083 | 0.086 |
| Phrasal association | 22 | 0.146 | 0.103 | 0.043 | 0.058 |

This lead to nine model instantiations, from which the optimal model was chosen to be used in the experiments (in this case the one with best results used the medium content selection and length path value functions, as seen in table 2.5).

The first experiment consisted of Single-Word Priming (which is the semantic similarity or dissimilarity between words) namely the fact that the presentation of a prime word will facilitate the pronunciation or lexical decision of a target word, according to Hodgson (Hodgson, 1991). This priming effect has been modeled by McDonald and Brew (McDonald & Brew, 2004) using a vector-based model that simulates the difference in effort by processing a target word preceded by a prime word, and by processing the same word preceded by an unrelated prime word.

Their experiment follows the methodology of McDonald and Brew, testing two hypotheses, one that the dependency-based model can simulate semantic priming, and the other that the model will be better at priming than a traditional word-based model. An ICE (Incremental Construction of Semantic Expectations) model was used to simulate the effort differences between processing a target word preceded by an unrelated or by a related prime word.

Their results indicate that the two hypothesis described above were proven true (semantic priming can be simulated with a dependency-based model and the results are better than those obtained from a word-based-model. This can be seen in table 2.6, where Related and Unrelated are the mean distance values when a previously processed word is related or unrelated, N is the number of times the relation appears on the corpus, and the dependency and ICE effects are the calculation of the Prime Effect size.

The second experiment was the detection of synonymy, i.e. identifying a synonym of a given target word out of a list of possibilities. Their methodology uses the TOEFL (Test of English as a Foreign Language) as benchmark that is a set of 80 multiple-choice questions, involving a target word in a sentence and a list of four potential synonyms. Several methods were tested alongside their Dependency space,

Tabela 2.7: Comparison of different models on the TOEFL synonymy task

| $Model$ | $Corpus$ | $Accuracy(\%)$ |
|---|---|---|
| Random Baseline | - | 25.0 |
| Word-based space | BNC | 61.3 |
| Dependency space | BNC | 73.0 |
| PMI-IR | BNC | 61.3 |
| PMI-IR | Web | 72.5 |
| LC-IR | Web | 81.3 |

such as the word-based space, the PMI-IR[7] of Turney (Turney, 2001), the LC-IR[8] of Higgins (Higgins, 2004) and random guessing[9].

Their results are shown in table 2.7, were we conclude that the semantic space model is better than the word-based space model and the PMI-IR method and a lot better than random guessing, but nonetheless achieved poorer results than the LC-IR method. It is to be noted that the LC-IR and PMI-IR methods can be applied to a web-based corpus, which is potentially much larger than the BNC.

Finally the last experiment was the Sense Ranking Model, which is the way senses (senses are no more than the set of words most similar to a target word) are scored according to their similarity to it's neighbors. The scoring is made using the "predominant sense score" (formulas 2.5 and 2.6) of McCarthy et al. (McCarthy et al., 2004):

$$PS(ws_i) = \sum_{n_j \in N(w)} sim_{distr}(w, n_j) \times \frac{sim_{sem}(ws_i, n_j)}{\sum_{ws'_i \in S(w)} sim_{sem}(ws'_i, n_j)} \tag{2.5}$$

$$sim_{sem}(ws_i, n_j) = \max_{ws_x \in S(n_j)} sim_{WN}(ws_i, ws_x) \tag{2.6}$$

The model possesses four different parameters:

- semantic space where the words are acquired;
- (sim(distr)): measure of distributional similarity;
- (k): Number of neighbors;
- (sim(WN)): measure of sense similarity.

The methodology followed was based on McCarthy's study using their optimal dependency-based model and the baseline word-based model. The results obtained are shown in table 2.8 (the $Acc_sr$ and

[7]Point wise mutual information retrieval.
[8]Local-context information retrieval.
[9]Assigns a random sense to each token.

Tabela 2.8: Sense Ranking and WSD tasks distance measure

| $Models$ | $Acc_sr$ | $Acc_wsd$ |
|---|---|---|
| Random Baseline | 31.0 | 25.4 |
| Word-based space | 49.3 | 49.9 |
| Dependency space | 54.3 | 54.3 |
| McCarthy et al. | 54.0 | 46.0 |
| Upper bound | - | 67.0 |

$Acc_wsd$ symbols represent the accuracy of the sense ranking and of the WSD process), and find the Dependency space model beating McCarthy's model, the word-based model and a random baseline.

There is no direct link between this work and my efforts to validate the matrices. This work serves mainly to demonstrate one of the methods used to represent lexical meaning in corpora.

## 2.5 Summary

In this section we present the matrices that we are going to validate (psychological verbs matrices), describing the properties shown on them. This description was accompanied by their division into two main categories: Distributional properties and transformational properties.

The distributional properties that we found, were the presence of human nouns (Nhum) and of subordinate clauses (*completivas*) in argument positions.

The transformational properties found were the existence of deverbal noun (nouns derived from a verb), adjectival constructions (adjectives derived from a verb), passive transformations (either with the auxiliary verb *ser* or *se*-passive) and the V "concret" (marking verbs that possess a non-psychological meaning).

An alternate method of retrieving lexical information was also explained (Semantic Space Models).

# 3 Strategy and Implementation

In this section we present the steps taken to solve each problem, dividing solutions in two parts: Strategy and Implementation.

The Strategy subsections show an overview of the problem while the Implementation subsections present, in loose terms, the solution process, as well as an example to show step-by-step the process.

Also, due to the fact that the process of identification and the underlying strategy behind the search for the passive patterns (*N-ser-Vpp* and *N-V-se*) is similar, both of them will be treated in a single section.

## 3.1 Corpus Processing

Before focusing on developing programs to search for specific information we must process the corpus data (CETEMPúblico) using the $L^2F$ chain in order to obtain the corpus in an xml tree form. This form contains all the relevant aspects in an easy to search format.

The first step taken towards a solution is the decomposition of the problem into smaller, more manageable ones, simplifying its complexity and allowing its sequential resolution in a step-by-step manner. The first problem detected is the processing of the corpus in a manageable time (less than 24 hours). This has been solved by dividing the corpus into several files processed concurrently in the GRID using the Condor AFS file system.

The process itself begins with the setting up of a condor environment, in order to process the information in the GRID effectively. Next we decided to process several files with a large size (6000 sentences long $\approx$ 800kb) using the GRID; given that we encountered memory issues, the size of the files was shortened in each test until we found the largest size that caused no problem: 2000 sentences $\approx$ 280kb. The processing of 200 of these files in the grid (10 parallel processings) allowed us to process 60Mb of information in under 21 minutes. Simple arithmetics allow us to conclude that given the corpus size, its processing will be concluded well within the 24 hours limit that has been set as this task goal (this limit was set because the processing chain is in constant upgrade, so the reprocessing of the corpus is a task that may have to be done several times, and therefore must not consume too much time).

The division and condor/hadoop running commands where done through a series of scripts developed specifically for this effect, in order to automatize the process for subsequent runs. These scripts

work by running a set of commands that process each file and place the results in a designated folder in the hadoop file system.

After being run, and placed in the hadoop file system, the xml results are accessed through the hadoop structure that allows programs to be run in a jar format.

Some of the problems we found where:

- Discovery of the correct size for each file,
- Encoding problems,
- Automatization of the process (script creation).

Each of these problems happened in a specific time and had a particular solution. The size discovery was a trial and error endeavor, with several sizes being tried (Starting at 6000 sentences / 800kb) until we found the 2000 sentences / 280kb value. The encoding problem occurred due to having the files in the ISO format that did not permit the correct display of certain characters like accents. This was resolved with a script that applied a conversion command (iconv) to all input files, transforming them into the UTF8 format.

The final problem was the automatization of the whole process, and that was solved by creating a series of scripts to run the commands.

- condor-submit.sh (3.1): That submits the results to the hadoop file system,
- run-xip.sh (3.2): Runs the xip-runner jar in an input file,
- run-xip.condor (3.3): Sets the parameters for running condor (like what machines are used),
- xip-runner.jar: A jar application that runs the $L^2F$ chain.

Basically, by running condor-submit.sh we set the process in motion since it calls run-xip.sh for each input file and run-xip.condor to specify the condor parameters.

## 3.2 Verbal Chains

### 3.2.1 Strategy

Verbal chains are strings of verb forms, eventually discontinuous, consisting in a sequence of auxiliary verbs and ending in a main verb. The main verb is a non-inflected verb form (infinitive, gerund, and past participle). Auxiliary verbs can operate on other auxiliaries, in a recursive (but limited) way. Each

```
#!/bin/bash

# get a short-term X.509 certificate
kx509
kxlist -p

# Define dir with source files
base_dir="/afs/l2f/projects/xip/Corpus/CETEMPublico"
script_dir="/afs/l2f/projects/xip/Corpus/scripts"

# Process each source file
for dir in ‘ls $base_dir‘; do
     for file in ‘ls $base_dir/$dir‘; do
     condor_submit -a "EXECUTABLE = $script_dir/run-xip.sh" -a "Arguments = $dir $base_dir/$dir/$file" -a
     "X509USERPROXY = /tmp/x509up_u‘id -u‘" $script_dir/run-xip.condor
 done
done
```

Figura 3.1: condor-submit.sh

```
#!/bin/bash

# $1 - target dir
# $2 - absolute path of the input file

# get the AFS token
/usr/bin/gssklog -server srv5.l2f.inesc-id.pt

data="20090412"
script_dir="/afs/l2f/projects/xip/Corpus/scripts"
hadoop_dir="/corpora/publico/$data"

export JAVA_HOME=/usr/lib64/jvm/java

/opt/hadoop/bin/hadoop jar $script_dir/xiprunner.jar UTF-8 $2 $hadoop_dir/$1/‘basename $2‘.out

# delete token
unlog
```

Figura 3.2: run-xip.sh

```
UNIVERSE = Vanilla
NOTIFICATION = Error
OUTPUT = run-xip.out.$(CLUSTER)-$(PROCESS)
ERROR = run-xip.err.$(CLUSTER)-$(PROCESS)
LOG = run-xip.log
REQUIREMENTS = (Arch == "X86_64") && ( (Machine == "wc01.l2f.inesc-id.pt") || (Machine == "wc02.l2f.inesc-id.pt")
|| (Machine == "wc04.l2f.inesc-id.pt") || (Machine == "wc05.l2f.inesc-id.pt") || (Machine == "wc06.l2f.inesc-id.pt")
|| (Machine == "wc08.l2f.inesc-id.pt") || (Machine == "no00.l2f.inesc-id.pt") || (Machine == "no01.l2f.inesc-id.pt")
|| (Machine == "no02.l2f.inesc-id.pt") || (Machine == "no03.l2f.inesc-id.pt") || (Machine == "no04.l2f.inesc-id.pt")
|| (Machine == "no05.l2f.inesc-id.pt") || (Machine == "no06.l2f.inesc-id.pt") || (Machine == "no07.l2f.inesc-id.pt") )
SHOULD_TRANSFER_FILES   = yes
WHEN_TO_TRANSFER_OUTPUT = on_exit
QUEUE
```

Figura 3.3: run-xip.condor

auxiliary may be directly connected to the following auxiliary, or to the main verb, or by way of a preposition. The non-inflected form and the choice of the preposition (if any) depends on the auxiliary construction (its syntactic structure). Auxiliary verbs are classified according to the main grammatical values they convey, that is, the way they contribute to the main verb general (lexical) meaning. This list of auxiliary verb constructions and their classification has been established by (Mamede & Baptista, 2009). The following sentence shows a string of auxiliary verbs:

*O Pedro pode ter começado a ler o livro ontem*

'Peter may have begun reading the book yesterday'

This example is analysed as :

*O Pedro pode*/VMOD *ter*/VTEMP *começado a*/VASP *ler*/VINF *o livro ontem*

Because of their adjectival nature, inflected forms of past participles (VCPART) are assimilated to adjectives in front of copula verbs (VCOP) *ser* or *estar* 'to be'; these are not to be confound with the non-inflected form of the past participle used with temporal auxiliary verbs *ter* 'to have' (and more rarely, **haver* *'there be'):

*O livro ainda não estava esgotado.*

*O livro ainda não tinha esgotado.*

'The book had not yet sold out'

These examples are analysed as :

*O livro ainda não estava*/VCOP *esgotado*/VCPART.

*O livro ainda não tinha*/VTEMP *esgotado*/VCPART.

However, only the first VCPART agrees with the sentence's subject, while the second remains non-inflected:

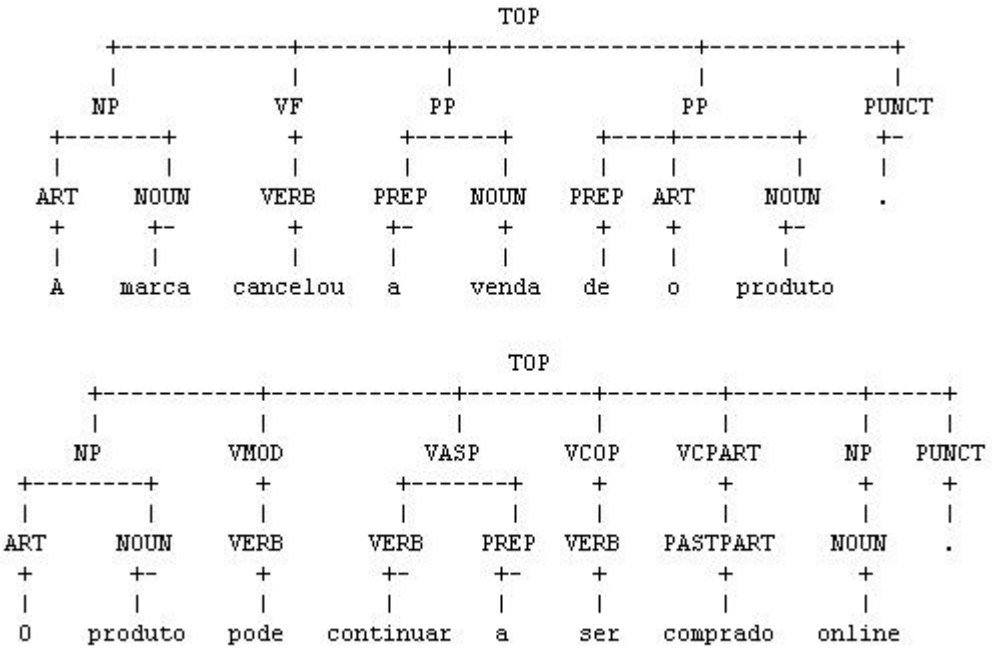
*A revista ainda não estava esgotada.*

*A revista ainda não tinha esgotado/*esgotada.*

'The magazine had not yet sold out'

Because verbal chains with auxiliary verbs constitute instances of the main verb equivalent to the instances where it appears as a single word (that is, without auxiliaries), it is necessary to identify these verbal chains before trying to validate the syntactic environment of all instances of a target verb. The processing chain retrieves the relevant information from two dependencies, VLINK and VDOMAIN. VDOMAIN shows the first and last verbs in a domain, and is used in the identification of one verb

chains (when only one verb appears in the dependency). VLINK is used to identify chains of more than one verb, in a simple case of a two verb chain, a single VLINK is needed, but on larger chains it is necessary to use the information of more VLINKs. This can be seen on the example in figure 3.4.




```
(...)
VDOMAIN(<VERB:cancelou|cancelar|8|15|V*is3s_***|hmm|>,<VERB:cancelou|cancelar|8|15|V*is3s_***|hmm|>)
(...)

(...)
VLINK(<VERB:pode|poder|47|50|V*ip3s_***|hmm|>,<VERB:continuar|continuar|52|60|V*n____***|hmm|>)
VLINK(<VERB:continuar|continuar|52|60|V*n____***|hmm|>,<VERB:ser|ser|64|66|V*n____***|hmm|>)
VLINK(<VERB:ser|ser|64|66|V*n____***|hmm|>,<PASTPART:comprado|comprar|68|75|V*p__sm***|hmm|>)
VDOMAIN(<VERB:pode|poder|47|50|V*ip3s_***|hmm|>,<PASTPART:comprado|comprar|68|75|V*p__sm***|hmm|>)
(...)
```

Figura 3.4: VLINK/VDOMAIN example

A program was also developed to find the sentences in which certain verbal chains are found: it receives as input a file with the chains and prints in the output the sentences in which they occur. With it, we can find the sentences in which a particular pattern occurs, allowing us to retrieve further information from it.

### 3.2.2 Implementation

#### 3.2.2.1 Verbal Chains

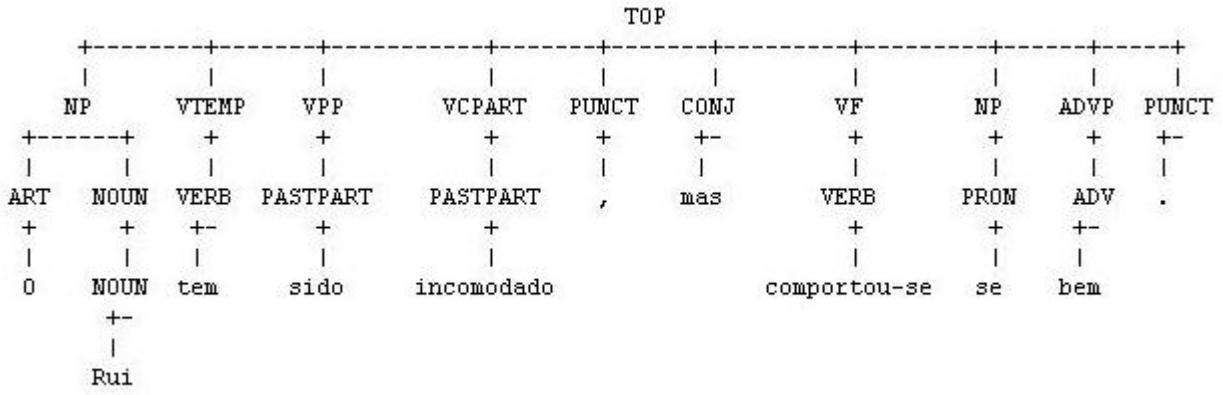
As it was said above, the verbal chain identifier is important to the *Nhum* validation process and functions by retrieving all the chains found in the form verb-type(verb lemma) (for instance: VASP (começar a) VCOP (ser) VCPART (ler)).

The verbal chain program functions in the following manner:

- Search for all NODE children of the tree for tags showing the verb types (VASP, VMOD, VTEMP, VPP, VCPART, VF, VINF, VCOP and VGER) and placing of the tag in a variable;
- The position of the words present in that tag is also taken, and all of these words that are not numbers, adverbs and punctuation marks (if the first word is a preposition it is omitted also) are concatenated into the variable (ex. VASP (começar a));
- The next step consists in placing the variable in a hashmap as a value, with the key being the position the variable has in the text (ex. VCOP (position of *ser* in the text));
- After completing these steps, we then search the dependencies for VLINK; each VLINK connects two verbs, and in some cases the final value of a VLINK is the beginning of another. This creates verbal chains with 2, 3 or more verbs; once the verbal chain is complete, i.e., when all verbs forms connected by the VLINK dependencies have been determined, the information associated to these verbs, and present in the hashmap, is then retrieved and subsequently concatenated and sent to the reducer;
- The case of simple verbs, i.e., verbs without auxiliaries, is checked by finding VDOMAIN that possess two identical child nodes, indicating that the domain is only populated by one verb, the process of retrieving the information from the hashmap is identical to the one described in the VLINK section;
- Finally the results are sent to the REDUCER and the process is repeated for the following trees in the input.

The main difficulties found in this program consisted in the retrieval of all words in each verb type and the identification of connected VLINK.

In the test sentence *O Rui tem sido incomodado, mas comportou-se bem* (as seen on figure 3.5) we can see that there are 4 verbs present *ter*, *ser*, *incomodar* and *comportar*, and that the first 2 are part of a chain. The program will then do the following:




```
MAIN(<PASTPART:sido|ser|10|13|V*p____***|hmm|>)
HEAD(<NOUN:Rui|Rui|2|4|Np***sm***|hmm|>,<NP:O|o|0|0|Td***sm***|hmm|,Rui|Rui|2|4|Np***sm***|hmm|>)
HEAD(<PRON:se|se|31|42|Pf**3__*a_|hmm|cli|>,<NP:se|se|31|42|Pf**3__*a_|hmm|cli|>)
HEAD(<VERB:tem|ter|6|8|V*ip3s_***|hmm|>,<VTEMP:tem|ter|6|8|V*ip3s_***|hmm|>)
HEAD(<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>,<VF:comportou-se|comportar|31|42|V*is3s_***|hmm|>)
HEAD(<PASTPART:sido|ser|10|13|V*p____***|hmm|>,<VPP:sido|ser|10|13|V*p____***|hmm|>)
HEAD(<PASTPART:incomodado|incomodar|15|24|V*p__sm***|hmm|>,<VCPART:incomodado|incomodar|15|24|V*p__sm***|hmm|>)
HEAD(<ADV:bem|bem|44|46|R******p**|hmm|>,<ADVP:bem|bem|44|46|R******p**|hmm|>)
HEAD(<CONJ:mas|mas|27|29|Cc********|hmm|>,<CONJ:mas|mas|27|29|Cc********|hmm|>)
DETD(<NOUN:Rui|Rui|2|4|Np***sm***|hmm|>,<ART:O|o|0|0|Td***sm***|hmm|>)
VLINK(<VERB:tem|ter|6|8|V*ip3s_***|hmm|>,<PASTPART:sido|ser|10|13|V*p____***|hmm|>)
VDOMAIN(<PASTPART:incomodado|incomodar|15|24|V*p__sm***|hmm|>,<PASTPART:incomodado|incomodar|15|24|V*p__sm***|hmm|>)
VDOMAIN(<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>,<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>)
VDOMAIN(<VERB:tem|ter|6|8|V*ip3s_***|hmm|>,<PASTPART:sido|ser|10|13|V*p____***|hmm|>)
MOD_POST(<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>,<ADV:bem|bem|44|46|R******p**|hmm|>)
SUBJ_PRE(<PASTPART:sido|ser|10|13|V*p____***|hmm|>,<NOUN:Rui|Rui|2|4|Np***sm***|hmm|>)
CDIR_POST(<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>,<PRON:se|se|31|42|Pf**3__*a_|hmm|cli|>)
CLITIC_POST(<VERB:comportou-se|comportar|31|42|V*is3s_***|hmm|>,<PRON:se|se|31|42|Pf**3__*a_|hmm|cli|>)
NE_INDIVIDUAL_PEOPLE(<NOUN:Rui|Rui|2|4|Np***sm***|hmm|>)
```

Figura 3.5: Syntactical tree and dependencies

- Places VTEMP in a string along with the verb and auxiliar words present (in this case VTEMP(ter)),
- Does the same with VPP(ser), VCPART(incomodar) and VF(comportar),
- Places each string as a value in a hashmap with it's position as key (key=25 / value=VTEMP(ter)),
- Then we find VLINKS and through their positions we find the values in the hashmap and concatenate them (a VLINK has the child nodes with the positions of VTEMP(ter) and VPP(ser) and so these values where fetched from the hashmap and concatenated in a string),
- Since no new VLINKS occur, the string with the chain is sent to the REDUCER,
- Then a search is made for VDOMAINS that have identical child nodes, since these indicate domains of only one verb (VCPART(incomodar) and VF(comportar) are examples of this), and send the respective strings to the REDUCER,
- When the mapping is finished the REDUCER will add all identical results sent to it and print them.

The output of running the program with the sentence as input would be:

- VTEMP (ter) VPP(ser) 1
- VCPART (incomodar) 1
- VF (comportar) 1

#### 3.2.2.2 Verbal Chain Pattern Recognition

Another program was developed in order to identify patterns given by the verbal chain results. Given a pattern it would show sentences in the corpus that possess that particular pattern.

The steps taken to develop this program where:

- First a list of patterns is fed from a file into a hashmap;
- Then all words from a sentence are concatenated into a string;
- The following steps are similar to 5 and 6 from the verbal chain program, with the difference being that only results that are present in the list of pattern hash are sent to the REDUCER, and the information sent is not only the pattern, but the string with the sentence as well;
- The process is then repeated for the following xml trees.

In this case, difficulties were felt in the use of a file with the patterns located in the hadoop file system. After all difficulties were addressed, it served as template for the *Nhum* searches that also have input files there.

For the same example *O Rui tem sido incomodado, mas comportou-se bem* 'Rui has been bothered, but he behaved well', we will search for the pattern VTEMP(ter) VPP(ser) in a file with 4 other sentences that do not present this pattern (*O João foi enganado* 'João has been fooled', *O Rui tem um carro novo* 'Rui has a new car', *A casa foi feita em tempo record* 'The house was completed in record time' and *A proposta acabou por nem chegar a ser votada* 'The proposition ended up not being voted'). The program will then do the following:

- The pattern VTEMP(ter) VPP(ser) is placed in a file called padroes.txt that will be placed in the hadoop file system,
- The program runs the file and places each pattern in a hashmap (e.g., key=value=VTEMP(ter) VPP(ser)),
- Each sentence is placed in a string, in case the pattern is found,
- The same steps are taken as seen in the verb chain identifier, and when each pattern is found it is compared to those on the hashmap, should they match, a string is created with the pattern and the sentence and sent to the REDUCER,
- The REDUCER will them add all identical results and print them.

The output of running the program with the sentence as input would be:

- PADRAO: VTEMP (ter) VPP (ser)
- FRASE: O Rui tem sido incomodado, mas comportou-se bem

## 3.3 Human Noun Identification

To discover if properties such as the selection of human nouns (*Nhum*) for the argument position of a given list of verbs in a corpus, an exact definition of *Nhum* must be used. The theoretical definition has been explained in chapter 2 but, from a programmer's point of view, it is necessary to determine how a human noun is identified in the input data.

This section shows the process of identifying human nouns on the xml trees resulting from running the CETEMPúblico Corpus on the processing chain, by identifying the various classes of words that are considered human nouns, and more specifically by identifying the xml tags that represent them.

The categories of nouns considered to be *Nhum* where personal pronouns, named entities (*NE*) and other classes of nouns like professions, proper names, affiliation status, organizations, nationalities, titles or family ties.

The attributes obtained from the processing chain (PROFESSION, PEOPLE, MEMBER, AFFILIATION, ORG, NATIONALITY, RELATIVE, HUMAN and TITLE) where used to identify the Nhum's. Personal pronouns where also classified as Nhum's (such as *eu*, *tu*, *ele*, and so on) and are identified by the PERS attribute. Although third person pronouns do not always correspond to the reduction of Nhum, they where coarsely generalized as such to simplify the identification process.

Named entities where also considered as valid candidates for human nouns. The NE type depends not only on the string of words it contains but also on its syntactic context. For instance, *Instituto Superior Técnico* is a named entity, since it is an organization (ORG) in the context of the sentence *Sou estudante do Instituto Superior Técnico* 'I am a student at Instituto Superior Técnico', and a location in the context of the sentence *A reunião decorreu no Instituto Superior Técnico* 'The meeting took place in Instituto Superior Técnico'. These attributes are represented in the processing chain by several tags:

The mentioned entities that interest us are those related to organizations (ENG1, ENG2, ENG3 and ENG4) and persons (ENP1, ENP2, ENP3, ENP4, ENP5 and ENP6), since only these represent entities that are considered Nhum.

Personal pronouns where also treated as Nhum; this is an approximation to the phenomena of Nhum pronouning since it is not always the case that a personal pronoun replaces (or reduces) a Nhum as it happens in *O Rui gosta de sopa* 'Rui likes soup' = *Ele gosta de sopa* 'He likes soup'. In this respect, first and second person pronouns, as they refer to the participants in the discourse, can unambiguously be associated with a Nhum distributional constraint, whereas third person pronouns may refer to non-human entities.

Other classes of nouns, besides named entities have been considered as human nouns. This is the case of nouns designating the professionals (e.g. carpenter '*carpinteiro*') since they can occupy the syntactic slot otherwise filled by a proper name: *O carpinteiro gosta da sopa* 'The carpenter likes soup'.

It should be noted that the HUMAN attribute was added to the processing chain after the first set of results was obtained. This happened because some words that where human nouns (like *criança* 'child') did not fit in any of the other attributes, so a new tag had to be introduced.

## 3.4 Identification of the *Nhum-V-Nhum* pattern

### 3.4.1 Strategy

This study has two main objectives: first, the validation of the data presented in the lexical matrices; secondly, the identification of nouns as human nouns when that information is not yet available for that particular lexical items. The results of this process can then be used to feed the processing chain with those new words, which had not been previously classified as human nouns. This has a feedback effect and will improve the results of the following iterations (an example of this is the HUMAN tag that was added to the processing chain to identify human nouns that where identified in the first iteration of the process).

It should be noted that the adequate identification of *Nhum* candidates is only carried out for relevant syntactic patterns, in particular, those that have been studied in the lexical matrix of psychological verbs. The basic structure of these verbs is *N0-V-N1*, where N1 is a obligatorily a human noun. Therefore, sentences with a psychological verb and an explicit direct object are relevant in the determination of human nouns.

Psychological verbs also allow for an equivalent sentence form *N1-V-Reflex*, also called *se* passive transformations. This property is indicated in the matrix and is dealt with in a later section. For psychological verbs allowing this sentence form, the identification of candidate human nouns applies to the subject position, which is then obligatorily filled by a *Nhum*.

### 3.4.2 Implementation

This program identifies sentences that have a psychological verb as a main verb, and retrieves the subject from it, classifying it as a *Nhum* or non-*Nhum* according to the presence or absence of the tags referred in the strategy section. Then only if the verb also possesses a direct object and this later is classifiable as a *Nhum*, can we retrieve the result.

The program itself works in the following way:

- First the list of psychological verbs is fed from a file into a hashmap;
- Then the program runs on the xml trees and retrieves the VMAIN tags, verifying if the main verb in each sentence is a psychological verb;
- For the sentences with psychological verbs, the SUBJ tag is taken and the lemma of its head is extracted;

- Should the word be a *Nhum* then the word is added with *Nhum+*, and *Nhum-* otherwise;
- Then only if a direct object exists and is considered a human noun can we send the information to the REDUCER;
- The respective counters (*Nhum+*/*Nhum-*) are incremented for the given verb.

After running the main program in the map-reduce paradigm, the results are fed into a program that gives the percentages of *Nhum*'s for each psychological verb. Basically it will divide the number of *Nhum*'s of a verb by all of its subjects (basically *Nhum+* + *Nhum-*) and multiply the result by 100.

Let us exemplify how a given property is validated. The psychological transitive verb construction *N-V-Nhum* may present a *Nhum* or a *N-hum* in the subject position. This property is lexically determined, i.e., it depends on the verb and therefore is explicitly represented in the matrix.

For the test sentences *A notícia abalou o Rui* 'The news shook Rui' and *O cliente enerva o sapateiro* 'The client annoys the shoemaker' we follow these steps:

- The verbs *abalar* a *enervar* are considered the main verbs in their sentences and are present in the list of psychological verbs, so the process continues to run for both sentences;
- The words *notícia* and *cliente* are found to be subjects to the above mentioned verbs and therefore they are marked as such (in this case *cliente* is marked as *Nhum* (*Nhum+*), while *notícias* is not (*Nhum-*));
- Then it is seen that the sentences have direct objects (CDIR) in relation to the verb; those are tested to see if they are human nouns (in this case both *sapateiro* and *Rui* are human nouns);
- Then all results that meet the pattern N0 V N1, with N1 being a human noun and V being a psychological verb, are sent to the REDUCER (each sentence will send two results, one with the verb and the other with the verb and the subject and its classification).

The results obtained from running the program on both these sentences are:

- VERBO: abalar 1
- VERBO: abalar NHUM-:notícia 1
- VERBO: enervar 1
- VERBO: enervar NHUM+:cliente 1

The results of this process on the entire corpus are then to be presented to the human specialist or to a learning algorithm. In this case, we confront the results with the data encoded in the matrix (see annex A, below).

## 3.5 Identification of the patterns *Nhum-V-se* and *Nhum-ser-Vpp*

### 3.5.1 Strategy

In the *Nhum-V-se* pattern (*-se* passive construction), a NP is followed by a verb with a reflexive (clitic) pronoun attached to it (*me*, *te*, *se*, *nos* and *vos*). The *Nhum-ser-Vpp* pattern (passive construction) consists of a NP, the auxiliary (copula) verb *ser* 'to be' and a past participle of a psychological verb.

These constructions are transforms (or alternations) of the basic direct transitive construction *N-V-Nhum*, where the obligatory *Nhum* direct object now appears in the subject position. These constructions are lexically dependent, i.e., their existence depends on the psychological verb and therefore, these transformational properties are explicitly represented in the matrices.

It should be noticed that clitic pronouns may appear in different syntactic positions, depending on many complex factors, such as:

- Sentence structure (main or subordinate clause):
  - *O público alegrou-se* 'The audience rejoiced'
  - *O João disse que o público se alegrou* 'João said that the audience rejoiced'
- Determination of the subject NP:
  - *(Ninguém/Toda a gente/Alguém/Só uma pessoa) se alegrou* '(No one/Everyone/Someone/Just one person) rejoiced'
- Adverbial modifiers, particularly those involving negation:
  - *O público (não/nunca) se alegrou* 'The audience (did not rejoice/never rejoiced)'
  - *O público (já/sempre) se alegrou* 'The audience (already rejoiced/always rejoices)'
- Verbal chains, the clitic can be raised from the main verb to one of its auxiliaries:
  - *O público começa a alegrar-se* 'The audience starts to rejoice'
  - *O público começa-se a alegrar* 'The audience starts to rejoice'

There are examples where both these patterns appear, and others where they cannot be applied. For example the verb *arrebatar* allows the passive transformation:

- *A actuação da Madonna arrebatou o público* 'Madonna's performance blew away the audience'

– *O público foi arrebatado pela actuação da Madonna* 'The audience was blown away by Madonna's performance'

But it does not seem to accept the *se*-passive transformation:

– **O público arrebatou-se (por/com/de) a actuação da Madonna* 'The audience blew due to Madonna's performance'

On the other hand, the verb *alegrar* has a symmetrical behavior regarding these two transformations:

– *A actuação da Madonna alegrou o público* 'Madonna's performance cheered the audience'

– **O público foi alegrado pela actuação da Madonna* 'The audience was cheered by Madonna's performance'

– *O público alegrou-se com a actuação da Madonna* 'The audience cheered with Madonna's performance'

### 3.5.2 Implementation of the *Nhum-V-se* search

The implementation of this pattern recognition is done sequentially. Each component is identified and only if all components are present and follow certain conditions (like the verb being a psychological verb) is the pattern placed in the output file. The process is executed in the following steps:

- First, the list of psychological verbs is fed from a file into a hashmap,
- Then, the program finds a noun and sees if the next word is a verb;
- The noun is checked to see if it is a *Nhum*;
- In case the next word is a verb and if it is present in the psychological verbs list, a search is made to see if it has a clitic reflexive pronoun attached;
- Should all these conditions be met, two results are sent to the REDUCER, one with the verb and the other with the verb and the sentence in which it appears;
- The process is then repeated for all xml trees present in the input file(s).

An example of this is an input file with the sentences *Os críticos enervam-se* 'The critics irk themselves, *A Joana arrebata-o* 'Joana blew him away', *A comida saciou-lhe o apetite* 'The food sated his appetite' and *Entusiasmaram-se as hostes* 'The hosts got carried away', in which we should identify the patterns only in the first sentence.

The second and third examples are there to verify if only the reflexive pronouns are being identified in the program (since *o* and *lhe* are not reflexive pronouns). In the fourth example we have an inverted subject transform that correspond to a *se* passive construction. However, as the subject is in a non-canonical position, it is not identified as such.

The process that is applied to this specific example is the following:

- For the first three sentences, a noun is identified and therefore the next step of the process is open (this is done by setting a variable to a given value);
- Next we see if the verbs that follow the nouns are identified as psychological, meaning that they pass to the next level of verification;
- The verb *enervar* shows the presence of clitics (although *saciar* and *arrebatar* also have clitics, they are not reflexive pronouns), therefore it fits the pattern and the results (just the verb and verb+sentence) are sent to the REDUCER.

The results obtained from this input file is:

- VERBO: enervar 1
- VERBO: arrebatar FRASE: Os críticos enervam-se 1

### 3.5.3 Implementation of the *Nhum-ser-Vpp* search

The implementation of this pattern identification has many aspects that mirror the previous program (*Nhum-V-se* pattern recognition), since both follow the same logic, differing mostly in what is identified in each step. The process is divided into the following parts:

- First the list of psychological verbs is fed from a file into a hashmap;
- Then the program finds a noun (and marks it as *Nhum* ou *N-hum*) and sees if the next word is the verb *ser*;
- Then we see if the next word in the pattern is a psychological verb in the past participle form (PASTPART tag);
- Should all of these conditions be met, two results are sent to the REDUCER, one with the verb and the other with the verb and the sentence in which it appears;
- The process is then repeated for all xml trees present in the input file(s).

An example of this is an input file with the sentences *Os críticos foram prejudicados* 'The critics were prejudicated', *O público foi arrebatado* 'The public was blown away', *A notícia chateia* 'The news annoy' and *O Rui foi-se galvanizando* 'Rui was getting rilled up', in which we should identify the patterns only in the first two sentences.

The process that is applied to this specific example is the following:

- For all sentences, a noun is identified and therefore the next step of the process is open (this is done by setting a variable to a given value);
- Next we see that the first two and the last sentences are followed by the verb *ser*, meaning that they pass to the next level of verifications;
- The first two verbs (*prejudicar* and *arrebatar*) are shown to be in the past participle form, therefor they fit the pattern and the results (just the verb and verb+sentence) are sent to the REDUCER.

The results obtained from this input file is:

- VERBO: arrebatar 1
- VERBO: prejudicar 1
- VERBO: arrebatar FRASE: O público foi arrebatado 1
- VERBO: prejudicar FRASE: Os críticos foram prejudicados 1

# 4 Evaluation and Results

## 4.1 Evaluation Process

The evaluation process serves the purpose of verifying if the programs are functioning according to their purpose and to validate the results achieved by them. The program verification is made by creating test input files that have a set of expected results, then after running the program, if the results found match the expected outcome we can assume the program is functioning properly. The result validation is a more complex procedure, since it involves more than the verification of the output given by the program, but also implies comparing the results with a sample of manually validated results.

### 4.1.1 Verbal Chains

The verbal chains identifier verification process started with choosing a list of one hundred (100) randomly chosen sentences and processed with the $L^2F$ Chain. The resulting output was manually scanned for the verbal chains it contained and afterward it was used to obtain the automatic program results. The results did not match due to some problems, mainly in the identification of all the words present in each verb category, and in the incorrect identification of chains. For instance identifying a VTEMP-VCOP and a VCOP-VCPART, when in reality it's a single chain VTEMP-VCOP-VCPART (like in the sentence *É por isso que essas raizes [...] vão ser lembradas nesta Quinzena Cultural [...]* 'Because of that, these roots [...] will be remembered in the cultural fortnight').

The expected results in the 100 sentence test are[1]:

- 1 VASP (acabar por) VINF (despistar)
- 1 VASP (ficar a) VINF (responder)
- 1 VASP (voltar a) VINF (fazer)
- 1 VCOP (estar) VCPART (concentrar)
- 1 VCOP (estar) VCPART (dividir)
- 1 VCOP (estar) VCPART (dotar)
- 1 VCOP (estar) VCPART (ligar)
- 1 VCOP (estar) VCPART (representar)
- 1 VCOP (ser) VCPART (declarar)
- 2 VCOP (ser) VCPART (deter)

[1] The chains with only one verb where omitted to simplify matters

- 1 VCOP (ser) VCPART (devolver)
- 1 VCOP (ser) VCPART (fiscalizar)
- 1 VCOP (ser) VCPART (hastear)
- 1 VCOP (ser) VCPART (obrigar)
- 1 VCOP (ser) VCPART (praticar)
- 1 VCOP (ser) VCPART (reabilitar)
- 1 VCOP (ser) VCPART (reunir)
- 1 VCOP (ser) VCPART (transaccionar)
- 1 VCOP (ser) VCPART (transformar)
- 1 VCOP (ser) VCPART (ultrapassar)
- 1 VMOD (dever) VINF (seguir)
- 1 VMOD (ficar de) VINF (responder)
- 1 VMOD (poder) VINF (escrever)
- 1 VMOD (poder)se) VINF (assistir)
- 1 VMOD (ter de) VINF (ser)
- 1 VTEMP (ir) VCOP (ser) VCPART (lembrar)
- 1 VTEMP (ir) VINF (estar)
- 1 VTEMP (ir) VINF (ultrapassar)
- 1 VTEMP (ter) VPP (abandonar)
- 1 VTEMP (ter) VPP (brilhar)
- 1 VTEMP (ter) VPP (encontrar)
- 1 VTEMP (ter) VPP (ganhar)
- 1 VTEMP (ter) VPP (ocupar)
- 1 VTEMP (ter) VPP (oferecer)
- 1 VTEMP (ter) VPP (prometer)
- 1 VTEMP (ter) VPP (recarregar)
- 1 VTEMP (ter) VPP (receber)
- 1 VTEMP (ter) VPP (ser)

### 4.1.2 Verb Chain Pattern Recognition

The pattern recognition program was also verified, using the same test file from the verbal chains. The 100 sentence file was run to find 1 pattern, and later it was run to find more than one.

The expected results in the 100 sentence test (for 1 pattern) are:

- PADRAO: 1 VCOP (ser) VCPART (transformar) FRASE: Assim os restos de uns pacíficos elefantes foram transformados pela imaginação humana em temíveis seres fantásticos.

The expected results in the 100 sentence test (for 3 pattern) are:

- PADRAO: 1 VCOP (ser) VCPART (transformar) FRASE: Assim os restos de uns pacíficos elefantes foram transformados pela imaginação humana em temíveis seres fantásticos.
- PADRAO: 1 VCOP (estar) VCPART (concentrar) FRASE: Esta está concentrada no Porto-Maria de Quiberon.

- PADRAO: 1 VMOD (dever) VINF (seguir) FRASE: As tropas italianas deviam seguir a partir de hoje para Balad (...)

Both searches (1 and 3 patterns) were applied to the test file, which resulted in no patterns/sentences being found. After resolving the issue that was causing this (a problem reading the pattern file), the results matched the expected data and the program was validated.

### 4.1.3 Human Noun Recognition

The human noun recognition applies both a program verification and a result validation, the first to discover program errors and the second to discover the inconsistencies between manual and automatic human noun identification that do not stem from the program. These inconsistencies happen because the chain does not identify all human nouns (either mentioned entities or other noun classes, such as certain professions).

For the first problem we used a list of one hundred (100) randomly chosen sentences and processed them in the $L^2F$ Chain, the resulting output was manually scanned for the patterns it contained and afterwards it was run in the *Nhum* recognition program. The results showed that certain hits were being duplicated due to catching subjects that were really in the patient role, and direct objects that were in the agent role (for instance in *O piloto foi ultrapassado pelo oponente* 'The pilot was overtaken by the adversary', *piloto* would be caught as a subject, despite not being it). After solving that problem, the results matched the test version, and the only thing left to test was the validity of the results.

The results showed 4 sentences with the pattern:

- *Rui marcaria Eunice tanto artística como pessoalmente.* 'Rui marked Eunice both artistically and personally' (for the verb *marcar* 'to mark')
- *Jogadores alugados preocupam a FIFA.* 'Loaned players worry FIFA' (for the verb *preocupar* 'to worry')
- *As mulheres vão ultrapassar os homens?* 'Will women overtake men?' (for the verb *ultrapassar* 'to overtake')
- *O brasileiro ainda ultrapassou o alemão [...]* 'The brasilian still overtook the german [...]' (also for the verb *ultrapassar* 'to overtake')

To validate the results found it was necessary to choose a list of verbs to analyze, it was decided to manually check the results of 10 verbs. The verbs chosen had less than 100 *N-V-Nhum* cases because of the difficulty of manually classifying such a large quantity of data.

Tabela 4.1: Results with Manual Checks

| Verb | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | % after manual check |
|---|---|---|---|---|
| abater | 99 | 79 | 79,80% | 90,91% |
| favorecer | 99 | 16 | 16,16% | 22,22% |
| impor | 69 | 33 | 47,83% | 56,52% |
| impressionar | 71 | 15 | 21,13% | 25,35% |
| influenciar | 75 | 28 | 37,33% | 38,67% |
| inspirar | 69 | 16 | 23,19% | 28,98% |
| orientar | 64 | 40 | 62,50% | 67,19% |
| reduzir | 85 | 37 | 43,53% | 51,76% |
| satisfazer | 87 | 7 | 8,05% | 10,34% |
| seduzir | 73 | 22 | 30,14% | 36,98% |

The results of these checks are present in table 4.1, and, for the most parts they coincide with the results found in the manual search. Yet some have discrepancies that stand in the 5-10% range. These are mostly caused by verbs that appeared with a large number of proper nouns (especially people's names) that the system does not identify as such.

### 4.1.4 *Nhum-V-se* and *Nhum-ser-Vpp* Patterns

The *Nhum-V-se* pattern recognition passed through a program validation step, to verify if the processed output matched the expected results. There is no need to make a result validation because there is no system error attached to this pattern.

The process is similar to those already described, a set number of sentences are manually scanned for results and then the program is run on the same data, if the results match then the program is considered to be validated. It was decided to use a set of 100 sentences that possess 3 cases where the pattern was present. After running the program in this input we found that the results matched the expected output and therefore the program was validated.

The expected results in the 100 sentence test are:

- 2 N-V-se:esgotar
- 1 N-V-se:esmagar
- 1 Nhum-V-se:esmagar

They show that 3 verbs have the *N-V-se* pattern (two times *esgotar* 'to drain' and one time *esmagar* 'to crush'), but only one of them (*esmagar*) has a human noun in the subject position[2].

[2]Notice that the verb *esmagar* is included in the psychological constructions of (Oliveira, 1984), even if this particular structure may be rare

Tabela 4.2: Verbal Chains by Length

| Verbal Chain | Nº of Appearances | % of Appearances |
|---|---|---|
| 1 verb | 40094 | 31,46% |
| 2 verbs | 56767 | 44,54% |
| 3 verbs | 28711 | 22,53% |
| 4 verbs | 1856 | 1,46% |
| 5 verbs | 21 | 0,01% |
| 6-7 verbs | 0 | 0,00% |

Like the *Nhum-V-se* pattern, the *Nhum-ser-Vpp* pattern needs only to pass a program validation, since these results are also without system error. Like above the same test set was used, although the number of patterns found is m, after running it we found that it too had matched the expected output.

The expected results in the 100 sentence test are:

- 1 N-ser-Vpp:reabilitar
- 1 N-ser-Vpp:transformar
- 1 N-ser-Vpp:ultrapassar

It shows that 3 verbs have the *N-V-se* pattern (*reabilitar* 'to rehabilitate', *transformar* 'to transform' and *ultrapassar* 'to overtake'), but none of them has an explicit human noun in the subject position.

## 4.2 Results Discussion

The results found for each program are mostly independent from each other and as such will be presented in distinct sections. The verbal chain results show information, like the number of times a chain length appears, while the various pattern results (*N-V-Nhum*, *N-V-se* and *N-ser-Vpp*) show cases that validate or contradict matrix values.

### 4.2.1 Verbal Chains

The verbal chains program allowed us to draw several conclusions, like the number of times each chain length occurs (4.2), what verbs appear the most (with or without auxiliaries) (table 4.3), what psychological verbs appear the most ( table 4.4[3]) and the most frequent auxiliary verbs (table 4.5).

Table 4.2 show us that only 31,46% of all verb instances are single forms, while verbal chains with auxiliaries represent 68,54%. This indicates the importance of taking into account verbal chains in order

[3]Notice that for the table 4.4 not all of these verbs are necessarily psychological verbs, nor do they appear in the corpus only in a psychological verb construction. These are just verbs represented in (Oliveira, 1984) lexical matrix.

Tabela 4.3: Most Frequent Verbs

| Main Verb | % with Auxiliaries | Nº with Auxiliaries | Nº without Auxiliaries |
|---|---|---|---|
| ter | 81.56% | 244595 | 55285 |
| fazer | 52.80% | 109257 | 97672 |
| estar | 91.45% | 187803 | 17559 |
| haver | 89.72% | 154117 | 17668 |
| dizer | 83.30% | 120717 | 24193 |
| dar | 61.30% | 59729 | 37711 |
| passar | 70.97% | 45453 | 18591 |
| ir | 87.00% | 60140 | 8990 |
| ver | 63.99% | 40101 | 22565 |
| ficar | 77.24% | 49994 | 14733 |

Tabela 4.4: Most Frequent Psychological Verbs

| Main Psychological Verb | % with Auxiliaries | Nº with Auxiliaries | Nº without Auxiliaries |
|---|---|---|---|
| perder | 61.30% | 18425 | 11631 |
| tentar | 63.72% | 18775 | 10689 |
| marcar | 45.60% | 10517 | 12545 |
| provocar | 55.78% | 8791 | 6970 |
| matar | 40.11% | 4828 | 7210 |
| atribuir | 37.54% | 4198 | 6986 |
| justificar | 76.10% | 8385 | 2633 |
| ultrapassar | 47.85% | 5178 | 5644 |
| transformar | 58.32% | 5174 | 3697 |
| preocupar | 43.69% | 3667 | 4726 |

Tabela 4.5: Most Frequent Auxiliary Verbs

| Auxiliary Verb | Verb Category | Nº of Appearances |
|---|---|---|
| ser | VCOP | 805676 |
| ter | VTEMP | 475242 |
| poder | VMOD | 374986 |
| ir | VTEMP | 239373 |
| estar | VCOP | 169245 |
| dever | VMOD | 162118 |
| estar a | VASP | 136404 |
| ter de | VMOD | 67234 |
| continuar a | VASP | 57065 |
| vir a | VASP | 48133 |

to analyse verb distribution and to determine verb senses and constructions. Verbal chains with just an auxiliary and a main verb constitute 44,54% of all verb chains and strings with 2 auxiliaries 22,53%; the remaining cases are about 2% (1,47%) of all verbal chains. This implies that, for the majority of verb chains, the basic (lexical) meaning of the verb is 'modified' by the adition of a single gramatical (temporal, aspectual or modal) value. Combinations of auxiliaries involve complex syntactical/semantic constraints that are out of the scope of this study. On the other hand, notice that the most frequent psychological verbs often appear in verbal chains (table 4.4). Therefore, it would be impossible to adequatly validate the data in the psychological verbs lexical matrix, without considering the verbal chains they are built in.

We now present table 4.3, with the most frequent verb instances in the corpus. For the most part they appear in verbal chains, i.e., with auxiliaries. As a side remark, notice that most of them may present grammatical status: *ter* 'to have', *fazer* 'to make', *haver* 'to be' and *dar* 'to give' are often used as support verbs (Gross, 1996), while *estar* 'to be' and *ficar* 'to stay' are copula verbs (with PPs).

Table 4.5 shows the most frequent type of auxiliary constructions, regardless of the length of the verbal chain they appear in. As one can see, the most frequent is VCOP *ser*, immediately followed by VTEMP *ter* and VMOD *poder*. Notice that VMOD *ter de* constitutes another, different construction.

### 4.2.2 Human Noun Identification

In this section we present the major results obtained in the process of *Nhum* identification and its contribution to validate the data in the lexical matrix of psychological verbs.

Table 4.6 shows results for a group of psychological verbs. In the first column we present the number of instances of each verb in the corpus followed by the number of times it appears in the *N-V-Nhum* pattern (and its percentage). Next the pattern *Nhum-V-Nhum* is shown (and the corresponding percentage in the *N-V-Nhum* pattern). It should be noticed that some verbs do not seem to appear at all in the psychological verb transitive construction in spite of their (feeble) frequency (*agoniar* 'to agonize' and *amargurar* 'to embitter'). An in-depth analysis should then be carried out in order to ascertain if the failure to recognize the transitive pattern is due to other factors (like incorrect parsing). Secondly, other verbs, even if they show up in the transitive pattern, do not seem to have a *Nhum* subject; on the other side we have verbs like *absorver* 'to absorb' that in the 47 patterns found, 25 of them have human nouns in their subjects, which allows some consistent data validation.

It is also interesting to consider the differences between verbs that have a similar number of *N-V-Nhum* patterns, like *abater* 'to shoot / to take down' (99) and *agitar* 'to shake / to agitate' (137). Despite appearing less times, *abater* has a far greater number of occurrences of the *Nhum-V-Nhum* pattern (79 times, which results in a 79,80% percentage of Nhum in the N0 position) than agitar (19 times / 13,87%).

Tabela 4.6: Noun-Verb-Noun Results

| Verb | Nº of occurrences | Nº of N-V-Nhum | % of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum |
|---|---|---|---|---|---|
| abalar | 794 | 67 | 8,44% | 20 | 29,85% |
| abater | 1035 | 99 | 9,57% | 79 | 79,80% |
| aborrecer | 303 | 11 | 3,63% | 2 | 18,18% |
| abrandar | 512 | 9 | 1,76% | 1 | 11,11% |
| absorver | 1003 | 33 | 3,29% | 19 | 57,58% |
| acalmar | 403 | 28 | 6,95% | 9 | 32,14% |
| admirar | 1883 | 30 | 1,59% | 22 | 73,33% |
| afectar | 3541 | 429 | 12,12% | 43 | 10,02% |
| afligir | 178 | 14 | 7,87% | 0 | 0,00% |
| agastar | 46 | 0 | 0,00% | 0 | 0,00% |
| agitar | 1367 | 137 | 10,02% | 19 | 13,87% |
| agoniar | 24 | 0 | 0,00% | 0 | 0,00% |
| alegrar | 253 | 10 | 3,95% | 1 | 10,00% |
| aliciar | 163 | 38 | 23,31% | 29 | 76,32% |
| alienar | 366 | 10 | 2,73% | 6 | 60,00% |
| alucinar | 23 | 0 | 0,00% | 0 | 0,00% |
| amaciar | 22 | 0 | 0,00% | 0 | 0,00% |
| amargurar | 27 | 0 | 0,00% | 0 | 0,00% |
| amedrontar | 30 | 5 | 16,67% | 1 | 20,00% |
| amenizar | 46 | 0 | 0,00% | 0 | 0,00% |

Next, in table 4.7 we compare these results with the property *Nhum-V-Nhum* encoded in the matrices. The value based on the corpus was attained in the following way: All results of 0,00% in the *Nhum-V-Nhum* pattern count are marked as (-), and all other results are marked as (+). This lead to some interesting results, as some values of both (-) and (+) conflict with the results presented in the matrices.

We cannot comment all cases here, but we discuss some of the most relevant aspects found from these results. Firstly, we consider the cases where *N0=Nhum* property described in the matrix is empirically confirmed by the pattern *Nhum-V-Nhum* found in the corpus (verbs *abalar, absorver, acalmar, aliciar, alienar and amedrontar*). Most of these verbs are unambiguously psychological verbs, the only exceptions are verbs *abalar* 'to leave / to shake' and *absorver* 'to absorb', which become an unambiguous psychological verb with a *Nhum* direct object.

On the other hand, where *N0≠Nhum* we cannot reach any conclusions, since the absence of the pattern may results from the fact that the corpus did not contain it, and not the non-existence of the pattern altogether. Therefore we conclude that only cases where the *N0=Nhum* property occur are relevant to the validation process.

In total, 302 verbs where shown to have the *Nhum-V-Nhum* property, and 68 verbs to lack it. Overall we found that for all relevant results (10 or more appearances of the verb in the corpus), 220 where correctly tagged and 150 showed results that contradicted the data in the matrices. This can also be due to the various meanings a verb can have, since some psychological verbs are known to have non-psychological uses.

Tabela 4.7: Noun-Verb-Noun Matrix Results

| Verb | % of Nhum-V-Nhum | Corpus Value | Matrix Value |
|---|---|---|---|
| abalar | 29,85% | + | + |
| abater | 79,80% | + | - |
| aborrecer | 18,18% | + | - |
| abrandar | 11,11% | + | - |
| absorver | 57,58% | + | + |
| acalmar | 32,14% | + | + |
| admirar | 73,33% | + | - |
| afectar | 10,02% | + | - |
| afligir | 0,00% | - | + |
| agastar | 0,00% | - | + |
| agitar | 13,87% | + | - |
| agoniar | 0,00% | - | + |
| alegrar | 10,00% | + | - |
| aliciar | 76,32% | + | + |
| alienar | 60,00% | + | + |
| alucinar | 0,00% | - | - |
| amaciar | 0,00% | - | + |
| amargurar | 0,00% | - | - |
| amedrontar | 20,00% | + | + |
| amenizar | 0,00% | - | + |

### 4.2.3 *Nhum-V-se* Pattern

This section shows the results obtained in the search for the *Nhum-V-se* pattern. Two sets of results were obtained in this case, the number of times the pattern occurs for each verb, and the corresponding percentage. These results are presented in table 4.8.

The analysis of these results reveals that some verbs do not possess the pattern, like *alucinar* 'to hallucinate', others, like *abalar* 'to shake', seem to have it in a low quantity (only 2 occurences). On the other hand, verbs like *admirar* 'to admire' are better represented (61 appearances) and have a high percentage (64,89%). Another interesting aspect of these results is the fact that some verbs, like *amargurar*, have N0=Nhum for all of their N-V-se patterns (basically 100,00% in Nhum-V-se).

Table 4.9 shows the list of verbs with the matrices classification, as well as the classifications we obtained from the results of table 4.8. The (-) classification (that marks the absence of the property) is placed in all cases where no patterns were found (0 ocurrences). The (+) classification (marks the presence of the pattern) on the other hand, is placed on all other verbs (1 or more appearances).

These tables show that not all results coincide with the data provided by the matrices. Some verbs classified as (+) do not present a single case like *amenizar* 'to smooth', while others, like *agitar* 'to agitate' defy their (-) classification by presenting a large number of examples of the *Nhum-V-se* pattern, such as: *(...) as hostes socialistas durienses agitam-se (...)* '(...) the socialist duriense hosts agitate themselves (...)' and *A JS agitava-se (...)* 'JS gets agitated (...)'.

Tabela 4.8: Nhum-V-se Results

| Verb | Nº of occurrences | Nº of N-V-se | % of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se |
|---|---|---|---|---|---|
| abalar | 794 | 13 | 1,64% | 2 | 15,38% |
| abater | 1035 | 186 | 17,97% | 13 | 6,99% |
| aborrecer | 303 | 73 | 24,09% | 28 | 38,36% |
| abrandar | 512 | 1 | 0,20% | 0 | 0,00% |
| absorver | 1003 | 22 | 2,19% | 6 | 27,27% |
| acalmar | 403 | 39 | 9,68% | 14 | 35,90% |
| admirar | 1883 | 94 | 4,99% | 61 | 64,89% |
| afectar | 3541 | 63 | 1,78% | 6 | 9,52% |
| afligir | 178 | 31 | 17,42% | 14 | 45,16% |
| agastar | 46 | 5 | 10,87% | 4 | 80,00% |
| agitar | 1367 | 255 | 18,65% | 80 | 31,37% |
| agoniar | 24 | 1 | 4,17% | 0 | 0,00% |
| alegrar | 253 | 29 | 11,46% | 8 | 27,59% |
| aliciar | 163 | 14 | 8,59% | 4 | 28,57% |
| alienar | 366 | 8 | 2,19% | 2 | 25,00% |
| alucinar | 23 | 1 | 4,35% | 0 | 0,00% |
| amaciar | 22 | 3 | 13,64% | 1 | 33,33% |
| amargurar | 27 | 3 | 11,11% | 3 | 100,00% |
| amedrontar | 30 | 6 | 20,00% | 4 | 66,67% |
| amenizar | 46 | 4 | 8,70% | 0 | 0,00% |

Tabela 4.9: Nhum-V-se Matrix Results

| Verb | Nº of occurrences | % of the Pattern | Corpus Value | Matrix Value |
|---|---|---|---|---|
| abalar | 2 | 15,38% | + | - |
| abater | 13 | 6,99% | + | - |
| aborrecer | 28 | 38,36% | + | + |
| abrandar | 0 | 0,00% | - | + |
| absorver | 6 | 27,27% | + | + |
| acalmar | 14 | 35,90% | + | + |
| admirar | 61 | 64,89% | + | + |
| afectar | 6 | 9,52% | + | - |
| afligir | 14 | 45,16% | + | + |
| agastar | 4 | 80,00% | + | + |
| agitar | 80 | 31,37% | + | - |
| agoniar | 0 | 0,00% | - | + |
| alegrar | 8 | 27,59% | + | + |
| aliciar | 4 | 28,57% | + | + |
| alienar | 2 | 25,00% | + | + |
| alucinar | 0 | 0,00% | - | - |
| amaciar | 1 | 33,33% | + | - |
| amargurar | 3 | 100,00% | + | + |
| amedrontar | 4 | 66,67% | + | + |
| amenizar | 0 | 0,00% | - | - |

Tabela 4.10: Nhum-ser-Vpp Results

| Verb | Nº of occurrences | Nº of N-ser-Vpp | % of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp |
|---|---|---|---|---|---|
| abalar | 794 | 149 | 18,77% | 21 | 14,09% |
| abater | 1035 | 515 | 49,76% | 160 | 31,07% |
| aborrecer | 303 | 32 | 10,56% | 2 | 6,25% |
| abrandar | 512 | 0 | 0,00% | 0 | 0,00% |
| absorver | 1003 | 342 | 34,10% | 21 | 6,14% |
| acalmar | 403 | 6 | 1,49% | 0 | 0,00% |
| admirar | 1883 | 40 | 2,12% | 13 | 32,50% |
| afectar | 3541 | 1026 | 28,97% | 193 | 18,81% |
| afligir | 178 | 2 | 1,12% | 1 | 50,00% |
| agastar | 46 | 1 | 2,17% | 0 | 0,00% |
| agitar | 1367 | 96 | 7,02% | 7 | 7,29% |
| agoniar | 24 | 1 | 4,17% | 0 | 0,00% |
| alegrar | 253 | 3 | 1,19% | 0 | 0,00% |
| aliciar | 163 | 39 | 23,93% | 32 | 82,05% |
| alienar | 366 | 150 | 40,98% | 19 | 12,67% |
| alucinar | 23 | 1 | 4,35% | 0 | 0,00% |
| amaciar | 22 | 0 | 0,00% | 0 | 0,00% |
| amargurar | 27 | 2 | 7,41% | 0 | 0,00% |
| amedrontar | 30 | 0 | 0,00% | 0 | 0,00% |
| amenizar | 46 | 3 | 6,52% | 0 | 0,00% |

Still some results maintain the classification presented by the matrices. *Alucinar* 'to hallucinate' (-) shows no cases of the pattern, while *aborrecer* 'to bore' (+) has, as seen in the sample sentences *(...) os jovens aborrecem-se* '(...) young people get bored' and *As pessoas aborrecem-se* 'People get bored'.

We have found 286 verbs with the *Nhum-V-se* pattern against 84 verbs (like *alucinar* 'to hallucinate') that did not appear in that construction. From these results, 234 verbs appear to confirm the *se*-passive property as it was encoded in the matrix, while 136 differed from it.

### 4.2.4 *Nhum-ser-Vpp* Pattern

Much like the above section, this also focuses in the search of a pattern (*Nhum-ser-Vpp*), and presents two sets of results, the number of cases and the percentage of cases for each verb. The results are displayed in table 4.10.

The analysis of the results shows verbs that do not possess the pattern (*abrandar* 'to slow'), verbs that have very low results (like *afligir* 'to afflict', which has 1 occurrence) and verbs that use the pattern extensively (at 160 occurrences and a percentage of 31,07%, *abater* is a prime example of this).

In table 4.11 we show the matrix classification for this pattern, as well as the classification obtained from the results in table 4.10. The classification process is the same as the one applied to the previous section, with cases with zero occurrences being classified as (-), with all other results being classified as (+).

Tabela 4.11: Nhum-ser-Vpp Matrix Results

| Verb | Nº of occurrences | % of the Pattern | Corpus Value | Matrix Value |
|---|---|---|---|---|
| abalar | 21 | 14,09% | + | + |
| abater | 160 | 31,07% | + | - |
| aborrecer | 2 | 6,25% | + | - |
| abrandar | 0 | 0,00% | - | - |
| absorver | 21 | 6,14% | + | + |
| acalmar | 0 | 0,00% | - | + |
| admirar | 13 | 32,50% | + | - |
| afectar | 193 | 18,81% | + | + |
| afligir | 1 | 50,00% | + | + |
| agastar | 0 | 0,00% | - | - |
| agitar | 7 | 7,29% | + | - |
| agoniar | 0 | 0,00% | - | + |
| alegrar | 0 | 0,00% | - | - |
| aliciar | 32 | 82,05% | + | + |
| alienar | 19 | 12,67% | + | + |
| alucinar | 0 | 0,00% | - | - |
| amaciar | 0 | 0,00% | - | + |
| amargurar | 0 | 0,00% | - | - |
| amedrontar | 0 | 0,00% | - | + |
| amenizar | 0 | 0,00% | - | + |

Like in the previous patterns the results obtained from running the search programs on the corpus do not always confirm the values encoded in the matrix. A case where the property was expected but did not appear is *amaciar* 'to soften', while *admirar* 'to admire' shows the opposite situation. Naturally, while the first situation may correspond to incorrect values in the matrix that need to be corrected, this latter case is often due to lexical polissemy, i.e. the sentence structure is possible but it corresponds to another construction of the verb. In the case of *admirar* the passive transform is associated to the transitive construction with the obligatory human subject *Nhum-V-N1* but not with the psychological construction of the same verb (the reverse occurs with the *se*-passive structure). As the sentences *(...) um país onde os desportistas são admirados e respeitados (...)* '(...) a country were sportsmen are admired and respected (...)' and *(...) Galvão de Melo era admirado na Força Aérea (...)* '(...) Galvão de Melo was admired in the Air Force' prove.

Examples where the properties are consistent with the data from the matrices are *abrandar* 'to slow' (-) and *aliciar* 'to entice' (+), with the latter being validated by sentences like *À partida, os jogadores serão aliciados com a oferta de crédito ilimitado (...)* 'At the start, gamblers will be enticed with the offer of ilimited credit (...)' and *Alguns empresários foram aliciados a participar no esquema (...)* 'Some entrepreneurs were enticed to participate in the scam (...)'.

We have found 202 verbs with the Nhum-ser-Vpp pattern against 168 verbs (like *amaciar* 'to soften') that did not appear in that construction. From these results, 236 verbs confirm the property as it was encoded in the matrix, while 134 differed from it.

Tabela 4.12: Overall Results

| Verb | Nº of occurrences | Nº of total patterns | % of total patterns | N-V-se Corpus Value | N-ser-Vpp Corpus Value | Nhum-V-Nhum Corpus Value |
|---|---|---|---|---|---|---|
| abalar | 794 | 43 | 5,42% | + | + | + |
| abater | 1035 | 252 | 24,35% | + | + | + |
| aborrecer | 303 | 32 | 10,56% | + | + | + |
| abrandar | 512 | 1 | 0,20% | - | - | + |
| absorver | 1003 | 46 | 4,59% | + | + | + |
| acalmar | 403 | 23 | 5,71% | + | - | + |
| admirar | 1883 | 96 | 5,10% | + | + | + |
| afectar | 3541 | 242 | 6,83% | + | + | + |
| afligir | 178 | 15 | 8,43% | + | + | - |
| agastar | 46 | 4 | 8,70% | + | - | - |
| agitar | 1367 | 106 | 7,75% | + | + | + |
| agoniar | 24 | 0 | 0,00% | - | - | - |
| alegrar | 253 | 9 | 3,56% | + | - | + |
| aliciar | 163 | 65 | 39,88% | + | + | + |
| alienar | 366 | 27 | 7,38% | + | + | + |
| alucinar | 23 | 0 | 0,00% | - | - | - |
| amaciar | 22 | 1 | 4,55% | + | - | - |
| amargurar | 27 | 3 | 11,11% | + | - | - |
| amedrontar | 30 | 5 | 16,67% | + | - | + |
| amenizar | 46 | 0 | 0,00% | - | - | - |

### 4.2.5 Overall Results

Some results cross the boundaries of each pattern and show a bigger picture of certain conclusions that where reached with this work. One of these is the number of times all these patterns appear as well as the relevance that number has in relation to the total. Another is the correlation that the presence of one property might have on the others. All these results are shown on table 4.12.

Only 41 verbs show an absence of all properties studied, among them are verbs like *descontentar* 'to baffle' and *esperançar* 'to hope'. On the other side of the spectrum are 150 verbs that present all the properties studied, with the emphasis in verbs with a large quantity of appearances like *marcar* 'to score / to mark' and *perder* 'to lose'. For those two verbs we observe some interesting facts, *marcar* does not present the *Nhum-V-se* pattern in the matrix, and *perder* lacks the *Nhum-V-Nhum* and *Nhum-ser-Vpp* patterns. One of the reasons for this is the ambiguity of the verbs themselves, since *perder* is known to rarely appear as a psychological verb, yet several instances of patterns with it were found. For instance,

Tabela 4.13: Matrix and Corpus Results

| Pattern | Matrix value | Corpus value | Nº of occurrences |
|---|---|---|---|
| Nhum-V-se | - | - | 34 |
| Nhum-V-se | - | + | 50 |
| Nhum-V-se | + | - | 86 |
| Nhum-V-se | + | + | 200 |
| Nhum-ser-Vpp | - | - | 90 |
| Nhum-ser-Vpp | - | + | 78 |
| Nhum-ser-Vpp | + | - | 56 |
| Nhum-ser-Vpp | + | + | 146 |
| Nhum-V-Nhum | - | - | 67 |
| Nhum-V-Nhum | - | + | 68 |
| Nhum-V-Nhum | + | - | 82 |
| Nhum-V-Nhum | + | + | 153 |

the sentence *O Benfica perdeu Maxi nos últimos quatro jogos* 'Benfica lost Maxi in the last four matches' does not present the verb in a psychological meaning, but the program does not distinguish this and therefore considers it a case of *Nhum-V-Nhum*. The same happens for the case of the *Nhum-V-se* pattern for the *marcar* verb, as seen in the example *O João marcou-me [...]* 'João scored/marked me', that may be used in a psychological sense *O João marcou-me pelo seu carácter* 'João marked me with his character', or in a non-psychological sense *O João marcou-me um golo* 'João scored me a goal'.

Finally, another aspect that is worth of remark is the fact that the cases where the corpus results differ from the values in the matrix consist mostly in (-) that turn into (+) and not the other way around, this is relevant since these are the cases that allow us to draw conclusions, since the opposite (- in corpus and + in matrix) can result from the size of the data studied, and not from the non-existence of a given pattern. This is seen in table 4.13.

In this comparison, we also see that the more significant results are those of pattern presence both in the corpus and on the matrix, while all other combinations (absence in one of them, or in both) have a smaller number of occurrences. Another relevant aspect is that the results that allow conclusions to be drawn (+ classifications in the corpus values) are roughly 2/3 of all cases (250/370 in *Nhum-V-se*, 224/370 in *Nhum-ser-Vpp* and 221/370 in *Nhum-V-Nhum*).

# 5 Conclusion and Future Work

## 5.1 Conclusion

We made a brief presentation of the tools used in the validation process, like the $L^2F$ processing chain, used to process the corpus, or the GRID computing paradigm, which provided enough processing power to deal with large amounts of text. We also used tools like Condor, to provide schedueling in the corpus processing stage, and Hadoop, to provide easy access and management of the large amount of information resulting from the processed CETEMPúblico corpus.

Strategies were then created to obtain results needed to verify the information in the matrices. The application of these strategies led to the creation of programs to that effect. The programs passed an evaluation process and yielded results that allow linguists to verify if the properties encoded in the matrices where or not confirmed by the empirical evidence retrieved from the corpus.

A conclusion to be drawn from this work is that the extraction of syntactical patterns from corpora is not sufficient to describe the lexicon, since numerous problems of ambiguity prevent the determination of what structural properties are observed in a given construction (or meaning) of a lexical item. Therefore, the fact that a verb presents a given structure (passive for instance) does not mean that sentence form corresponds to all the meanings that verb could present. Just a case-by-case analysis allows a rigorous confirmation of the relation between sentence forms and their meaning. In this sense, the work here developed proved to be an important tool to aid lexicographical descriptions, mainly in the construction of computational lexicons.

## 5.2 Future Work

Here the future possibilities to enrich this work are discussed, and these vary from direct approaches (implementing searches for properties that where not dealt with) to indirect ones (like improving the lexicon of the processing chain with the information retrieved). The main aspects that can be improved in future endeavors are:

- The creation of programs to validate the information in other matrices,

- Use of the conclusions of this work to improve the lexical information in the processing chain (mainly by identifying words with tags),
- Improving the programs usability, by adding features to it (like capturing different tags that were not present in the current version of the processing chain).

Of these, the enrichment of the lexicon seems the most important one, since it allows the improvement of several other projects that use the processing chain. The improvement of the program's usability may be not only relevant, but also necessary, since the processing chain is a constantly evolving entity, and a drastic change in the way certain words are cataloged with tags could modify the results obtained.

# Bibliography


Baptista, J. (2005). *Sintaxe dos Nomes Predicativos com verbo-suporte SER DE*. Lisbon, Portugal: Fundação para a Ciência e a Tecnologia/Fundação Calouste Gulbenkian.

Casteleiro, J. (1981). *Sintaxe Transformacional do Adjectivo: Regência das Construções Completivas.* Lisbon, INIC.

Condor Team. (2008, November). *Condor Version 7.0.5 Manual.* Madison, USA.

Dean, J. (2006). Experiences with MapReduce, an abstraction for large-scale computation. In *Pact '06: Proceedings of the 15th international conference on parallel architectures and compilation techniques.* New York, USA: ACM.

Freire, H. (1994). *Determinação e formalização das propriedades sintácticas de adjectivos terminados em -vel.* Lisbon, FLUL.

Grefenstette, G. (1994). *Exploration in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Gross, M. (1996). Lexicon Grammar. In *Concise encyclopedia of syntactic theories* (p. 244-258). Cambridge, MA, USA: Pergamon.

Hadoop Team. (2008). *Hadoop.* (http://hadoop.apache.org/core/)

Harris, Z. (1964). The elementary Transformations. In *Transformations and discourse analysis papers 54* (p. 482-532). Dordrecht, Netherlands: Dreidel.

Higgins, D. (2004). Which statistics reflect semantics? Rethinking synonymy and word similarity. In *Proceedings of the international conference on linguistic evidence* (p. 265-284). Berlin, Germany: Mouton de Gruyter.

Hodgson, J. (1991). Informational Constraints on Pre-lexical Priming. In *Language and cognitive processes* (p. 169-205). Hove, UK: Psychology Press.

Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. In *Bulletin de la société vaudoise des sciences naturelles* (Vol. 37, p. 241-272). Lausanne, Switzerland: La Société Vaudoise des Sciences Naturelles.

Kesselman, C., & Foster, I. (1998). *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA, USA: Morgan Kaufmann Publishers.

Levin, B. (1993). *English Class Verbs and Alternations: A preliminary Investigation*. Chicago, IL, USA: University of Chicago Press.

Lowe, W. (2001). Towards a Theory of Semantic Space. In *Proceedings of the 23rd annual conference of the cognitive science society* (p. 576-581). Cambridge, MA, USA: Lawrence Erlbaum Associates.

Mamede, N. (2007). *A cadeia de processamento XIP em Maio de 2007*.

Mamede, N., & Baptista, J. (2009). *Cadeias Verbais*.

McCarthy, D., Koeling, R., Weeds, J., & Carroll, J. (2004). Finding predominant senses in untagged text. In *Proceedings of the 42rd annual meeting of the association for computer linguistics* (p. 280-287). Morristown, NJ, USA: Association for Computational Linguistics.

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42rd annual meeting of the association for computer linguistics* (p. 17-24). Morristown, NJ, USA: Association for Computational Linguistics.

Medeiros, J. (1995). *Processamento morfológico e correcção ortográfica do português.* Lisbon, Portugal.

Oliveira, M. (1984). *Syntaxe des Verbes Psychologiques du Portugais.* Lisbon, INIC.

Padó, S., & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. In *Computational linguistics* (Vol. 33, p. 161-199). Cambridge, MA, USA: MIT Press.

Pardal, J. (2007). *Manual do utilizador do RuDriCo.* Laboratório de Sistemas de Língua Falada.

Ranchhod, E. (1990). *Sintaxe dos predicados nominais com Estar.* Lisbon, INIC.

R.C.E. Xerox. (2003). *XIP user guide.* (http://www.xrce.xerox.com)

Ribeiro, R., Mamede, N., & Trancoso, I. (2003). Using morphossyntactic information in TTS systems: comparing strategies for european portuguese. In *Computational processing of the portuguese language: 6th international workshop.* London, UK: Springer-Verlag.

Rubenstein, H., & Goodenough, J. (1965). Contextual Correlates of Synonymy. In *Communications of the acm* (Vol. 8, p. 627-633). New York, NY, USA: ACM.

Salton, G. (1989). *Automatic text processing - The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley.

Santos, D., & Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th annual meeting of the association for computational linguistics* (p. 442-449). Morristown, NJ, USA: Association for Computational Linguistics.

Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th european conference on machine learning* (p. 491-502). London, UK: Springer-Verlag.

# I Appendices

# Annex: Matrix with the full results

Tabela A.1: Full Results

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| abalar | 794 | 13 | 2 | 15,38% | + | - | 149 | 21 | 14,09% | + | + | 67 | 20 | 29,85% | + | + |
| abater | 1035 | 186 | 13 | 6,99% | + | - | 515 | 160 | 31,07% | + | - | 99 | 79 | 79,80% | + | - |
| aborrecer | 303 | 73 | 28 | 38,36% | + | + | 32 | 2 | 6,25% | + | - | 11 | 2 | 18,18% | + | - |
| abrandar | 512 | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 9 | 1 | 11,11% | + | - |
| absorver | 1003 | 22 | 6 | 27,27% | + | + | 342 | 21 | 6,14% | + | + | 33 | 19 | 57,58% | + | + |
| acalmar | 403 | 39 | 14 | 35,90% | + | + | 6 | 0 | 0,00% | - | + | 28 | 9 | 32,14% | + | + |
| admirar | 1883 | 94 | 61 | 64,89% | + | + | 40 | 13 | 32,50% | + | - | 30 | 22 | 73,33% | + | - |
| afectar | 3541 | 63 | 6 | 9,52% | + | - | 1026 | 193 | 18,81% | + | + | 429 | 43 | 10,02% | + | - |
| afligir | 178 | 31 | 14 | 45,16% | + | + | 2 | 1 | 50,00% | + | + | 14 | 0 | 0,00% | - | + |
| agastar | 46 | 5 | 4 | 80,00% | + | + | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + |
| agitar | 1367 | 255 | 80 | 31,37% | + | - | 96 | 7 | 7,29% | + | - | 137 | 19 | 13,87% | + | - |
| agoniar | 24 | 1 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| alegrar | 253 | 29 | 8 | 27,59% | + | + | 3 | 0 | 0,00% | - | - | 10 | 1 | 10,00% | + | - |
| aliciar | 163 | 14 | 4 | 28,57% | + | + | 39 | 32 | 82,05% | + | + | 38 | 29 | 76,32% | + | + |
| alienar | 366 | 8 | 2 | 25,00% | + | + | 150 | 19 | 12,67% | + | + | 10 | 6 | 60,00% | + | + |
| alucinar | 23 | 1 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| amaciar | 22 | 3 | 1 | 33,33% | + | - | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| amargurar | 27 | 3 | 3 | 100,00% | + | + | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| amedrontar | 30 | 6 | 4 | 66,67% | + | + | 0 | 0 | 0,00% | - | + | 5 | 1 | 20,00% | + | + |
| amenizar | 46 | 4 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| amolecer | 20 | 2 | 1 | 50,00% | + | + | 0 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| amuar | 19 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| angustiar | 70 | 20 | 9 | 45,00% | + | + | 1 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + |
| animar | 2201 | 188 | 82 | 43,62% | + | + | 242 | 25 | 10,33% | + | + | 136 | 29 | 21,32% | + | + |
| aniquilar | 86 | 8 | 2 | 25,00% | + | - | 41 | 11 | 26,83% | + | + | 3 | 2 | 66,67% | + | + |
| apaixonar | 494 | 214 | 134 | 62,62% | + | + | 26 | 11 | 42,31% | + | - | 13 | 4 | 30,77% | + | - |
| apavorar | 33 | 4 | 1 | 25,00% | + | + | 0 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | + |
| apaziguar | 71 | 12 | 1 | 8,33% | + | - | 7 | 0 | 0,00% | - | + | 7 | 2 | 28,57% | + | + |
| aperfeiçoar | 143 | 19 | 5 | 26,32% | + | + | 21 | 2 | 9,52% | + | + | 9 | 5 | 55,56% | + | + |
| aplacar | 16 | 4 | 1 | 25,00% | + | + | 0 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| apoquentar | 23 | 2 | 1 | 50,00% | + | + | 0 | 0 | 0,00% | - | + | 3 | 2 | 66,67% | + | + |
| aquietar | 14 | 5 | 1 | 20,00% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| arrebatar | 153 | 11 | 5 | 45,45% | + | + | 25 | 6 | 24,00% | + | - | 13 | 5 | 38,46% | + | - |
| arrepiar | 102 | 20 | 4 | 20,00% | + | + | 7 | 2 | 28,57% | + | - | 2 | 0 | 0,00% | - | - |
| assombrar | 79 | 3 | 1 | 33,33% | + | + | 7 | 1 | 14,29% | + | - | 8 | 2 | 25,00% | + | - |
| atordoar | 17 | 0 | 0 | 0,00% | - | + | 2 | 1 | 50,00% | + | + | 5 | 3 | 60,00% | + | + |
| atrair | 1098 | 52 | 9 | 17,31% | + | - | 104 | 43 | 41,35% | + | + | 277 | 52 | 18,77% | + | + |
| atrapalhar | 170 | 51 | 19 | 37,25% | + | + | 2 | 1 | 50,00% | + | + | 11 | 1 | 9,09% | + | + |
| atribuir | 7285 | 757 | 272 | 35,93% | + | - | 4027 | 1106 | 27,46% | + | - | 205 | 108 | 52,68% | + | - |
| azedar | 133 | 18 | 2 | 11,11% | + | + | 0 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | - |
| baralhar | 266 | 19 | 5 | 26,32% | + | + | 5 | 0 | 0,00% | - | + | 17 | 0 | 0,00% | - | + |
| beneficiar | 3325 | 15 | 2 | 13,33% | + | - | 168 | 63 | 37,50% | + | + | 247 | 64 | 25,91% | + | + |
| bloquear | 1044 | 17 | 4 | 23,53% | + | + | 106 | 10 | 9,43% | + | - | 22 | 11 | 50,00% | + | - |
| cansar | 1271 | 214 | 138 | 64,49% | + | + | 1 | 0 | 0,00% | - | - | 11 | 2 | 18,18% | + | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| captar | 406 | 12 | 5 | 41,67% | + | - | 161 | 9 | 5,59% | + | + | 13 | 7 | 53,85% | + | + |
| cativar | 188 | 17 | 2 | 11,76% | + | - | 12 | 5 | 41,67% | + | + | 32 | 7 | 21,88% | + | + |
| cegar | 77 | 13 | 1 | 7,69% | + | - | 0 | 0 | 0,00% | - | - | 3 | 2 | 66,67% | + | - |
| chagar | 47 | 2 | 1 | 50,00% | + | - | 0 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| chatear | 176 | 39 | 20 | 51,28% | + | + | 1 | 0 | 0,00% | - | + | 2 | 2 | 100,00% | + | + |
| chocar | 1173 | 70 | 12 | 17,14% | + | + | 2 | 1 | 50,00% | + | - | 38 | 3 | 7,89% | + | - |
| civilizar | 26 | 6 | 1 | 16,67% | + | + | 13 | 3 | 23,08% | + | + | 1 | 0 | 0,00% | - | + |
| comover | 260 | 57 | 25 | 43,86% | + | + | 5 | 0 | 0,00% | - | - | 21 | 5 | 23,81% | + | - |
| compadecer | 151 | 16 | 2 | 12,50% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| comprometer | 3437 | 2504 | 1392 | 55,59% | + | + | 28 | 2 | 7,14% | + | + | 38 | 7 | 18,42% | + | + |
| confortar | 145 | 14 | 3 | 21,43% | + | + | 3 | 1 | 33,33% | + | + | 7 | 3 | 42,86% | + | + |
| confundir | 1193 | 428 | 65 | 15,19% | + | + | 103 | 20 | 19,42% | + | - | 59 | 14 | 23,73% | + | - |
| conquistar | 3030 | 59 | 15 | 25,42% | + | - | 218 | 21 | 9,63% | + | + | 243 | 139 | 57,20% | + | + |
| consolar | 340 | 33 | 17 | 51,52% | + | + | 3 | 1 | 33,33% | + | + | 15 | 5 | 33,33% | + | + |
| consternar | 22 | 2 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - |
| constranger | 46 | 6 | 1 | 16,67% | + | + | 19 | 14 | 73,68% | + | - | 2 | 2 | 100,00% | + | - |
| contagiar | 142 | 6 | 1 | 16,67% | + | - | 27 | 13 | 48,15% | + | + | 35 | 4 | 11,43% | + | + |
| contaminar | 319 | 11 | 1 | 9,09% | + | - | 123 | 43 | 34,96% | + | + | 30 | 9 | 30,00% | + | + |
| contentar | 620 | 233 | 93 | 39,91% | + | + | 0 | 0 | 0,00% | - | + | 16 | 3 | 18,75% | + | + |
| conter | 2546 | 34 | 7 | 20,59% | + | - | 47 | 3 | 6,38% | + | + | 36 | 2 | 5,56% | + | + |
| contrariar | 1802 | 9 | 2 | 22,22% | + | + | 181 | 10 | 5,52% | + | + | 46 | 22 | 47,83% | + | + |
| corromper | 53 | 7 | 2 | 28,57% | + | - | 6 | 2 | 33,33% | + | + | 1 | 0 | 0,00% | - | + |
| danar | 159 | 1 | 0 | 0,00% | - | + | 8 | 4 | 50,00% | + | - | 13 | 6 | 46,15% | + | - |
| decepcionar | 188 | 7 | 4 | 57,14% | + | + | 1 | 1 | 100,00% | + | + | 17 | 8 | 47,06% | + | + |
| degradar | 319 | 216 | 18 | 8,33% | + | + | 9 | 2 | 22,22% | + | + | 1 | 0 | 0,00% | - | + |
| deleitar | 58 | 34 | 16 | 47,06% | + | + | 0 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| deliciar | 169 | 91 | 39 | 42,86% | + | + | 0 | 0 | 0,00% | - | - | 12 | 8 | 66,67% | + | - |
| denegrir | 28 | 9 | 0 | 0,00% | - | - | 13 | 3 | 23,08% | + | + | 1 | 0 | 0,00% | - | + |
| deprimir | 136 | 4 | 1 | 25,00% | + | + | 7 | 0 | 0,00% | - | - | 5 | 0 | 0,00% | - | - |
| derreter | 74 | 26 | 2 | 7,69% | + | + | 8 | 1 | 12,50% | + | - | 2 | 0 | 0,00% | - | - |
| desacreditar | 85 | 9 | 2 | 22,22% | + | - | 6 | 0 | 0,00% | - | + | 6 | 0 | 0,00% | - | + |
| desalentar | 26 | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| desanimar | 164 | 3 | 1 | 33,33% | + | + | 0 | 0 | 0,00% | - | + | 12 | 0 | 0,00% | - | + |
| desapontar | 215 | 5 | 3 | 60,00% | + | + | 0 | 0 | 0,00% | - | + | 17 | 6 | 35,29% | + | + |
| desarmar | 370 | 14 | 5 | 35,71% | + | + | 46 | 16 | 34,78% | + | - | 12 | 6 | 50,00% | + | - |
| desassossegar | 26 | 1 | 1 | 100,00% | + | + | 0 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | - |
| desconcertar | 72 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 2 | 1 | 50,00% | + | - |
| desconsolar | 12 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| descontentar | 147 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| desculpar | 707 | 202 | 114 | 56,44% | + | + | 3 | 1 | 33,33% | + | + | 17 | 13 | 76,47% | + | + |
| desenganar | 171 | 23 | 3 | 13,04% | + | - | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| desenrascar | 38 | 18 | 6 | 33,33% | + | + | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + |
| desesperar | 402 | 9 | 2 | 22,22% | + | + | 34 | 1 | 2,94% | + | - | 16 | 1 | 6,25% | + | - |
| desgostar | 53 | 10 | 4 | 40,00% | + | + | 0 | 0 | 0,00% | - | - | 2 | 1 | 50,00% | + | - |
| desgraçar | 31 | 3 | 0 | 0,00% | - | + | 5 | 1 | 20,00% | + | - | 1 | 0 | 0,00% | - | - |
| desiludir | 767 | 39 | 8 | 20,51% | + | + | 0 | 0 | 0,00% | - | + | 59 | 22 | 37,29% | + | + |
| deslumbrar | 117 | 22 | 10 | 45,45% | + | + | 1 | 0 | 0,00% | - | - | 9 | 3 | 33,33% | + | - |
| desmascarar | 63 | 5 | 1 | 20,00% | + | - | 22 | 4 | 18,18% | + | + | 3 | 1 | 33,33% | + | + |
| desmoralizar | 39 | 2 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + | 2 | 0 | 0,00% | - | + |
| desnortear | 23 | 7 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | - |
| desolar | 40 | 0 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| desonrar | 10 | 1 | 1 | 100,00% | + | + | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| desorientar | 57 | 25 | 9 | 36,00% | + | + | 1 | 0 | 0,00% | - | - | 8 | 0 | 0,00% | - | - |
| desprestigiar | 16 | 2 | 1 | 50,00% | + | + | 1 | 1 | 100,00% | + | + | 1 | 0 | 0,00% | - | + |
| desvanecer | 148 | 93 | 5 | 5,38% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| dilacerar | 38 | 3 | 0 | 0,00% | - | + | 12 | 4 | 33,33% | + | - | 2 | 1 | 50,00% | + | - |
| distinguir | 2044 | 471 | 152 | 32,27% | + | - | 328 | 152 | 46,34% | + | + | 195 | 98 | 50,26% | + | + |
| distrair | 211 | 61 | 19 | 31,15% | + | + | 10 | 7 | 70,00% | + | - | 14 | 3 | 21,43% | + | + |
| divertir | 773 | 375 | 125 | 33,33% | + | + | 87 | 8 | 9,20% | + | + | 14 | 3 | 21,43% | + | + |
| electrizar | 15 | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 3 | 1 | 33,33% | + | - |
| embaraçar | 200 | 9 | 3 | 33,33% | + | + | 2 | 0 | 0,00% | - | + | 59 | 20 | 33,90% | + | + |
| embasbacar | 12 | 1 | 1 | 100,00% | + | + | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| embatucar | 10 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + |
| embebedar | 26 | 19 | 7 | 36,84% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| embriagar | 29 | 7 | 2 | 28,57% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| emocionar | 196 | 67 | 34 | 50,75% | + | + | 3 | 0 | 0,00% | - | - | 10 | 3 | 30,00% | + | - |
| empobrecer | 49 | 3 | 1 | 33,33% | + | + | 1 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | - |
| empolgar | 61 | 12 | 5 | 41,67% | + | + | 0 | 0 | 0,00% | - | - | 12 | 4 | 33,33% | + | - |
| emudecer | 38 | 2 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | - | 1 | 1 | 100,00% | + | - |
| enaltecer | 267 | 9 | 4 | 44,44% | + | - | 32 | 5 | 15,63% | + | + | 13 | 9 | 69,23% | + | + |
| enamorar | 22 | 8 | 2 | 25,00% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| encadear | 27 | 14 | 2 | 14,29% | + | - | 4 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| encantar | 244 | 31 | 13 | 41,94% | + | + | 1 | 0 | 0,00% | - | - | 24 | 8 | 33,33% | + | - |
| encarniçar | 10 | 4 | 1 | 25,00% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| encurralar | 52 | 1 | 0 | 0,00% | - | - | 6 | 5 | 83,33% | + | + | 6 | 1 | 16,67% | + | + |
| endiabrar | 11 | 0 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| endoidecer | 13 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| endurecer | 243 | 8 | 0 | 0,00% | - | + | 4 | 0 | 0,00% | - | + | 2 | 2 | 100,00% | + | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| enervar | 109 | 44 | 18 | 40,91% | + | + | 0 | 0 | 0,00% | - | + | 23 | 4 | 17,39% | + | + |
| enfadar | 11 | 5 | 2 | 40,00% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| enfeitiçar | 15 | 3 | 0 | 0,00% | - | - | 2 | 1 | 50,00% | + | + | 1 | 0 | 0,00% | - | + |
| enfurecer | 159 | 34 | 19 | 55,88% | + | + | 0 | 0 | 0,00% | - | + | 51 | 7 | 13,73% | + | + |
| enganar | 1486 | 337 | 150 | 44,51% | + | - | 107 | 58 | 54,21% | + | + | 41 | 21 | 51,22% | + | + |
| engrandecer | 12 | 8 | 1 | 12,50% | + | + | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| enjoar | 42 | 5 | 2 | 40,00% | + | + | 0 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| enlouquecer | 84 | 6 | 1 | 16,67% | + | - | 0 | 0 | 0,00% | - | - | 7 | 2 | 28,57% | + | - |
| enraivecer | 10 | 2 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + | 2 | 1 | 50,00% | + | + |
| enredar | 87 | 54 | 19 | 35,19% | + | - | 3 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| enriquecer | 347 | 47 | 4 | 8,51% | + | - | 63 | 7 | 11,11% | + | - | 12 | 8 | 66,67% | + | - |
| enrolar | 80 | 30 | 7 | 23,33% | + | - | 5 | 1 | 20,00% | + | + | 3 | 0 | 0,00% | - | + |
| entalar | 48 | 7 | 2 | 28,57% | + | - | 2 | 1 | 50,00% | + | + | 7 | 6 | 85,71% | + | + |
| entediar | 10 | 3 | 1 | 33,33% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| enternecer | 15 | 7 | 3 | 42,86% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - |
| entreter | 247 | 144 | 75 | 52,08% | + | + | 5 | 1 | 20,00% | + | + | 7 | 4 | 57,14% | + | + |
| entristecer | 47 | 18 | 4 | 22,22% | + | + | 0 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | - |
| entusiasmar | 627 | 118 | 71 | 60,17% | + | + | 2 | 1 | 50,00% | + | + | 63 | 7 | 11,11% | + | + |
| envelhecer | 171 | 4 | 1 | 25,00% | + | - | 2 | 1 | 50,00% | + | - | 3 | 2 | 66,67% | + | - |
| envenenar | 98 | 7 | 3 | 42,86% | + | - | 30 | 16 | 53,33% | + | + | 4 | 3 | 75,00% | + | + |
| envergonhar | 183 | 23 | 8 | 34,78% | + | + | 5 | 1 | 20,00% | + | + | 11 | 3 | 27,27% | + | + |
| enxovalhar | 19 | 0 | 0 | 0,00% | - | + | 10 | 7 | 70,00% | + | + | 4 | 1 | 25,00% | + | + |
| escandalizar | 149 | 18 | 10 | 55,56% | + | + | 0 | 0 | 0,00% | - | + | 18 | 2 | 11,11% | + | + |
| esfalfar | 11 | 7 | 2 | 28,57% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| esfriar | 72 | 1 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| esgotar | 1613 | 342 | 31 | 9,06% | + | + | 53 | 1 | 1,89% | + | - | 20 | 3 | 15,00% | + | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| esmagar | 413 | 48 | 13 | 27,08% | + | - | 136 | 38 | 27,94% | + | + | 62 | 30 | 48,39% | + | + |
| espantar | 1128 | 111 | 52 | 46,85% | + | + | 17 | 1 | 5,88% | + | - | 25 | 5 | 20,00% | + | - |
| esperançar | 66 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| espevitar | 34 | 7 | 2 | 28,57% | + | + | 2 | 1 | 50,00% | + | + | 4 | 0 | 0,00% | - | + |
| espezinhar | 32 | 3 | 1 | 33,33% | + | - | 25 | 6 | 24,00% | + | + | 2 | 1 | 50,00% | + | + |
| espicaçar | 42 | 3 | 2 | 66,67% | + | - | 4 | 0 | 0,00% | - | + | 7 | 1 | 14,29% | + | + |
| estafar | 18 | 0 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| estarrecer | 20 | 1 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| estorvar | 15 | 3 | 1 | 33,33% | + | - | 2 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| estragar | 454 | 65 | 17 | 26,15% | + | + | 13 | 1 | 7,69% | + | + | 4 | 2 | 50,00% | + | + |
| exacerbar | 64 | 10 | 2 | 20,00% | + | + | 18 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| exaltar | 267 | 108 | 25 | 23,15% | + | + | 8 | 1 | 12,50% | + | + | 8 | 3 | 37,50% | + | + |
| exasperar | 49 | 13 | 10 | 76,92% | + | + | 0 | 0 | 0,00% | - | + | 9 | 2 | 22,22% | + | + |
| exaurir | 12 | 1 | 1 | 100,00% | + | - | 2 | 1 | 50,00% | + | - | 2 | 0 | 0,00% | - | - |
| excitar | 124 | 22 | 8 | 36,36% | + | + | 4 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + |
| extasiar | 37 | 6 | 4 | 66,67% | + | + | 0 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| extenuar | 11 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| extraviar | 17 | 7 | 1 | 14,29% | + | + | 3 | 1 | 33,33% | + | + | 1 | 1 | 100,00% | + | + |
| falsear | 29 | 1 | 0 | 0,00% | - | - | 14 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| fartar | 384 | 176 | 110 | 62,50% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| fascinar | 288 | 67 | 14 | 20,90% | + | + | 6 | 3 | 50,00% | + | - | 24 | 8 | 33,33% | + | - |
| fatigar | 36 | 4 | 1 | 25,00% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| favorecer | 930 | 24 | 3 | 12,50% | + | - | 73 | 22 | 30,14% | + | + | 99 | 16 | 16,16% | + | + |
| ferir | 1048 | 69 | 25 | 36,23% | + | + | 224 | 163 | 72,77% | + | - | 183 | 84 | 45,90% | + | - |
| flagelar | 53 | 5 | 1 | 20,00% | + | + | 10 | 5 | 50,00% | + | - | 4 | 1 | 25,00% | + | - |
| formalizar | 814 | 7 | 1 | 14,29% | + | + | 261 | 16 | 6,13% | + | - | 16 | 8 | 50,00% | + | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fortalecer | 204 | 66 | 4 | 6,06% | + | + | 13 | 2 | 15,38% | + | + | 3 | 1 | 33,33% | + | + |
| fortificar | 18 | 2 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| fraudar | 50 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| frustrar | 249 | 14 | 2 | 14,29% | + | + | 46 | 1 | 2,17% | + | - | 6 | 3 | 50,00% | + | - |
| fulminar | 28 | 4 | 3 | 75,00% | + | - | 7 | 5 | 71,43% | + | + | 3 | 3 | 100,00% | + | + |
| fustigar | 132 | 6 | 1 | 16,67% | + | - | 25 | 9 | 36,00% | + | + | 10 | 6 | 60,00% | + | + |
| galvanizar | 53 | 11 | 6 | 54,55% | + | + | 1 | 0 | 0,00% | - | + | 18 | 3 | 16,67% | + | + |
| gelar | 125 | 5 | 1 | 20,00% | + | - | 7 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - |
| guiar | 436 | 67 | 30 | 44,78% | + | - | 56 | 7 | 12,50% | + | + | 32 | 25 | 78,13% | + | + |
| harmonizar | 71 | 19 | 3 | 15,79% | + | + | 15 | 1 | 6,67% | + | + | 1 | 0 | 0,00% | - | + |
| hipnotizar | 16 | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + |
| honrar | 217 | 15 | 7 | 46,67% | + | - | 22 | 9 | 40,91% | + | + | 10 | 6 | 60,00% | + | + |
| horrorizar | 62 | 15 | 5 | 33,33% | + | + | 0 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - |
| hostilizar | 42 | 3 | 1 | 33,33% | + | + | 7 | 1 | 14,29% | + | + | 5 | 2 | 40,00% | + | + |
| humanizar | 44 | 15 | 4 | 26,67% | + | + | 2 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + |
| ilibar | 356 | 15 | 5 | 33,33% | + | + | 111 | 62 | 55,86% | + | + | 139 | 63 | 45,32% | + | - |
| iludir | 384 | 21 | 3 | 14,29% | + | + | 17 | 8 | 47,06% | + | + | 17 | 10 | 58,82% | + | + |
| iluminar | 372 | 89 | 6 | 6,74% | + | - | 68 | 4 | 5,88% | + | + | 8 | 3 | 37,50% | + | + |
| ilustrar | 1281 | 27 | 21 | 77,78% | + | - | 114 | 1 | 0,88% | + | + | 49 | 11 | 22,45% | + | + |
| imolar | 78 | 35 | 26 | 74,29% | + | - | 14 | 7 | 50,00% | + | + | 3 | 3 | 100,00% | + | + |
| imortalizar | 41 | 9 | 2 | 22,22% | + | + | 12 | 2 | 16,67% | + | + | 3 | 1 | 33,33% | + | + |
| impacientar | 75 | 28 | 14 | 50,00% | + | + | 0 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | - |
| impor | 3823 | 1286 | 366 | 28,46% | + | - | 660 | 295 | 44,70% | + | + | 69 | 33 | 47,83% | + | + |
| importunar | 25 | 1 | 0 | 0,00% | - | - | 13 | 7 | 53,85% | + | + | 5 | 1 | 20,00% | + | + |
| impressionar | 821 | 48 | 15 | 31,25% | + | + | 2 | 0 | 0,00% | - | + | 71 | 15 | 21,13% | + | + |
| incapacitar | 13 | 1 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| incendiar | 374 | 124 | 10 | 8,06% | + | + | 159 | 18 | 11,32% | + | - | 13 | 9 | 69,23% | + | - |
| incensar | 16 | 1 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| inchar | 37 | 2 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| incomodar | 745 | 82 | 23 | 28,05% | + | + | 84 | 26 | 30,95% | + | + | 109 | 22 | 20,18% | + | + |
| incriminar | 104 | 5 | 2 | 40,00% | + | - | 44 | 28 | 63,64% | + | + | 37 | 14 | 37,84% | + | + |
| indignar | 487 | 173 | 81 | 46,82% | + | + | 2 | 0 | 0,00% | - | - | 32 | 3 | 9,38% | + | - |
| indispor | 12 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | + |
| infamar | 10 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| inferiorizar | 11 | 1 | 0 | 0,00% | - | + | 2 | 1 | 50,00% | + | + | 0 | 0 | 0,00% | - | + |
| inflamar | 61 | 17 | 3 | 17,65% | + | + | 5 | 0 | 0,00% | - | - | 6 | 2 | 33,33% | + | - |
| influenciar | 1140 | 44 | 13 | 29,55% | + | - | 293 | 70 | 23,89% | + | + | 75 | 28 | 37,33% | + | + |
| inibir | 179 | 26 | 5 | 19,23% | + | + | 11 | 7 | 63,64% | + | + | 22 | 3 | 13,64% | + | + |
| injuriar | 20 | 2 | 2 | 100,00% | + | - | 2 | 0 | 0,00% | - | + | 2 | 1 | 50,00% | + | + |
| inocentar | 135 | 2 | 2 | 100,00% | + | + | 3 | 1 | 33,33% | + | + | 4 | 2 | 50,00% | + | + |
| inquietar | 196 | 81 | 27 | 33,33% | + | + | 3 | 2 | 66,67% | + | + | 19 | 5 | 26,32% | + | - |
| inspirar | 1077 | 326 | 99 | 30,37% | + | + | 157 | 24 | 15,29% | + | + | 69 | 16 | 23,19% | + | - |
| insultar | 242 | 57 | 28 | 49,12% | + | - | 69 | 32 | 46,38% | + | + | 52 | 40 | 76,92% | + | + |
| insurgir | 745 | 673 | 456 | 67,76% | + | + | 0 | 0 | 0,00% | - | - | 3 | 1 | 33,33% | + | - |
| interessar | 4381 | 498 | 197 | 39,56% | + | + | 3 | 2 | 66,67% | + | - | 52 | 16 | 30,77% | + | - |
| intimidar | 135 | 20 | 9 | 45,00% | + | + | 19 | 13 | 68,42% | + | + | 22 | 7 | 31,82% | + | + |
| intoxicar | 37 | 1 | 1 | 100,00% | + | - | 10 | 6 | 60,00% | + | + | 10 | 1 | 10,00% | + | + |
| intrigar | 238 | 10 | 2 | 20,00% | + | - | 0 | 0 | 0,00% | - | - | 25 | 4 | 16,00% | + | - |
| irar | 45 | 1 | 1 | 100,00% | + | + | 0 | 0 | 0,00% | - | - | 2 | 2 | 100,00% | + | - |
| irritar | 738 | 186 | 117 | 62,90% | + | + | 1 | 0 | 0,00% | - | + | 129 | 36 | 27,91% | + | + |
| isolar | 1252 | 458 | 181 | 39,52% | + | - | 119 | 14 | 11,76% | + | - | 68 | 26 | 38,24% | + | - |
| jubilar | 33 | 4 | 2 | 50,00% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| justiçar | 55 | 0 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | + |
| justificar | 12197 | 1418 | 560 | 39,49% | + | + | 889 | 46 | 5,17% | + | + | 528 | 57 | 10,80% | + | + |
| lenir | 49 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 2 | 2 | 100,00% | + | - |
| lesar | 134 | 1 | 0 | 0,00% | - | + | 60 | 34 | 56,67% | + | + | 27 | 12 | 44,44% | + | + |
| liquidar | 302 | 11 | 5 | 45,45% | + | - | 113 | 14 | 12,39% | + | + | 22 | 14 | 63,64% | + | + |
| lixar | 182 | 11 | 9 | 81,82% | + | + | 5 | 2 | 40,00% | + | + | 1 | 0 | 0,00% | - | + |
| ludibriar | 22 | 0 | 0 | 0,00% | - | - | 12 | 6 | 50,00% | + | + | 8 | 4 | 50,00% | + | + |
| maçar | 38 | 3 | 2 | 66,67% | + | + | 3 | 0 | 0,00% | - | + | 2 | 2 | 100,00% | + | + |
| magoar | 153 | 48 | 16 | 33,33% | + | + | 3 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| maravilhar | 110 | 21 | 11 | 52,38% | + | + | 0 | 0 | 0,00% | - | - | 8 | 4 | 50,00% | + | - |
| marcar | 16095 | 142 | 24 | 16,90% | + | - | 2256 | 116 | 5,14% | + | + | 738 | 336 | 45,53% | + | + |
| martirizar | 17 | 3 | 0 | 0,00% | - | + | 11 | 5 | 45,45% | + | + | 3 | 1 | 33,33% | + | + |
| massacrar | 349 | 12 | 5 | 41,67% | + | + | 138 | 95 | 68,84% | + | + | 51 | 35 | 68,63% | + | + |
| matar | 6314 | 383 | 193 | 50,39% | + | - | 3266 | 2400 | 73,48% | + | - | 1721 | 1060 | 61,59% | + | - |
| melhorar | 2441 | 19 | 3 | 15,79% | + | - | 153 | 6 | 3,92% | + | + | 28 | 5 | 17,86% | + | + |
| melindrar | 30 | 4 | 1 | 25,00% | + | + | 0 | 0 | 0,00% | - | + | 4 | 0 | 0,00% | - | + |
| metamorfosear | 29 | 20 | 8 | 40,00% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| minar | 335 | 10 | 1 | 10,00% | + | - | 20 | 3 | 15,00% | + | - | 10 | 3 | 30,00% | + | - |
| mobilizar | 1274 | 119 | 57 | 47,90% | + | + | 274 | 156 | 56,93% | + | + | 333 | 118 | 35,44% | + | + |
| moderar | 441 | 7 | 3 | 42,86% | + | + | 210 | 10 | 4,76% | + | + | 5 | 2 | 40,00% | + | + |
| modificar | 661 | 149 | 10 | 6,71% | + | + | 133 | 7 | 5,26% | + | + | 15 | 3 | 20,00% | + | + |
| moer | 40 | 3 | 1 | 33,33% | + | + | 2 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + |
| monopolizar | 143 | 1 | 0 | 0,00% | - | - | 18 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + |
| motivar | 1752 | 32 | 8 | 25,00% | + | + | 391 | 28 | 7,16% | + | + | 59 | 12 | 20,34% | + | + |
| neutralizar | 138 | 15 | 8 | 53,33% | + | - | 52 | 6 | 11,54% | + | + | 12 | 6 | 50,00% | + | + |
| notabilizar | 96 | 68 | 40 | 58,82% | + | + | 0 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ofender | 224 | 35 | 21 | 60,00% | + | + | 21 | 5 | 23,81% | + | + | 24 | 13 | 54,17% | + | + |
| ofuscar | 145 | 3 | 0 | 0,00% | - | + | 24 | 2 | 8,33% | + | - | 12 | 4 | 33,33% | + | - |
| orgulhar | 416 | 261 | 133 | 50,96% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| orientar | 842 | 157 | 32 | 20,38% | + | - | 236 | 24 | 10,17% | + | + | 64 | 40 | 62,50% | + | + |
| pacificar | 67 | 5 | 1 | 20,00% | + | + | 4 | 1 | 25,00% | + | + | 6 | 3 | 50,00% | + | + |
| paralizar | 18 | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 2 | 0 | 0,00% | - | + |
| pasmar | 179 | 47 | 9 | 19,15% | + | + | 10 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| penalizar | 661 | 18 | 2 | 11,11% | + | + | 317 | 116 | 36,59% | + | - | 92 | 37 | 40,22% | + | - |
| perder | 18556 | 986 | 156 | 15,82% | + | + | 125 | 9 | 7,20% | + | - | 1053 | 424 | 40,27% | + | - |
| perseguir | 747 | 110 | 36 | 32,73% | + | - | 227 | 118 | 51,98% | + | + | 126 | 65 | 51,59% | + | + |
| personalizar | 71 | 7 | 1 | 14,29% | + | + | 19 | 3 | 15,79% | + | + | 1 | 1 | 100,00% | + | + |
| perturbar | 622 | 51 | 5 | 9,80% | + | + | 98 | 5 | 5,10% | + | + | 45 | 5 | 11,11% | + | + |
| perverter | 37 | 3 | 0 | 0,00% | - | + | 5 | 2 | 40,00% | + | + | 1 | 1 | 100,00% | + | + |
| petrificar | 14 | 3 | 1 | 33,33% | + | + | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| picar | 164 | 17 | 10 | 58,82% | + | + | 13 | 5 | 38,46% | + | + | 11 | 6 | 54,55% | + | + |
| polir | 39 | 4 | 0 | 0,00% | - | + | 10 | 1 | 10,00% | + | + | 0 | 0 | 0,00% | - | + |
| politizar | 32 | 9 | 1 | 11,11% | + | + | 5 | 2 | 40,00% | + | + | 0 | 0 | 0,00% | - | + |
| popularizar | 48 | 18 | 3 | 16,67% | + | - | 8 | 3 | 37,50% | + | + | 3 | 1 | 33,33% | + | + |
| prejudicar | 1498 | 81 | 13 | 16,05% | + | + | 505 | 171 | 33,86% | + | + | 161 | 35 | 21,74% | + | + |
| prender | 4236 | 2779 | 187 | 6,73% | + | + | 230 | 171 | 74,35% | + | - | 484 | 316 | 65,29% | + | - |
| preocupar | 5395 | 906 | 377 | 41,61% | + | + | 1 | 1 | 100,00% | + | - | 533 | 40 | 7,50% | + | - |
| prestigiar | 57 | 2 | 0 | 0,00% | - | + | 5 | 1 | 20,00% | + | + | 4 | 1 | 25,00% | + | + |
| profanar | 50 | 0 | 0 | 0,00% | - | - | 23 | 3 | 13,04% | + | + | 1 | 1 | 100,00% | + | + |
| prostrar | 13 | 6 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| proteger | 1089 | 188 | 59 | 31,38% | + | + | 148 | 31 | 20,95% | + | + | 107 | 44 | 41,12% | + | + |
| provocar | 9379 | 158 | 29 | 18,35% | + | - | 633 | 36 | 5,69% | + | + | 361 | 30 | 8,31% | + | + |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pulverizar | 95 | 10 | 3 | 30,00% | + | - | 23 | 3 | 13,04% | + | + | 3 | 2 | 66,67% | + | + |
| purificar | 30 | 10 | 1 | 10,00% | + | - | 8 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| ralar | 52 | 5 | 4 | 80,00% | + | + | 0 | 0 | 0,00% | - | - | 6 | 1 | 16,67% | + | - |
| reabilitar | 209 | 15 | 7 | 46,67% | + | + | 65 | 14 | 21,54% | + | + | 12 | 6 | 50,00% | + | + |
| reanimar | 98 | 15 | 2 | 13,33% | + | + | 12 | 2 | 16,67% | + | + | 2 | 2 | 100,00% | + | + |
| rebaixar | 27 | 7 | 4 | 57,14% | + | - | 13 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | + |
| rebelar | 59 | 46 | 21 | 45,65% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| recalcar | 14 | 0 | 0 | 0,00% | - | - | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| recompor | 127 | 55 | 16 | 29,09% | + | + | 3 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| reconciliar | 151 | 88 | 38 | 43,18% | + | + | 1 | 1 | 100,00% | + | + | 2 | 1 | 50,00% | + | + |
| reconfortar | 28 | 3 | 1 | 33,33% | + | + | 1 | 1 | 100,00% | + | + | 3 | 1 | 33,33% | + | + |
| reconquistar | 151 | 5 | 4 | 80,00% | + | - | 10 | 0 | 0,00% | - | + | 8 | 4 | 50,00% | + | + |
| recrear | 20 | 9 | 2 | 22,22% | + | + | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| redimir | 145 | 107 | 46 | 42,99% | + | - | 8 | 1 | 12,50% | + | + | 3 | 1 | 33,33% | + | + |
| reduzir | 5093 | 693 | 50 | 7,22% | + | - | 1683 | 96 | 5,70% | + | + | 85 | 37 | 43,53% | + | + |
| refinar | 58 | 9 | 0 | 0,00% | - | + | 6 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| refrear | 99 | 2 | 0 | 0,00% | - | + | 13 | 2 | 15,38% | + | + | 2 | 2 | 100,00% | + | + |
| regalar | 48 | 9 | 3 | 33,33% | + | + | 1 | 0 | 0,00% | - | - | 3 | 1 | 33,33% | + | - |
| regenerar | 28 | 6 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| regozijar | 192 | 121 | 64 | 52,89% | + | + | 0 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | - |
| regrar | 123 | 0 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| rejubilar | 160 | 1 | 1 | 100,00% | + | + | 0 | 0 | 0,00% | - | - | 3 | 2 | 66,67% | + | - |
| rejuvenescer | 30 | 4 | 1 | 25,00% | + | - | 1 | 0 | 0,00% | - | + | 2 | 1 | 50,00% | + | + |
| relaxar | 63 | 6 | 1 | 16,67% | + | + | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| renovar | 1172 | 87 | 13 | 14,94% | + | - | 210 | 28 | 13,33% | + | - | 26 | 13 | 50,00% | + | - |
| repousar | 266 | 3 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 4 | 0 | 0,00% | - | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| reprimir | 134 | 8 | 2 | 25,00% | + | + | 45 | 3 | 6,67% | + | + | 9 | 6 | 66,67% | + | + |
| restabelecer | 254 | 18 | 3 | 16,67% | + | + | 178 | 0 | 0,00% | - | - | 4 | 2 | 50,00% | + | - |
| retalhar | 33 | 3 | 1 | 33,33% | + | - | 10 | 2 | 20,00% | + | - | 0 | 0 | 0,00% | - | - |
| reter | 703 | 32 | 10 | 31,25% | + | - | 106 | 21 | 19,81% | + | + | 29 | 10 | 34,48% | + | + |
| retrair | 93 | 77 | 35 | 45,45% | + | + | 1 | 0 | 0,00% | - | - | 2 | 1 | 50,00% | + | - |
| revigorar | 12 | 3 | 1 | 33,33% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| revoltar | 351 | 219 | 127 | 57,99% | + | + | 0 | 0 | 0,00% | - | - | 6 | 2 | 33,33% | + | - |
| ridicularizar | 77 | 9 | 2 | 22,22% | + | + | 30 | 5 | 16,67% | + | + | 4 | 1 | 25,00% | + | + |
| saciar | 40 | 6 | 2 | 33,33% | + | + | 2 | 1 | 50,00% | + | - | 1 | 1 | 100,00% | + | - |
| safar | 151 | 71 | 26 | 36,62% | + | + | 2 | 1 | 50,00% | + | - | 2 | 0 | 0,00% | - | + |
| sangrar | 46 | 1 | 1 | 100,00% | + | - | 3 | 0 | 0,00% | - | - | 1 | 1 | 100,00% | + | - |
| satisfazer | 2647 | 39 | 13 | 33,33% | + | + | 300 | 15 | 5,00% | + | - | 87 | 7 | 8,05% | + | - |
| saturar | 104 | 3 | 0 | 0,00% | - | + | 3 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| secar | 179 | 8 | 0 | 0,00% | - | - | 19 | 1 | 5,26% | + | - | 1 | 0 | 0,00% | - | - |
| seduzir | 258 | 40 | 6 | 15,00% | + | - | 19 | 10 | 52,63% | + | + | 73 | 22 | 30,14% | + | + |
| sensibilizar | 292 | 23 | 9 | 39,13% | + | + | 26 | 19 | 73,08% | + | - | 59 | 33 | 55,93% | + | - |
| serenar | 109 | 1 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 7 | 3 | 42,86% | + | + |
| seringar | 57 | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| sobressaltar | 79 | 17 | 7 | 41,18% | + | + | 9 | 2 | 22,22% | + | + | 4 | 0 | 0,00% | - | + |
| sofisticar | 99 | 10 | 2 | 20,00% | + | + | 21 | 3 | 14,29% | + | + | 1 | 0 | 0,00% | - | + |
| sossegar | 327 | 24 | 13 | 54,17% | + | + | 7 | 2 | 28,57% | + | + | 48 | 25 | 52,08% | + | + |
| suavizar | 90 | 12 | 0 | 0,00% | - | + | 14 | 1 | 7,14% | + | + | 2 | 0 | 0,00% | - | + |
| subjugar | 31 | 5 | 2 | 40,00% | + | - | 20 | 4 | 20,00% | + | + | 1 | 0 | 0,00% | - | + |
| sublimar | 91 | 4 | 0 | 0,00% | - | - | 7 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| subverter | 153 | 7 | 1 | 14,29% | + | + | 25 | 3 | 12,00% | + | + | 1 | 0 | 0,00% | - | + |
| sufocar | 71 | 6 | 0 | 0,00% | - | + | 16 | 1 | 6,25% | + | + | 7 | 2 | 28,57% | + | + |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| supreender | 23 | 2 | 2 | 100,00% | + | + | 3 | 3 | 100,00% | + | - | 8 | 5 | 62,50% | + | - |
| suster | 77 | 4 | 1 | 25,00% | + | - | 10 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + |
| temperar | 200 | 8 | 0 | 0,00% | - | - | 32 | 0 | 0,00% | - | - | 4 | 1 | 25,00% | + | - |
| tentar | 18107 | 100 | 22 | 22,00% | + | + | 223 | 69 | 30,94% | + | + | 54 | 33 | 61,11% | + | + |
| tocar | 4359 | 171 | 60 | 35,09% | + | - | 232 | 45 | 19,40% | + | - | 216 | 105 | 48,61% | + | - |
| tolher | 32 | 11 | 1 | 9,09% | + | + | 3 | 1 | 33,33% | + | - | 3 | 0 | 0,00% | - | - |
| tonificar | 11 | 2 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| torturar | 181 | 11 | 3 | 27,27% | + | + | 103 | 66 | 64,08% | + | + | 9 | 6 | 66,67% | + | + |
| trair | 321 | 69 | 26 | 37,68% | + | - | 88 | 37 | 42,05% | + | + | 42 | 20 | 47,62% | + | + |
| tramar | 93 | 11 | 3 | 27,27% | + | + | 10 | 4 | 40,00% | + | + | 23 | 4 | 17,39% | + | + |
| tranquilizar | 229 | 45 | 20 | 44,44% | + | + | 2 | 1 | 50,00% | + | + | 61 | 25 | 40,98% | + | + |
| transcender | 111 | 22 | 5 | 22,73% | + | - | 2 | 0 | 0,00% | - | - | 3 | 0 | 0,00% | - | - |
| transfigurar | 137 | 91 | 23 | 25,27% | + | + | 11 | 2 | 18,18% | + | + | 3 | 1 | 33,33% | + | + |
| transformar | 5509 | 3173 | 504 | 15,88% | + | + | 621 | 69 | 11,11% | + | + | 122 | 41 | 33,61% | + | + |
| transir | 11 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| transportar | 3053 | 127 | 32 | 25,20% | + | - | 1508 | 765 | 50,73% | + | - | 468 | 227 | 48,50% | + | - |
| transtornar | 30 | 5 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + | 1 | 0 | 0,00% | - | + |
| traumatizar | 44 | 3 | 0 | 0,00% | - | + | 1 | 1 | 100,00% | + | + | 2 | 0 | 0,00% | - | - |
| trespassar | 58 | 4 | 0 | 0,00% | - | - | 24 | 9 | 37,50% | + | - | 4 | 2 | 50,00% | + | + |
| triturar | 20 | 0 | 0 | 0,00% | - | + | 16 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | + |
| trucidar | 45 | 1 | 0 | 0,00% | - | + | 28 | 14 | 50,00% | + | + | 10 | 9 | 90,00% | + | + |
| turbar | 47 | 3 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 6 | 0 | 0,00% | - | - |
| turvar | 20 | 0 | 0 | 0,00% | - | + | 0 | 0 | 0,00% | - | - | 1 | 0 | 0,00% | - | - |
| ufanar | 13 | 3 | 1 | 33,33% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| ultrajar | 12 | 0 | 0 | 0,00% | - | + | 3 | 1 | 33,33% | + | + | 1 | 0 | 0,00% | - | + |
| ultrapassar | 5448 | 108 | 29 | 26,85% | + | - | 973 | 88 | 9,04% | + | - | 284 | 147 | 51,76% | + | - |

| Verb | Nº of Verbs | Nº of N-V-se | Nº of Nhum-V-se | % of Nhum-V-se | Nhum-V-se Corpus Value | Nhum-V-se Matrix Value | Nº of N-ser-Vpp | Nº of Nhum-ser-Vpp | % of Nhum-ser-Vpp | Nhum-ser-Vpp Corpus Value | Nhum-ser-Vpp Matrix Value | Nº of N-V-Nhum | Nº of Nhum-V-Nhum | % of Nhum-V-Nhum | Nhum-V-Nhum Corpus Value | Nhum-V-Nhum Matrix Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| valorizar | 952 | 446 | 25 | 5,61% | + | + | 96 | 19 | 19,79% | + | + | 11 | 6 | 54,55% | + | + |
| vangloriar | 70 | 64 | 22 | 34,38% | + | + | 0 | 0 | 0,00% | - | - | 0 | 0 | 0,00% | - | - |
| varar | 27 | 1 | 0 | 0,00% | - | - | 3 | 2 | 66,67% | + | - | 0 | 0 | 0,00% | - | - |
| vergar | 80 | 17 | 7 | 41,18% | + | + | 5 | 1 | 20,00% | + | + | 4 | 2 | 50,00% | + | + |
| vexar | 10 | 0 | 0 | 0,00% | - | + | 3 | 2 | 66,67% | + | + | 0 | 0 | 0,00% | - | + |
| viciar | 132 | 13 | 6 | 46,15% | + | + | 29 | 10 | 34,48% | + | + | 4 | 2 | 50,00% | + | + |
| zangar | 487 | 140 | 98 | 70,00% | + | + | 1 | 1 | 100,00% | + | - | 2 | 1 | 50,00% | + | - |