



HAL
open science

ViPER: A Lexicon-Grammar of European Portuguese Verbs

Jorge Baptista

► **To cite this version:**

Jorge Baptista. ViPER: A Lexicon-Grammar of European Portuguese Verbs. 31e Colloque International sur le Lexique et la Grammaire, Sep 2012, Nové Hradý, Czech Republic. pp.10-16. hal-01414265

HAL Id: hal-01414265

<https://hal.science/hal-01414265>

Submitted on 12 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ViPEr*: A Lexicon-Grammar of European Portuguese Verbs

Jorge Baptista

Univ. Algarve/CECL, Faro, Portugal
INESC-ID Lisboa/L2F, Lisboa, Portugal
jbaptis@ualg.pt

Abstract

This paper presents the current state of ViPEr, the Lexicon-Grammar of European Portuguese verbs, a database with distributional, syntactic and semantic properties of the most frequently occurring verbs. The classification follows the theoretical framework of the Lexicon-Grammar. The paper presents the main linguistic criteria that were adopted in the classification of 5052 frequently occurring verbs, which yield 6,059 different constructions or word senses. The paper concludes with some preliminary results on the application of ViPEr to texts and plans for future work.

1. Introduction

The use of reliable, large-coverage language resources is key to the performance of many Natural Language Processing (NLP) systems. To our knowledge, while some syntactic descriptions of European Portuguese verbs exist, most have not been made publicly available to the NLP community or just consist in human-oriented dictionaries, not having been built for computational processing (FERNANDES 2008, BORBA 1991; BUSSE 1994). On the other hand, partial linguistic studies have produced throughout the last three decades, with major efforts in the late 90s (OLIVEIRA 1981; NASCIMENTO 1997; RODRIGUES 1997), but little, if any, use was made by the NLP community of such data, and very little effort has been addressed to validate or test those, mostly introspective-sourced, and theoretically-oriented, linguistic descriptions. A recent attempt in that direction is that of GOMES (2011), that highlighted the difficulties of the task.

For this project, a practical approach to the lexicon was adopted. For many NLP tasks, but specially for any task where a fine-grained semantic distinction is required of ambiguous lexical forms, being able to identify the meaning of the verb (and of the surrounding elements as well) can be facilitated by the knowledge of the syntactic and semantic constraints the verb imposes on the lexical fulfillment of its argument positions. In particular, the number of verb arguments; their structural and distributional type; the prepositions the verb selects to introduce its essential complements; the main shape-changes that these structures can undergo; and other relevant linguistic information; besides its intrinsic linguistic interest, all this data can be put to use to improve parsing strategies, word-sense disambiguation, question-answer systems, computer-assisted language learning systems, among other applications. Above all, an inventory of basic word senses and their corresponding structures is necessary, and this is the aim of the *ViPEr* project.

This paper presents the current state of the Lexicon-Grammar of European Portuguese verbs. The classification of European Portuguese verb constructions is largely based on the methodology presented by M. GROSS (1975, 1981, 1996) and his collaborators under the Lexicon-Grammar theoretical framework (see LAMIROY (1998) for an overview). The classification proper is directly inspired in the synthesis of LECLÈRE (2002). In the following, the main linguistic criteria that were adopted in the classification are presented. For lack of space, only the most salient classes and properties are presented. The paper concludes with preliminary results from applying this new resource to real texts, and prepares future work.

2. General classification principles

This section presents and briefly discusses the main principles applied to the classification verbal constructions. The classification excludes auxiliary verbs (BAPTISTA *et al.* 2010), support-verbs (BAPTISTA 2005a, RANCHHOD 1990), and operator verbs (M. GROSS 1981). For lack of space, the definition of these verbal constructions is omitted.

The main taxonomical principle is the notion of *simple* (or *elementary*) sentence, which can be defined as the syntactic expression of a semantic predicate. In this sense, adverbial and adjectival (i.e. relative-restrictive) subordinate clauses, as well as coordinated clauses, forming complex sentences, are out of the scope of the classification procedure. The term *simple* (or *elementary*) is to be taken here in the more precise sense of a sentence resulting from the most basic of constraints forming the

* This paper is dedicated to Christian Leclère.

kernel of a language (HARRIS 1991). Hence, verbs with completive arguments are considered second-order operators and are included in the base set of simple sentences.

Each simple sentence is described following the principle of *maximal projection of a predicate*, i.e. including all essential arguments of the predicate. Each semantic predicate is defined by a fixed number of arguments, usually no less than one and no larger than three. Exceptions to this general rule, by virtue of their specialized meaning, are: (i) *meteorological predicates*, forming impersonal constructions e.g. *chover* ‘to rain’; part-of-the-day verbs, e.g. *amanhecer* ‘to dawn’ (class **31I**); and (ii) *object transference predicates*, e.g. *importar* ‘import’ (**38LT**); a few other verbs, such as verb *apostar* ‘to bet’ (**10**), with more than two complements. These are, in a way, exceptional predicates and, as they are not very productive, they may all be defined extensionally.

Obligatory complements are always essential arguments. However, even essential arguments can often be reduced in discourse; therefore, it is difficult to ascertain if a given complement should be considered as essential or circumstantial to that particular predicate. Furthermore, not only some complements expressing so-called circumstances of the predicate should be considered as their essential arguments, e.g. the locative complement of *viver* ‘to live, reside’ (**35LS**) as in *O Pedro vive em Lisboa* ‘Peter lives in Lisbon’ (place); or the manner complement of *portar-se* (to behave’ (**33MV**) in *O Pedro portou-se bem* ‘Peter behaved well’ (manner).

However, some complements are also transformationally derived from the splitting and reanalysis of larger constituents (BAPTISTA 1997, 2000; GUILLET & LECLÈRE 1992; LECLÈRE 1995), e.g. *O cão mordeu as canelas do carteiro* ‘The dog bite the postman heels’ = *O cão mordeu as canelas ao carteiro* ‘The dog bite the heels to-the postman’ = *O cão mordeu o carteiro nas canelas* ‘The dog bite the postman on the heels’ (**32CL**). In these examples, the part-whole (metonymic) relation between the body-part noun (*Nbp*) *canelas* (legs) and the human noun (*Nhum*) *carteiro* (postman) allows for two other alternative syntactic configurations, where the human noun becomes the dative complement or the direct object, while the body-part noun is the direct object or a locative, respectively. In this case, all sentence forms are considered transformationally equivalent, and a single lexicon-syntactic entry has been construed.

In *neutral* constructions, that is, verbs that are diathetically neuter regarding transitivity and intransitivity, e.g. *engordar* ‘to become fat/to put some weight’: (1) *O Pedro engordou* vs. (2) *Os doces engordaram o Pedro* ‘Peter get-fat-ed/the candies get fat-ed Peter’; preference was given in the classification to the transitive structure (class **32C**), for this being the longest structure. Nevertheless, the transitive construction (2) is considered to be derived from a complex structure with causative operator-verb (GROSS 1981), operating on the intransitive sentence form (1), after this being reduced to an infinitive object sub-clause: (1a) *Os doces fizeram o Pedro engordar* ‘The candies made Peter get fat’; these embedding is followed by a subject reversion in the infinitive sub-clause, that moves to a post-verbal position, and its reanalysis as an object of the infinitive verb: (1b) *Os doces fizeram engordar o Pedro* ‘The candies made get fat Peter’; and, finally, by *fusioning* the causative operator-verb with the main, intransitive verb, yielding the (superficial) direct transitive structure of (2): *Os doces engordaram o Pedro* ‘The candies get fat-ed Peter’. In the lexical entry of these verbs, the *N_I V* transformational property yielding the intransitive form (1) was then explicitly encoded.

Finally, for intrinsic reflexive constructions, like *suicidar-se* ‘to commit suicide’ (**31H**), e.g. *O Pedro suicidou-se* ‘Peter suicide himself = commit suicide’, the reflex pronoun is treated as part of the verb and not as an autonomous constituent, as it can not be zeroed: **O Pedro suicidou* ‘Peter suicide’, nor replaced by a distributionally free constituent: *O Pedro suicidou o João* ‘Peter suicide John’.

This last example could be said ironically, but such discursive phenomena, that is, productive, figurative uses, were disregarded in the classification, since they correspond to the expression of the speakers’ creativity, and not to the language basic structure from which they are drawn. On the other hand, conventional figurative uses may give rise to splitting a verb into two lexical entries, as with *ralar* ‘grate/worry’ e.g. *O Pedro ralou o queijo* ‘Peter grated the cheese’ (**32C**) vs. *Isso ralava o Pedro* ‘That worried Peter’ (**4**).

Since there is no general rule to define the maximal projection of a given predicate, and because the semantic and syntactic complexity of the language structure is such, often only through a case-by-case decision had one to proceed. The observation of extensive data from corpora was key to help deciding the most adequate solution for each case.

The classification of verbal constructions may be structured in two layers, based on the structural complexity and semantic compositionality of the sentences: (i) *completive constructions*, i.e. simple sentences whose verb selects at least one sentential argument (a completive, subordinate clause; these have precedence over (ii) *non-completive constructions*, i.e. simple sentences whose verb selects only noun phrases as their arguments. As far as subordinate clauses are concerned, only completive clauses are here considered. All adverbial and adjectival (relative) subordinate clauses are excluded. To make matters more complex, though, completive subordinate clauses are often replaced by complex noun phrases built around predicative nouns, functioning in as much the same way as a completive: *Que o Pedro tenha vindo/a vinda do Pedro entristeceu o João* ‘That Peter has come/the coming of Peter has saddened John’ (4). In some cases, the completive is rarely observed in corpora, but if a propositional content can be semantically defined for a given argument position, which is usually reflected in the predicative nature of the nouns selected for that position, that particular construction has been included in completive constructions, even if only these complex noun phrases are observed in that position. Non-completive constructions, therefore, exclusively present nominal-arguments, whose head noun cannot be a predicative noun.

The preposition introducing the complement(s) is an important taxonomical criterion, since each prepositional construction usually presents only one, invariable, preposition for a given argument position, e.g. *telefonar a* ‘phone to’ (33; RODRIGUES 1997), *gostar de* ‘like of’ (8), *confiar em* ‘trust in’ (35R), *casar com* ‘marry with’ (35S), *ansiar por* ‘expect’ (8). Basic prepositions in European Portuguese are *a* ‘to’, *com* ‘with’, *de* ‘of/from’ and *em* ‘in/at’; locative prepositions have been collapsed under the notation *Loc*; all other prepositions, v.g. *para* ‘to/towards’, *por* ‘by/for’, *sobre* ‘about/on/over/upon’, etc., are collapsed under the same taxon, though they have always been explicitly encoded in the lexical entry. Some constructions present, however, more than one preposition introducing its complement positions, e.g. *lutar com/contra* ‘fight with/against’ (35R): *O Pedro luta contra/com a injustiça* ‘Peter fights against/with the injustice’. If the meaning of the sentence cannot be clearly distinguished, nor the distributional constraints differ significantly depending on this preposition variation, a single lexical-entry has been construed, using the basic preposition (in this case, *com* ‘with’) for its classification and explicitly registering the other variants in the entry’s syntactic properties.

A special case of prepositional construction consists of *symmetric* constructions (BAPTISTA 2005), usually presenting a complement introduced by *com* ‘with’, e.g. *O Pedro conversou com o João* ‘Peter talked with John’ (35S); *O Pedro confundiu o João com o Paulo* ‘Peter mistook John for Paul’ (36S1); *O Pedro debateu esse assunto com o João* ‘Peter debated this topic with John’ (36S2); and *O Pedro concordou com o João em fazerem isso* ‘Peter agreed with John in doing that’ (42S). Because this set of constructions is also defined by particular syntactic-semantic and transformational properties, it had to be singled out from other prepositional constructions.

A rarer case of preposition alternation consist of constructions where the verb presents not only a direct object structure but also a prepositional one, while the meaning and the distribution do not change: *O Pedro namora a Sara* ‘Peter is dating Sara’ (32H) = *O Pedro namora com a Sara* ‘Peter is dating with Sara’ (35S). In this case, the longer, prepositional construction is given precedence over the direct-transitive construction. The transitive construction is noted by the property [pcz], corresponding to the zeroing of the complement preposition.

The distributional constraints imposed by the verb on its arguments positions (subject and essential complements) are the next major criterion for verb classification. The main distributional classes here considered are:

- *Nhum/Nnhum*: Human/non-human noun opposition. This property is only tested with proper names; animal denoting nouns are excluded, except for specialized verbs such as *animal voices*, e.g. *bramir* ‘roar’ (exclusively applied to elephants; 31R). Verb constructions with strictly human arguments (31H, 32H) have priority over those that allow both human and non human (31R, 32x).
- *Npl*: plural noun. This abstract notion describes the distributional constraints of verbs imposing a conceptual plural in a given argument position (NASCIMENTO 1997), e.g. the subject of *abundar* ‘abound’ (31PL), or the object of *dispersar* ‘disperse’ and *coleccionar* ‘collect’ (32PL).

- *Nbp*: body-part nouns. Nouns designating a body-part, that is, nouns that intrinsically imply a metonymic relation, usually with a human noun in the sentence, and hence allow for several types of sentence restructuring (see examples above; BAPTISTA 1997, 2000).
- *Nloc*: locative nouns, i.e. nouns designating places or locations. Locative verbs form a complex syntactic-semantic system and a wide range of syntactic constructions can express the predicates they denote, depending on whether the verb can be defined as a *stative* or *dynamic* predicate, or the particular orientation that can be accorded in the case of *dynamic* verbs. Hence, one considers *source-* and *destination-*oriented dynamic verbs; more rarely, a *traject-*oriented predicate can be considered; e.g. *O Pedro vive em Lisboa* ‘Peter lives in Lisbon’ (stative; **35LS**); *O Pedro vai a/para Lisboa* ‘Peter goes to Lisbon’ (dynamic, destination-oriented; **35LD**); *O Pedro veio de Lisboa* ‘Peter came from Lisbon (source-oriented; **35LD**); and *O Pedro passou por Lisboa* ‘Peter passed through Lisbon’ (traject-oriented; **35LD**). Locatives can appear not only in prepositional phrases, as in the examples above, but they can also appear as noun phrases, usually in object position, e.g. *O Pedro atravessou o pátio* ‘Peter crossed the yard’ (dynamic, traject-oriented; **38L1**). *Locative-Fused* verbs can be considered cases of *Fusion* (M. GROSS 1981): for example, an object is fused with verbs like *pôr* ‘put’, *deitar* ‘throw’ or *meter* ‘insert’ and its locative prepositional complement becomes the direct object of the resulting verb in *O Pedro envenenou o vinho* ‘Peter poisoned the wine’ = *O Pedro deitou/meteu/pôs veneno no vinho* ‘Peter put poison in/on the wine’ (destination-oriented; **38L4**); in another class, the verb is fused with the locative: *O Pedro enjaulou o leão* ‘Peter caged the lion’ = *O Pedro meteui/pôs o leão na jaula* ‘Peter put the lion in the cage’ (destination-oriented; **38L2**).
- *R*: Constraint direct object constructions: Some verb constructions, without being frozen or idiomatic (BAPTISTA *et al.* 2004; VALE 2008), present such narrow distributional constraints that the general distributional properties stated above fail to capture the precise choice of words for a given syntactic position. For the most part, these are direct transitive constructions where the object must be selected from a very limited word set: *O Pedro estrelou uns ovos* ‘Peter fried some eggs’ (**32R**). In this example, only *ovo* can fill the object distributional constraints; the overall meaning is clearly compositional; otherwise, number and determiner variation is free. These type of distributional constraint is harder to discover and to describe, so they constitute residual, though large, lexical classes (DIAS *et al.* 2006, LECLÈRE 2002).

Transformational properties are considered, in order to distinguish structurally similar constructions that correspond to different types, semantically homogenous, sets of predicates. Another use of these properties is to allow for the constitution of amenable-sized verb classes, so that if a set becomes too large (usually over 200) it is then advisable to split it in two, more treatable subsets. Here, only some of these operations are presented:

- *Passive*: For example, the large set of direct transitive verbs with measurement nouns allow the constitution of a highly homogenous subclass of *measurement* verbs, since none of the elements of this set allows the *Passive* transformation: *As batatas pesam três quilos* ‘The potatoes weight 3 Kg’ vs. **Três quilos são pesados pelas batatas* ‘3Kg are weighed by the potatoes’ (**32NM**).
- *Fusion* of causative operator verb and adjectival sentences. In another example, the de-adjectival causative verbs consist of a set of direct-transitive verbal constructions that is distinguished by systematically having paraphrases with adjectival basic sentences under a causative-operator verb e.g. *clarificar* = *tornar (mais) claro* ‘clarify, make clear(er)’: *O Pedro clarificou a sua posição* ‘Peter clarified his position’ = *O Pedro tornou (mais) clara a sua posição* ‘Peter made his position clear(er)’ (**32TA**).

Finally, and as it is clear from the above remarks, the meaning of the verb in the simple sentence that constitutes its construction is always present during its classification, and it is a *datum* that the operations the sentence may undergo do not alter. Therefore, one cannot speak of syntactic classification alone, since meaning is always implied. The use of semantic concepts as declarative propositions in formal classification, however, has been avoided, and whenever needed, only as a last resource; these propositions can only be adopted when the intuitions about the meaning involved are highly reproducible. Usually, these intuitions should be backed by clear syntactic properties, as is the case of symmetric constructions (BAPTISTA 2005b), where the special relation between the verb and

two of its arguments determines a set of syntactic properties, which are unique to that particular class of predicates.

Only seldom was any semantic concept used so far in the classification here proposed. The concept of **apparition** serves to distinguish the direct-transitive constructions with concrete objects where the object appears after/during the process, e.g. *construir* ‘build’, class **32A**; as opposed to the predicates where the object preexists the process e.g. *sublinhar* ‘underline’, class **32C**. (in this later case, the object usually it undergoes some transformation/manipulation). Other properties may have to be devised, especially for the large and very heterogeneous **32C** class.

3. Results and future work

A selection of verbs taken from the CETEMPúblico, a large-sized (approx. 190 million words), journalistic text, European Portuguese corpus (ROCHA & SANTOS 2000) underwent the classification process. The corpus was processed through the STRING NLP chain (MAMEDE *et al.* 2012)² for part-of-speech tagging and disambiguation and for verbal chains parsing. The most frequent verbs (frequency above 5 instances in the corpus, in descending order) were studied to determine their basic constructions using the criteria briefly explained above. From an initial list of about 7,000 different verb lemmas, 5,052 have already been classified, yielding 6,059 different lexicon-syntactic entries (or clear-cut verb senses). The classification of the remaining verbs is still on going. While not considered in the classification, 259 support and operator-verbs’ constructions (M. Gross 1981) have also been identified. Table 1 presented the breakdown of the ambiguous verbs (ws=word senses):

Table 1. Number of different word senses per lemma

ws	lemmas	ws	lemmas	ws	lemmas	ws	lemmas
1	4293	3	145	5	9	7	3
2	565	4	41	6	8	8	2

Fifty-five lexicon-syntactic classes were established so far (see Appendix). Most of these classes correspond to the French Lexicon-Grammar tables (LECLÈRE 2002) and the same conventional code was adopted for easier comparison. For lack of space these cannot be presented here in full (see BAPTISTA 2012 for a complete overview). As one can see, most verbs have been encoded for only one sense. A very conservative approach was adopted here, describing, at this stage, only the most common word senses, thus the still small number of duplicates. For example, the verb *apontar* ‘to point/aim/take note/indicate/signal’ yields 4 different lexicon-grammar entries: (**36DT**) *O bandido apontou uma faca ao polícia* ‘The bandit aimed a knife at the policeman’, (**38LD**) *O Pedro apontou os números premiados num papel* ‘Peter took note of the lotary numbers in a piece of paper’, (**39**) *O Pedro apontou o João como sério candidato ao prémio* ‘Peter indicated John as a serious candidate’ and (**9**) *O Pedro apontou ao João quais os defeitos que devia corrigir* ‘Peter signaled to John which issues he should correct’. For each entry, the structural, distributional, semantic and transformational properties were encoded in a single binary matrix, and an illustrative example was provided.

In order to do a preliminary assessment of ViPEr, two small articles were retrieved from the online daily edition of *Público*, with about 1,000 words. From the 85 verb forms, 41 correspond to auxiliaries (modal, temporal and aspectual verb auxiliaries; support and operator verbs; and copula verbs). The output of the STRING system looks like the following sentences:

"A Europa deve{**dever(VMOD)**} *cumprir*{**cumprir(05,35R,32R#)**} os acordos com a maior celeridade possível. Espero{**esperar(06,35R#)**} que a Europa *esteja*{**estar(VSUP#)**} a_a altura de_as circunstâncias", *afirmou*{**afirmar(31H,09#)**}.

‘Europe must fulfill the agreements as fast as possible. I hope that Europe is up to the job, he said.’

Notice the correct parsing of the modal auxiliary *dever* ‘must’ (BAPTISTA *et al.* 2010) and the support verb *estar à altura de* (RANCHHOD 1990). In most cases, at least one the verb classes encoded in ViPEr match the actual uses found in these texts. Thus, *cumprir* ‘fulfil’ corresponds to the **32R** entry, *esperar* ‘hope’ is an instance of the **06** use, and *afirmar* is the **09** *verbum dicendi* (Baptista). However, for the remaining 43 verbs, even in such a small sample, it was possible to find lexical lacuna and many other problems:

As médias totais têm{**ter(VOPL#)**} em conta os resultados de os alunos internos – aqueles que frequentam{**frequentar(38L1#)**} as aulas até a o final de o ano lectivo e

² <https://string.l2f.inesc-id.pt> [22/07/2012].

vão{ir(35LD#)} a exame com uma classificação interna igual ou superior a 10 - e as de os externos , que anularam{anular(32TA#)} a matrícula e se autopropuseram{autopropuserir(No viper data)} a exame .

‘The overall averages take into account the results of internal students - those who attend classes until the end of the school year and take a final exam with an internal grade equal to or greater than 10 - and the results of the external students, which cancelled their registration and presented themselves to exam.’

Thus, *ter em conta* and *ir a exame* may be considered frozen sentences (classes **CNP2** and **CP1**, respectively; BAPTISTA *et al.* 2004); this type of structure has not been integrated in the system yet. The expression *anular matrícula* can be considered a support verb construction, but this use had not been identified yet. The only verb for which there is no entry in ViPER is also an unknown, derived word, formed on the base verb *propor* ‘propose’ with prefix *auto-* ‘self’. An extension of the morphologic module LexMan (Diniz & Mamede 2011) of the STRING system will enable in the near future (FREITAS 2012).

In future work, the lexical coverage of ViPER needs to be further assessed. All the verbs of a 290K words, POS-disambiguated corpus (RIBEIRO 2004, DINIZ & MAMEDE 2011) has been annotated with the ViPER classes and is currently being manually reviewed. Verb forms associated to 136 lemmas had not a ViPER tag. Many of them correspond to lemmas that were changed by the new Portuguese, orthographic reform, but only 46 had to be added to the database. This corpus, manually annotated with the verb senses, will enable the use and evaluation of rule-based and machine-learning techniques in word-sense disambiguation tasks (Travanca 2012).

Acknowledgments

Research for this paper was partially funded by project REAP.PT (CMU-Portugal/HuMach/0053/2008). The author wishes to thanks, in chronological order, the students who, in different ways, have contributed to this work: David Monteiro (UNL), Rui Santos (UALG) to Tiago Travanca (IST) and Cláudio Diniz (INESC-ID Lisboa) for their help; and, finally, to Nuno Mamede (IST/INESC-ID Lisboa), for his constant support and friendship.

References

- BAPTISTA, Jorge. 1997. Conversão, nomes parte-do-corpo e reestruturação dativa. in Castro, Ivo (ed.) *Actas do XII Encontro da APL*, Lisboa, 1996, pp. 51-59. Lisboa: APL/Colibri.
- BAPTISTA, Jorge. 2000. Boby-part nouns and local grammars. *Révue d'Informatique et Statistiques en Sciences Humaines*, 36, pp. 53-66. Liège: Univ. Liège.
- BAPTISTA, Jorge; CORREIA, Anabela; FERNANDES, Graça. 2004. Frozen Sentences of Portuguese: Formal Descriptions for NLP. Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain, July 26, 2004, pp.72-79. Barcelona: ACL.
- BAPTISTA, Jorge. 2005a. *Sintaxe dos Predicados Nominais com SER DE*. Lisboa: FCT/FCG.
- BAPTISTA, Jorge. 2005b. Construções simétricas: argumentos e complementos. in Figueiredo, O; and Rio-Torto, Graça; and Silva, F. (eds.). *Estudos de Homenagem a Mário Vilela*, pp. 353-367. Porto: Campo das Letras.
- Baptista, Jorge. 2010. *Verba dicendi*: a structure looking for verbs. in Nakamura, Takuya; Laporte, Éric; Dister, Anne; Fairon, Cédric (eds.). *Les Tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*. Cahiers du CENTAL 6, pp. 11-20. Louvain-la-Neuve: CENTAL.
- BAPTISTA, Jorge; MAMEDE, Nuno; GOMES, Fernando. 2010. Auxiliary verbs and verbal chains in European Portuguese. *Computational Processing of the Portuguese Language*. LNAI/LNCS 6001, pp.
- BAPTISTA, Jorge. 2012. *Verb Classification Guidelines*. Technical Report. Lisboa: L2F-Spoken Language Lab/INESC-ID Lisboa.
- BOONS, Jean-Paul; GUILLET, Alain & LECLÈRE, Christian. 1976a. *La structure des phrases simples en français: Les constructions intransitives*. Paris: Droz.
- BOONS, Jean-Paul; GUILLET, Alain & LECLÈRE, Christian. 1976b. *La structure des phrases simples en français: Les constructions transitives*. Rapport de Recherches 10. Paris: LADL.
- BORBA, Francisco. 1991. *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*. São Paulo, Brasil: UNESP.
- BUSSE, Winfried. 1994. *Dicionário Sintático de Verbos*. Coimbra: Livraria Almedina.
- CHABY, Teresa. 1997. *Construções de segmentação. Propriedades léxico-sintáticas*. MA Thesis). Universidade de Lisboa - Faculdade de Letras.

- DIAS, Maria Carmelita P.; LAPORTE, Éric & LECLÈRE, Christian. 2006. Verbs with very strictly selected complements. In: *Collocations and Idioms: The First Nordic Conference on Syntactic Freezes*. Finland: University of Joensuu.
- DINIZ, Cláudio; MAMEDE, Nuno. 2011. *LexMan - Lexical Morphological Analyser*. Technical Report. Lisboa: L2F/INESC-ID Lisboa.
- FERNANDES, Francisco. 2008. *Dicionário de Verbos e Regimes* (45th ed). São Paulo, Brasil: Globo.
- GOMES, Fernando. 2011. *Validation of Lexical-Syntactical Matrices*. MSc Dissertation. Instituto Superior Técnico, Universidade Técnica de Lisboa.
- GROSS, Gaston. 1996a. *Les expressions figées en français*. Paris: Ophrys.
- GROSS, Maurice. 1975. *Méthodes en Syntaxe*. Paris: Hermann.
- GROSS, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63, 7–52.
- GROSS, Maurice. 1982. Une classification des phrases «figées» du français. *Revue Québécoise de Linguistique*, 12-2, pp–pp.16
- GROSS, Maurice. 1996b. Lexicon-Grammar. in Brown, K. and Miller, J. (eds.), *Concise Encyclopedia of Syntactic Theories*. Cambridge: Pergamon, pp. 244–259.
- GUILLET, Alain & LECLÈRE, Christian. 1981. Restructuration deu group nominal. *Langages* 63, 99–125.
- GUILLET, Alain & LECLÈRE, Christian. 1992. *La structure des phrases simples en français: 2 - Les constructions transitives locatives*. Paris: Droz.
- HARRIS, Zellig S. 1991. *A Theory of Language and Information. A Mathematical Approach*. Oxford: Clarendon.
- LAMIROY, Béatrice (ed.) 1998. *Le Lexique-Grammaire. Travaux de Linguistique*, Vol. 37. Duculot.
- LECLÈRE, Christian. 1995. Sur une restructuration dative. *Language Research*, 179–198.
- LECLÈRE, Christian. 2002. Organization of the Lexicon-Grammar of French Verbs. *Linguisticae Investigationes*, 25-1, 29–48.
- MAMEDE, Nuno; BAPTISTA, Jorge; DINIZ, Cláudio. 2012. STRING - An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. Proceedings of PROPOR'2012, Coimbra, Portugal. <http://www.propor2012.org/demos/DemoSTRING.pdf> (on-line)
- OLIVEIRA, Maria Elisa. 1981. *Syntaxe des verbes psychologiques du Portugais*. Lisboa: INIC.
- RIBEIRO, Ricardo. 2004. *Anotação Morfosintáctica Desambiguada do Português* (MSc Thesis). Lisboa: IST/UTL.
- RANCHHOD, Elisabete. 1990. *Sintaxe dos predicados nominais com ESTAR*. Lisboa: INIC.
- ROCHA, Paulo & SANTOS, Diana. 2000 (November). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR'2000*, pp. 131–140.
- RODRIGUES, Rosinda. 1997. *Determinação das propriedades sintáticas das frases simples com a forma N0 V a N1*. Master thesis, Universidade de Lisboa - Faculdade de Letras.
- TRAVANCA, Tiago. 2012. *Disambiguation of Verb Senses*. MA Project. Lisboa: IST.
- VALE, Oto. 2001. *Expressões Cristalizadas do Português do Brasil: Uma Proposta de Tipologia*. Araraquara: UEJMF(PhD Thesis).

Appendix. ViPer : Verb classes of European Portuguese.

Class	Count	Structure	verb	example
01I	13	QueF ₀ V	adiantar 'matter'	<i>Não adianta fazer isso</i> 'doesn't matter to do that'
01T	61	QueF ₀ V QueF ₁	evitar 'avoid'	<i>Fazer isso evitava ter de fazer aquilo</i> 'To do this avoids having to do that'
02	15	QueF ₀ V Prep ₁ QueF ₁	obrigar 'force'	<i>Isto obriga a fazer aquilo</i> 'This forces to do that'
03	1	N ₀ V N ₁ (Loc ₁ Nloc ₁) Vinf ^f	mandar 'send'	<i>O Pedro mandou o João à loja comprar café</i> 'Peter sent John to the shop to buy some coffee'
04	330	Nnr ₀ V Nhum ₁	irritar 'irritate'	<i>Isso irrita o Pedro</i> 'That irritates Peter'
05	24	Nnr ₀ V a Nhum ₁	agradar 'please'	<i>Isso agrada ao Pedro</i> 'That pleases Peter'
06	221	Nhum ₀ V QueF ₁	pensar 'think'	<i>O Pedro pensa que o João é inteligente</i> 'Peter thinks that John is intelligent'
07	46	Nhum ₀ V a (Vinf ^f) ₁	aprender 'learn'	<i>O Pedro aprendeu a fazer isso</i> 'Peter learn to do that'
08	96	N ₀ V Prep ₁ QueF ₁	depender 'depend'	<i>O Pedro dependia da autorização do João</i> 'Peter depended on John's authorization'
09	162	Nhum ₀ V QueF ₁ a Nhum ₂	dizer 'say'	<i>O Pedro disse ao João que está feliz</i> 'Peter said to John that [he] is happy'
10	4	Nhum ₀ V QueF ₁ Prep ₂ Nhum ₂	apostar 'bet'	<i>O Pedro apostou com o João que ganhava a corrida</i> 'Peter bet with John that [he] would win the race'
11	43	N ₀ V N ₁ a QueF ₂	obrigar 'force'	<i>O Pedro obrigou o João a fazer isso</i> 'Peter forced John to do that'
12	34	N ₀ V N ₁ de ₂ (Vinf ^f) ₂	impedir 'prevent'	<i>O Pedro impediu o João de fazer isso</i> 'Peter prevented John from doing that'

Class	Count	Structure	verb	example
13	21	$N_0 V N_1 de_2 QueF_2$	<i>informar</i> 'inform'	<i>O Pedro informou o João de que ia fazer isso</i> 'Peter informed John that [he] was going to do that'
14	11	$N_0 V N_1 Prep_2 N_2, (a Nhum_3)$	<i>pagar</i> 'pay'	<i>O Pedro pagou 20€ por isso ao João</i> 'Peter paid 20€ to John for that'
16	9	$N_0 V QueF_1 Prep_2 QueF_2$	<i>deduzir</i> 'deduce'	<i>O Pedro deduziu isso daquilo</i> 'Peter deduced this from that'
31CL	20	$Nbp_0 V$	<i>suar</i> 'sweat'	<i>Os pés do Pedro suam</i> 'Peter's feet sweat'
31H	287	$Nhum_0 V$	<i>espirrar</i> 'sneeze'	<i>O Pedro espirrou</i> 'Peter's sneezed'
31I	24	$0 V$	<i>chover</i> 'rain'	<i>Chove</i> '[it] rains'
31PL	11	$Npl_0 V$	<i>proliferar</i> 'proliferate'	<i>As bactérias proliferam</i> 'Bacteria proliferate'
31R	275	$(Nhum+Nnhum)_0 V$	<i>morrer</i> 'die'	<i>O Pedro morreu</i> 'Peter died'
32A	35	$N_0 V Nnhum_1$ {apparition}	<i>preparar</i> 'prepare'	<i>O Pedro preparou o almoço</i> 'Peter prepared lunch'
32C	1,148	$N_0 V Nnhum_1$ *{apparition}	<i>ler</i> 'read'	<i>O Pedro leu um livro</i> 'Peter read a book'
32CL	193	$N_0 V Nbp_1$	<i>partir</i> 'break'	<i>O Pedro partiu um braço</i> 'Peter broke an arm'
32CV	13	$N_0 N-v N_1$ [= converter N_1 em N]	<i>cristalizar</i> 'crystallize'	<i>Isso cristalizou o açúcar</i> 'That crystallized the sugar'
32H	440	$Nhum_0 V Nhum_1$	<i>amar</i> 'love'	<i>O Pedro ama a Ana</i> 'Peter loves Ana'
32NM	29	$N_0 V Nmeas_1$	<i>medir</i> 'measure'	<i>O Pedro mede 1,80 m</i> 'Peter measures 1.80 m'
32PL	55	$N_0 V Npl_1$	<i>ordenar</i> 'order'	<i>O Pedro ordenou os alunos</i> 'Peter ordered the students'
32R	243	$N_0 V Nc_1$	<i>estrelar</i> 'fry'	<i>O Pedro estrelou um ovo</i> 'Peter fried an egg'
32TA	289	$N_0 Adj-v N_1$ [V=tornar $Adj N_1$]	<i>amaciar</i> 'soften'	<i>O sabonete amacia a pele</i> 'The soap softens the skin'
33	77	$N_0 V a N_1$	<i>telefonar</i> 'phone'	<i>O Pedro telefonou ao João</i> 'Peter phoned to John'
33MV	4	$N_0 V Advmanner$	<i>portar-se</i> 'behave'	<i>O Pedro portou-se mal</i> 'Peter behaved badly'
33NM	1	$N_0 V Prep_1 Nmeas_1$	<i>ascender</i> 'ascend'	<i>O PIB ascende a 1B\$</i> 'example'
34	53	$N_0 V Prep_1 N_1 Prep_2 N_2$	<i>saber</i> 'know'	<i>O Pedro sabe muito de futebol</i> 'Peter knows a lot about football'
35LD	273	$N_0 V-dyn Loc_1 Nloc_1$	<i>entrar</i> 'enter'	<i>O Pedro entrou na sala</i> 'Peter entered into the room'
35LS	40	$N_0 V-stat Loc_1 Nloc_1$	<i>viver</i> 'live'	<i>O Pedro vive em Lisboa</i> 'Peter lives in Lisbon'
35R	173	$N_0 V Prep_1 N_1$	<i>confiar</i> 'trust'	<i>O Pedro confia no João</i> 'Peter trusts in John'
35S	112	$N_0 V com N_1$	<i>conversar</i> 'talk'	<i>O Pedro conversou com o João</i> 'Peter talked with John'
36DT	108	$Nhum_0 V Nobj_1 a Nhum_2$	<i>dar</i> 'give'	<i>O Pedro deu um livro ao João</i> 'Peter gave a book to John'
36R	89	$N_0 V N_1 Prep_2 N_2$	<i>transformar</i> 'transform'	<i>O Pedro transforma barro em arte</i> 'Peter transforms clay in art'
36S1	82	$N_0 V Nobj_1 com_2 Nobj_2$	<i>misturar</i> 'mix'	<i>O Pedro mistura o açúcar com a farinha</i> 'Peter mixes sugar with flour'
36S2	15	$Nhum_0 V Nobj_1 com_2 Nhum_2$	<i>combinar</i> 'arrange'	<i>O Pedro combinou com o João uma ida ao cinema</i> 'Peter arranged with John to go to the movies'
36TA	6	$N_0 Adj-v N_1 Prep_2 N_2$ [V=tornar $Adj N_1$]	<i>adequar</i> 'adjust'	<i>O Pedro adequa o discurso ao público</i> 'Peter adjusts his talk to the public'
38L1	193	$N_0 V Nloc_1$	<i>invadir</i> 'invade'	<i>O Pedro invadiu a sala</i> 'Peter invaded the room'
38L2	35	$N_0 Nloc-v Nobj_1$ [V=put in $Nloc$]	<i>enjaular</i> 'cage'	<i>O Pedro enjaulou o leão</i> 'Peter caged the lion'
38L3	9	$Nloc_0 V Nobj_1$	<i>encerrar</i> 'enclose'	<i>A jaula encerrava a fera</i> 'The cage enclosed the beast'
38L4	86	$N_0 Nobj-v Nloc-d_1$ [V=pôr $Nobj$]	<i>envenenar</i> 'poison'	<i>O Pedro envenenou a bebida</i> 'Peter poisoned the drink'
38L5	10	$N_0 Nobj-v Nloc-s_1$ [V=tirar $Nobj$]	<i>desengordurar</i> 'ungrease'	<i>O Pedro desengordurou o prato</i> 'Peter ungreased the dish'
38LD	73	$N_0 Vdyn N_1 Loc-d_2 Nloc_2$	<i>pousar</i> 'put'	<i>O Pedro pousou o livro na mesa</i> 'Peter put the book on the table'
38LS	73	$N_0 Vdyn N_1 Loc-s_2 Nloc_2$	<i>retirar</i> 'remove'	<i>O Pedro retirou o livro da mesa</i> 'Peter removed the book from the table'
38LT	41	$N_0 Vdyn N_1 Loc-s_2 Nloc_2 Loc-d_3 Nloc_3$	<i>transferir</i> 'transfer'	<i>O Pedro transferiu o livro daqui para ali</i> 'Peter transferred the book from here to there'
38LR	5	$N_0 Vstat N_1 Prep_2 N_2$	<i>situar</i> 'place'	<i>O Pedro situou a casa no mapa</i> 'Peter placed the house in the map'
38PL	57	$N_0 V N_1 Prep_2 Npl_2$	<i>dividir</i> 'divide'	<i>O Pedro dividiu o bolo em fatias</i> 'Peter divided the cake in tranches'
38R	7	$N_0 V N_1 Loc-d_2 N_2$	<i>remeter</i> 'send'	<i>O Pedro remeteu o João para a Ana</i> 'Peter sent John to Anna'
38TD	9	$N_0 V N_1 Loc-s_2 N_2$	<i>receber</i> 'receive'	<i>O Pedro recebeu uma prenda do João</i> 'Peter received a gift from John'
39	58	$N_0 V N_1 (Prep_2) N_2$	<i>nomear</i> 'appoint'	<i>O Pedro nomeou o João (como) seu representante</i> 'Peter appointed John (as) his representative'
40	11	$N_0 V Prep_1 N_1 Prep_2 N_2$	<i>dar</i> 'hit'	<i>O Pedro deu com um livro na cabeça do João</i> 'Peter hit with a book on John's head'
41	11	$N_0 V Prep_1 N_1 Prep_2 QueF_2$	<i>apelar</i> 'appeal'	<i>O Pedro apelou ao João para que fizesse isso</i> 'Peter appealed to John for [him] to do that'
42S	5	$N_0 V com N_1 Prep_2 N_2$	<i>comungar</i> 'commune'	<i>O Pedro comungava com o João dos mesmos ideais</i> 'Peter communed with John from the same ideals'
Total: 6,059				

Notations

N_0, N_1, N_2, N_3 : subject and complements; *Prep*: preposition; *Adj*: adjective; *Adv*: adverb; *Nhum*: human noun; *Nnhum*: non-human nouns; *Nbp*: body-part noun; *Npl*: plural noun; *Nloc*: locative noun; *Nnr*: non-constraint noun; *Nobj*: "object" noun (semantic role); *QueF*: completive sub-clause; *Loc*: locative preposition; *V*: verb; *Vdyn*: dynamic locative verb; *Vstat*: static locative verb; *-v*: verb ending.