Identifying and Analyzing Brazilian Portuguese Complex Predicates

Magali Sanches Duran[♡] **Carlos Ramisch**[♠ ◇] **Sandra Maria Aluísio**[♡] **Aline Villavicencio**[♠]
[♡] Center of Computational Linguistics (NILC), ICMC, University of São Paulo, Brazil
[♠] Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
[◇] GETALP – LIG, University of Grenoble, France
magali.duran@uol.com.br ceramisch@inf.ufrgs.br
sandra@icmc.usp.br avillavicencio@inf.ufrgs.br

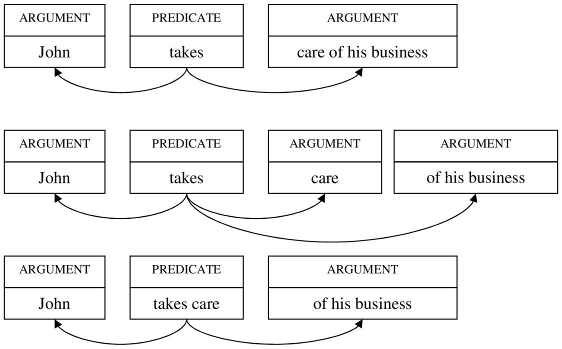
Abstract

Semantic Role Labeling annotation task depends on the correct identification of predicates, before identifying arguments and assigning them role labels. However, most predicates are not constituted only by a verb: they constitute Complex Predicates (CPs) not yet available in a computational lexicon. In order to create a dictionary of CPs, this study employs a corpus-based methodology. Searches are guided by POS tags instead of a limited list of verbs or nouns, in contrast to similar studies. Results include (but are not limited to) light and support verb constructions. These CPs are classified into idiomatic and less idiomatic. This paper presents an in-depth analysis of this phenomenon, as well as an original resource containing a set of 773 annotated expressions. Both constitute an original and rich contribution for NLP tools in Brazilian Portuguese that perform tasks involving semantics.

1 Introduction

Semantic Role Labeling (SRL), independently of the approach adopted, comprehends two steps before the assignment of role labels: (a) the delimitation of argument takers and (b) the delimitation of arguments. If the argument taker is not correctly identified, the argument identification will propagate the error and SRL will fail. Argument takers are predicates, frequently represented only by a verb and occasionally by Complex Predicates (CPs), that is, "predicates which are multi-headed: they are composed of more than one grammatical element" (Alsina et al., 1997, p. 1), like *give a try, take care, take a shower*. In SRL, the verbal phrases (VPs) identified by a parser are usually used to automatically identify argument takers, but do no suffice. A lexicon of CPs, as well as the knowledge about verbal chains composition, would complete a fully automatic identification of argument takers. Consequently, the possibility of disagreement between SRL annotators would rely only on the assignment of role labels to arguments. This paper reports the investigation of such multi-word units, in order to meet the needs arisen from an SRL annotation task in a corpus of Brazilian Portuguese[1].

To stress the importance of these CPs for SRL, consider the sentence *John takes care of his business* in three alternatives of annotation:



The first annotation shows *care of his business* as a unique argument, masking the fact that this segment is constituted of a predicative noun, *care*, and its internal argument, *of his business*. The second annotation shows *care* and *of his business* as arguments of *take*, which is incorrect because *of his business* is clearly an argument of *care*. The third annotation is the best for SRL purposes: as a unique predicate — *take care*, *take* shares its external argu-

[1] CPs constituted by verbal chains (e.g. *have been working*) are not focused here.

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 74–82,
Portland, Oregon, USA, 23 June 2011. ©2011 Association for Computational Linguistics

ment with *care* and *care* shares its internal argument with *take*.

The goal of this paper is twofold: first, we briefly describe our computer-aided corpus-based method used to build a comprehensive machine-readable dictionary of such expressions. Second and most important, we analyze these expressions and their behavior in order to shed some light on the most adequate lexical representation for further integration of our resource into an SRL annotation task. The result is a database of 773 annotated CPs, that can be used to inform SRL and other NLP applications.

In this study we classify CPs into two groups: idiomatic CPs and less idiomatic CPs. Idiomatic CPs are those whose sense may not be inferred from their parts. Examples in Portuguese are *fazer questão (make a point), ir embora (go away), dar o fora (get out), tomar conta (take care), dar para trás (give up), dar de ombros (shrug), passar mal (get sick)*. On the other hand, we use "less idiomatic CPs" to refer to those CPs that vary in a continuum of different levels of compositionality, from fully compositional to semi-compositional sense, that is, at least one of their lexical components may be litterally understood and/or translated. Examples of less idiomatic CPs in Portuguese are: *dar instrução (give instructions), fazer menção (make mention), tomar banho (take a shower), tirar foto (take a photo), entrar em depressão (get depressed), ficar triste (become sad)*.

Less idiomatic CPs headed by a predicative noun have been called in the literature "light verb constructions" (LVC) or "support verb constructions" (SVC). Although both terms have been employed as synonyms, "light verb" is, in fact, a semantic concept and "support verb" is a syntactic concept. The term "light verb" is attributed to Jespersen (1965) and the term "support verb" was already used by Gross in 1981. A light verb is the use of a polysemous verb in a non prototypical sense or "with a subset of their [its] full semantic features", North (2005). On the other hand, a support verb is the verb that combines with a noun to enable it to fully predicate, given that some nouns and adjectives may evoke internal arguments, but need to be associated with a verb to evoke the external argument, that is, the subject. As the function of support verb is almost always performed by a light verb, attributes of LVCs and SVCs have been merged, making them near synonyms. Against this tendency, this study will show cases of SVCs without light verbs (*trazer prejuízo* = *damage*, lit. *bring damage*) and cases of LVCs without support verbs (*dar certo* = *work well*, lit. *give correct*).

To the best of our knowledge, to date, there is no similar study regarding these complex predicates in Brazilian Portuguese, focusing on the development of a lexical resource for NLP tasks, such as SRL. The remainder of this paper is organized as follows: in §2 we discuss related work, in §3 we present the corpus and the details about our methodology, in §4 we present and discuss the resulting lists of candidates, in §5 we envisage further work and draw our conclusions.

2 Related Work

Part of the CPs focused on here are represented by LVCs and SVCs. These CPs have been studied in several languages from different points of view: diacronic (Ranchhod, 1999; Marchello-Nizia, 1996), language contrastive (Danlos and Samvelian, 1992; Athayde, 2001), descriptive (Butt, 2003; Langer, 2004; Langer, 2005) and for NLP purposes (Salkoff, 1990; Stevenson et al., 2004; Barreiro and Cabral, 2009; Hwang et al., 2010). Closer to our study, Hendrickx et al. (2010) annotated a Treebank of 1M tokens of European Portuguese with almost 2,000 CPs, which include LVCs and verbal chains. This lexicon is relevant for many NLP applications, notably for automatic translation, since in any task involving language generation they confer fluency and naturalness to the output of the system.

Work focusing on the automatic extraction of LVCs or SVCs often take as starting point a list of recurrent light verbs (Hendrickx et al., 2010) or a list of nominalizations (Teufel and Grefenstette, 1995; Dras, 1995; Hwang et al., 2010). These approaches are not adopted here because our goal is precisely to identify which are the verbs, the nouns and other lexical elements that take part in CPs.

Similar motivation to study LVCs/SVCs (for SRL) is found within the scope of Framenet (Atkins et al., 2003) and Propbank (Hwang et al., 2010). These projects have taken different decisions on how to annotate such constructions. Framenet annotates the head of the construction (noun or adjective) as argument taker (or frame evoker) and the light verb separately; Propbank, on its turn, first annotates separately light verbs and the predicative nouns (as ARG-PRX) and then merges them, annotating the whole construction as an argument taker.

We found studies regarding Portuguese LVCs/SVCs in both European (Athayde, 2001; Rio-Torto, 2006; Barreiro and Cabral, 2009; Duarte et al., 2010) and Brazilian Portuguese (Neves, 1996; Conejo, 2008; Silva, 2009; Abreu, 2011). In addition to the variations due to dialectal aspects, a brief comparison between these papers enabled us to verify differences in combination patterns of both variants. In addition, Brazilian Portuguese studies do not aim at providing data for NLP applications, whereas in European Portuguese there are at least two studies focusing on NLP applications: Barreiro and Cabral (2009), for automatic translation and Hendrickx et al. (2010) for corpus annotation.

3 Corpus, Extraction Tool and Methods

We employ a corpus-based methodology in order to create a dictionary of CPs. After a first step in which we use a computer software to automatically extract candidate n-grams from a corpus, the candidate lists have been analyzed by a linguist to distinguish CPs from fully compositional word sequences.

For the automatic extraction, the PLN-BR-FULL[2] corpus was used, consisting of news texts from *Folha de São Paulo* from 1994 to 2005, with 29,014,089 tokens. The corpus was first preprocessed for sentence splitting, case homogenization, lemmatization and POS tagging using the PALAVRAS parser (Bick, 2000).

Differently from the studies referred to in Section 2, we did not presume any closed list of light verbs or nouns as starting point to our searches. The search criteria we used contain seven POS patterns observed in examples collected during previous corpus annotation tasks[3]:

1. V + N + PRP: *abrir mão de* (*give up*, lit. *open hand of*);
2. V + PRP + N: *deixar de lado* (*ignore*, lit. *leave at side*);
3. V + DET + N + PRP: *virar as costas para* (*ignore*, lit. *turn the back to*);
4. V + DET + ADV: *dar o fora* (*get out*, lit. *give the out*);
5. V + ADV: *ir atrás* (*follow*, lit. *go behind*);
6. V + PRP + ADV: *dar para trás* (*give up*, lit. *give to back*);
7. V + ADJ: *dar duro* (*work hard*, lit. *give hard*).

This strategy is suitable to extract occurrences from active sentences, both affirmative and negative. Cases which present intervening material between the verb and the other element of the CP are not captured, but this is not a serious problem considering the size of our corpus, although it influences the frequencies used in candidate selection. In order to facilitate human analysis of candidate lists, we used the `mwetoolkit`[4]: a tool that has been developed specifically to extract MWEs from corpora, which encompasses candidate extraction through pattern matching, candidate filtering (e.g. through association measures) and evaluation tools (Ramisch et al., 2010). After generating separate lists of candidates for each pattern, we filtered out all those occurring less than 10 times in the corpus. The entries resulting of automatic identification were classified by their frequency and their annotation is discussed in the following section.

4 Discussion

Each pattern of POS tags returned a large number of candidates. Our expectation was to identify CPs among the most frequent candidates. First we annotated "interesting" candidates and then, in a deep analysis, we judged their idiomaticity. In the Table 1, we show the total number of candidates extracted before applying any threshold, the number of analyzed candidates using a threshold of 10 and the number of CPs by pattern divided into two columns: idiomatic and less idiomatic CPs. Additionally, each CP was annotated with one or more single-verb

[2] `www.nilc.icmc.usp.br/plnbr`

[3] V = VERB, N = NOUN, PRP = PREPOSITION, DET = DETERMINER, ADV = ADVERB, ADJ = ADJECTIVE.

[4] `www.sf.net/projects/mwetoolkit`

Pattern	Extracted	Analyzed	Less idiomatic	Idiomatic
V + N + PRP	69,264	2,140	327	8
V + PRP + N	74,086	1,238	77	8
V + DET + N + PRP	178,956	3,187	131	4
V + DET + ADV	1,537	32	0	0
V + ADV	51,552	3,626	19	41
V + PREP + ADV	5,916	182	0	2
V + ADJ	25,703	2,140	145	11
Total	407,014	12,545	699	74

Table 1: Statistics for the Patterns.

paraphrases. Sometimes it is not a simple task to decide whether a candidate constitutes a CP, specially when the verb is a very polysemous one and is often used as support verb. For example, *fazer exame em/de alguém/alguma coisa* (lit. *make exam in/of something/somebody*) is a CP corresponding to *examinar* (*exam*). But *fazer exame* in another use is not a CP and means to submit oneself to someone else's exam or to perform a test to pass examinations (take an exam). In the following sections, we comment the results of our analysis of each of the patterns.

4.1 Verb + Noun + Preposition

The pattern V + N is very productive, as every complement of a transitive verb not introduced by preposition takes this form. For this reason, we restricted the pattern, adding a preposition after the noun with the aim of capturing only nouns that have their own complements.

We identified 335 complex predicates, including both idiomatic and less idiomatic ones. For example, *bater papo* (*shoot the breeze*, lit. *hit chat*) or *bater boca* (*have an argument*, lit. *hit mouth*) are idiomatic, as their sense is not compositional. On the other side, *tomar consciência* (*become aware*, lit. *take conscience*) and *tirar proveito* (*take advantage*) are less idiomatic, because their sense is more compositional. The candidates selected with the pattern V + N + PRP presented 29 different verbs, as shown in Figure 1[5].

Sometimes, causative verbs, like *causar* (*cause*) and *provocar* (*provoke*) give origin to constructions paraphrasable by a single verb. In spite of taking them into consideration, we cannot call them LVCs, as they are used in their full sense. Examples:

- *provocar alteração* (*provoke alteration*)= *alterar* (*alter*);
- *causar tumulto* (*cause riot*) = *tumultuar* (*riot*).

Some of the candidates returned by this pattern take a deverbal noun, that is, a noun created from the verb, as stated by most works on LVCs and SVCs; but the opposite may also occur: some constructions present denominal verbs as paraphrases, like *ter simpatia por* (*have sympathy for*) = *simpatizar com* (*sympathize with*) and *fazer visita* (lit. *make visit*) = *visitar* (*visit*). These results oppose the idea about LVCs resulting only from the combination of a deverbal noun and a light verb. In addition, we have identified idiomatic LVCs that are not paraphrasable by verbs of the same word root, like *fazer jus a* (lit. *make right to*) = *merecer* (*deserve*).

Moreover, we have found some constructions that have no correspondent paraphrases, like *fazer sucesso* (lit. *make success*) and *abrir exceção* (lit. *open exception*). These findings evidence that, the most used test to identify LVCs and SVC — the existence of a paraphrase formed by a single verb, has several exceptions.

We have also observed that, when the CP has a paraphrase by a single verb, the prepositions that introduce the arguments may change or even be suppressed, like in:

- *Dar apoio* **a** *alguém* = *apoiar alguém* (*give support* **to** *somebody* = *support somebody*);

[5] We provide one possible (most frequent sense) English translation for each Portuguese verb.

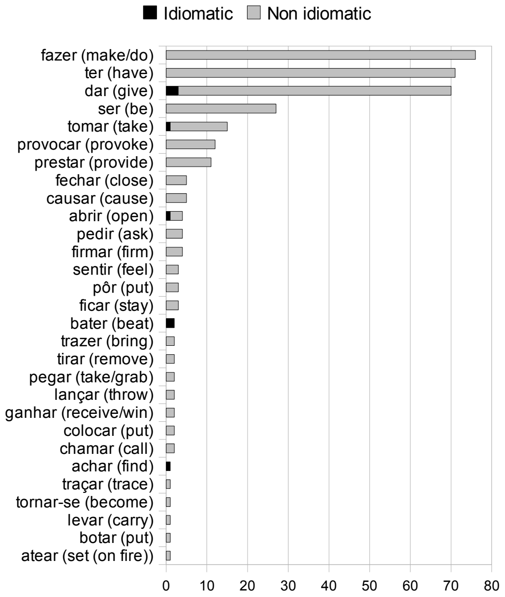


Figure 1: Distribution of verbs involved in CPs, considering the pattern V + N + PRP.

- *Dar cabo* **de** *alguém ou* **de** *alguma coisa* = *acabar* **com** *alguém ou* **com** *alguma coisa* (*give end* **of** *somebody or* **of** *something* = *end* **with** *somebody or* **with** *something*).

Finally, some constructions are polysemic, like:

- *Dar satisfação a alguém* (lit. *give satisfaction to somebody*) = make somebody happy or provide explanations to somebody;

- *Chamar atenção de alguém* (lit. *call the attention of somebody*) = attract the attention of somebody or reprehend somebody.

4.2 Verb + Preposition + Noun

The results of this pattern have too much noise, as many transitive verbs share with this CP class the same POS tags sequence. We found constructions with 12 verbs, as shown in Figure 2. We classified seven of these constructions as idiomatic CPs: *dar de ombro* (*shrug*), *deixar de lado* (*ignore*), *pôr de lado* (*put aside*), *estar de olho* (*be alert*), *ficar de olho* (*stay alert*), *sair de férias* (*go out on vacation*). The later example is very interesting, as *sair de férias* is synonym of *entrar em férias* (*enter on vacation*), that is, two antonym verbs are used to express the same idea, with the same syntactic frame. In the remaining constructions, the more frequent verbs are used to give an aspectual meaning to the noun: *cair em*, *entrar em*, *colocar em*, *pôr em* (*fall in*, *enter in*, *put in*) have inchoative meaning, that is, indicate an action starting, while *chegar a* (*arrive at*) has a resultative meaning.

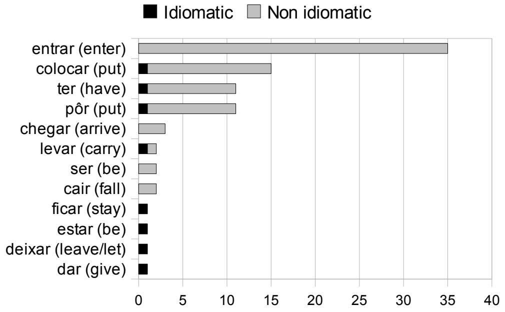


Figure 2: Distribution of verbs involved in CPs, considering the pattern V + PRP + N.

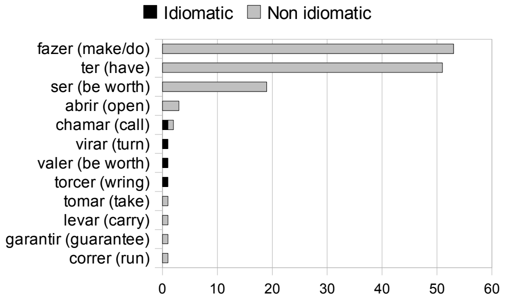


Figure 3: Distribution of verbs involved in CPs, considering the pattern V + DET + N + PRP.

4.3 Verb + Determiner + Noun + Preposition

This pattern gave us results very similar to the pattern V + N + PRP, evidencing that it is possible to have determiners as intervening material between the verb and the noun in less idiomatic CPs. The verbs involved in the candidates validated for this pattern are presented in Figure 3.

The verbs *ser* (*be*) and *ter* (*have*) are special cases. Some *ter* expressions are paraphrasable by an expression with *ser* + ADJ, for example:

- *Ter a responsabilidade por* = *ser responsável por* (*have the responsibility for* = *be responsible for*);

- *Ter a fama de* = *ser famoso por* (*have the fame of* = *be famous for*);

- *Ter a garantia de* = *ser garantido por* (*have the guarantee of* = *be guaranteed for*).

Some *ter* expressions may be paraphrased by a single verb:

- *Ter a esperança de* = *esperar* (*have the hope of* = *hope*);
- *Ter a intenção de* = *tencionar* (*have the intention of* = *intend*);
- *Ter a duração de* = *durar* (*have the duration of* = *last*).

Most of the *ser* expressions may be paraphrased by a single verb, as in *ser uma homenagem para* = *homenagear* (*be a homage to* = *pay homage to*). The verb *ser*, in these cases, seems to mean "to constitute". These remarks indicate that the patterns *ser* + DET + N and *ter* + DET + N deserve further analysis, given that they are less compositional than they are usually assumed in Portuguese.

4.4 Verb + Determiner + Adverb

We have not identified any CP following this pattern. It was inspired by the complex predicate *dar o fora* (*escape*, lit. *give the out*). Probably this is typical in spoken language and has no similar occurrences in our newspaper corpus.

4.5 Verb + Adverb

This pattern is the only one that returned more idiomatic than less idiomatic CPs, for instance:

- *Vir abaixo* = *desmoronar* (lit. *come down* = *crumble*);
- *Cair bem* = *ser adequado* (lit. *fall well* = *be suitable*);
- *Pegar mal* = *não ser socialmente adequado* (lit. *pick up bad* = *be inadequate*);
- *Estar de pé*[6] = *estar em vigor* (lit. *be on foot* = *be in effect*);
- *Ir atrás (de alguém)* = *perseguir* (lit. *go behind (somebody)* = *pursue*);
- *Partir para cima (de alguém)* = *agredir* (lit. *leave upwards* = *attack*);
- *Dar-se bem* = *ter sucesso* (lit. *give oneself well* = *succeed*);
- *Dar-se mal* = *fracassar* (lit. *give oneself bad* = *fail*).

In addition, some CPs identified through this pattern present a pragmatic meaning: *olhar lá* (*look there*), *ver lá* (*see there*), *saber lá* (*know there*), *ver só* (*see only*), *olhar só* (*look only*), provided they are employed in restricted situations. The adverbials in these expressions are expletives, not contributing to the meaning, exception made for *saber lá*, (lit. *know there*) which is only used in present tense and in first and third persons. When somebody says "Eu sei lá" the meaning is "I don't know".

4.6 Verb + Preposition + Adverb

This is not a productive pattern, but revealed two verbal expressions: *deixar para lá (put aside)* and *achar por bem (decide)*.

4.7 Verb + Adjective

Here we identified three interesting clusters:

1. **Verbs of double object**, that is, an object and an attribute assigned to the object. These verbs are: *achar* (*find*), *considerar* (*consider*), *deixar* (*let/leave*), *julgar* (*judge*), *manter* (*keep*), *tornar* (*make*) as in: *Ele acha você inteligente* (lit. *He finds you intelligent* = *He considers you intelligent*). For SRL annotation, we will consider them as full verbs with two internal arguments. The adjective, in these cases, will be labeled as an argument. However, constructions with the verbs *fazer* and *tornar* followed by adjectives may give origin to some deadjectival verbs, like *possibilitar* = *tornar possível* (*possibilitate* = *make possible*). Other examples of the same type are: *celebrizar* (*make famous*), *esclarecer* (*make clear*), *evidenciar* (*make evident*), *inviabilizar* (*make unfeasible*), *popularizar* (*make popular*), *responsabilizar* (*hold responsible*), *viabilizar* (*make feasible*).

[6] The POS tagger classifies *de pé* as ADV.

2. **Expressions involving predicative adjectives**, in which the verb performs a functional role, in the same way as support verbs do in relation to nouns. In contrast to predicative nouns, predicative adjectives do not select their "support" verbs: they combine with any verb of a restrict set of verbs called copula. Examples of copula verbs are: *acabar* (*finish*), *andar* (*walk*), *continuar* (*continue*), *estar* (*be*), *ficar* (*stay*), *parecer* (*seem*), *permanecer* (*remain*), *sair* (*go out*), *ser* (*be*), *tornar-se* (*become*), *viver* (*live*). Some of these verbs add an aspect to the predicative adjective: durative (*andar*, *continuar*, *estar*, *permanecer*, *viver*) and resultative (*acabar*, *ficar*, *tornar-se*, *sair*).

    - The resultative aspect may be expressed by an infix, substituting the combination of V + ADJ by a full verb: *ficar triste* = *entristecer* (*become sad*) or by the verbalization of the adjective in reflexive form: *ficar tranquilo* = *tranquilizar-se* (*calm down*); *estar incluído* = *incluir-se* (*be included*).
    - In most cases, adjectives preceded by copula verbs are formed by past participles and inherit the argument structure of the verb: *estar arrependido de* = *arrepender-se de* (lit. *be regretful of* = *regret*).

3. **Idiomatic CPs**, like *dar duro* (lit. *give hard* = *make an effort*), *dar errado* (lit. *give wrong* = *go wrong*), *fazer bonito* (lit. *make beautiful* = *do well*), *fazer feio* (*make ugly* = *fail*), *pegar leve* (lit. *pick up light* = *go easy*), *sair errado* (lit. *go out wrong* = *go wrong*), *dar certo* (lit. *give correct* = *work well*).

4.8 Summary

We identified a total of 699 less idiomatic CPs and observed the following recurrent pairs of paraphrases:

- V = V + Deverbal N, e.g. *tratar* = *dar tratamento* (*treat* = *give treatment*);
- Denominal V = V + N, e.g. *amedrontar* = *dar medo* (*frighten* = *give fear*);

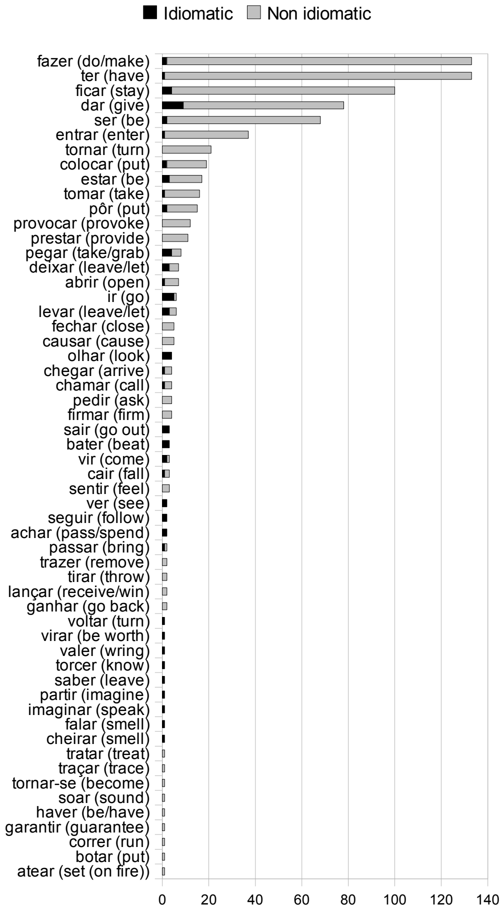


Figure 4: Distribution of verbs involved in CPs, considering the total number of CPs (i.e. all patterns).

- Deadjectival V = V + ADJ, e.g. *responsabilizar* = *tornar responsável* (lit. *responsibilize* = *hold responsible*).

This will help our further surveys, as we may search for denominal and deadjectival verbs (which may be automatically recognized through infix and suffix rules) to manually identify corresponding CPs. Moreover, the large set of verbs involved in the analyzed CPs, summarized in Figure 4, shows that any study based on a closed set of light verbs will be limited, as it cannot capture common exceptions and non-prototypical constructions.

5 Conclusions and Future Work

This study revealed a large number of CPs and provided us insights into how to capture them with more precision. Our approach proved to be very useful to identify verbal MWEs, notably with POS tag patterns that have not been explored by other studies (patterns not used to identify LVCs/SVCs). However, due to the onus of manual annotation, we assume an arbitrary threshold of 10 occurrences that removes potentially interesting candidates. Our hypothesis is that, in a machine-readable dictionary, as well as in traditional lexicography, rare entries are more useful than common ones, and we would like to explore two alternatives to address this issue. First, it would be straightforward to apply more sophisticated filtering techniques like lexical association measures to our candidates. Second, we strongly believe that our patterns are sensitive to corpus genre, because the CPs identified are typical of colloquial register. Therefore, the same patterns should be applied on a corpus of spoken Brazilian Portuguese, as well as other written genres like web-crawled corpora. Due to its size and availability, the latter would also allow us to obtain better frequency estimators.

We underline, however, that we should not underestimate the value of our original corpus, as it contains a large amount of unexplored material. We observed that only the context can tell us whether a given verb is being used as a full verb or as a light and/or support verb[7]. As a consequence, it is not possible to build a comprehensive lexicon of light and support verbs, because there are full verbs that function as light and/or support verbs in specific constructions, like *correr* (*run*) in *correr risco* (*run risk*). As we discarded a considerable number of infrequent lexical items, it is possible that other unusual verbs participate in similar CPs which have not been identified by our study.

For the moment, it is difficult to assess a quantitative measure for the quality and usefulness of our resource, as no similar work exists for Portuguese. Moreover, the lexical resource presented here is not complete. Productive patterns, the ones involving nouns, must be further explored to enlarge the aimed lexicon. A standard resource for English like DANTE[8], for example, contains 497 support verb constructions involving a fixed set of 5 support verbs, and was evaluated extrinsically with regard to its contribution in complementing the FrameNet data (Atkins, 2010). Likewise, we intend to evaluate our resource in the context of SRL annotation, to measure its contribution in automatic argument taker identification. The selected CPs will be employed in an SRL project and, as soon as we receive feedback from this experience, we will be able to report how many CPs have been annotated as argument takers, which will represent an improvement in relation to the present heuristic based only on parsed VPs.

Our final goal is to build a broad-coverage lexicon of CPs in Brazilian Portuguese that may contribute to different NLP applications, in addition to SRL. We believe that computer-assisted language learning systems and other Portuguese as second language learning material may take great profit from it. Analysis systems like automatic textual entailment may use the relationship between CPs and paraphrases to infer equivalences between propositions. Computational language generation systems may also want to choose the most natural verbal construction to use when generating texts in Portuguese. Finally, we believe that, in the future, it will be possible to enhance our resource by adding more languages and by linking the entries in each language, thus developing a valuable resource for automatic machine translation.

Acknowledgements

We thank the Brazilian research foundation FAPESP for financial support.

References

Débora Taís Batista Abreu. 2011. A semântica de construções com verbos-suporte e o paradigma Framenet. Master's thesis, São Leopoldo, RS, Brazil.

1997. *Complex Predicates*. CSLI Publications, Stanford, CA, USA.

Maria Francisca Athayde. 2001. *Construções com verbo-suporte (funktionsverbgefüge) do português e do alemão*. Number 1 in Cadernos do CIEG Centro Interuniversitário de Estudos Germanísticos. Universidade de Coimbra, Coimbra, Portugal.

Sue Atkins, Charles Fillmore, and Christopher R. Johnson. 2003. Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3):251–280.

Sue Atkins, 2010. *The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data*. Menha Publishers, Kampala, Uganda.

[7] A verb is not light or support in the lexicon, it is light and/or support depending on the combinations in which it participates.

[8] www.webdante.com

Anabela Barreiro and Luís Miguel Cabral. 2009. ReEscreve: a translator-friendly multi-purpose paraphrasing software tool. In *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*, pages 1–8, Ottawa, Canada, Aug.

Eckhard Bick. 2000. *The parsing system Palavras*. Aarhus University Press.

Miriam Butt. 2003. The light verb jungle. In *Proceedings of the Workshop on Multi-Verb Constructions*, pages 243–246, Trondheim, Norway.

Cássia Rita Conejo. 2008. O verbo-suporte fazer na língua portuguesa: um exercício de análise de base funcionalista. Master's thesis, Maringá, PR, Brazil.

Laurence Danlos and Pollet Samvelian. 1992. Translation of the predicative element of a sentence: category switching, aspect and diathesis. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 21–34, Montréal, Canada.

Mark Dras. 1995. Automatic identification of support verbs: A step towards a definition of semantic weight. In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence*, pages 451–458, Canberra, Australia. World Scientific Press.

Inês Duarte, Anabela Gonçalves, Matilde Miguel, Amália Mendes, Iris Hendrickx, Fátima Oliveira, Luís Filipe Cunha, Fátima Silva, and Purificação Silvano. 2010. Light verbs features in European Portuguese. In *Proceedings of the Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features (Verb 2010)*, Pisa, Italy, Nov.

Iris Hendrickx, Amália Mendes, Sílvia Pereira, Anabela Gonçalves, and Inês Duarte. 2010. Complex predicates annotation in a corpus of Portuguese. In *Proceedings of the ACL 2010 Fourth Linguistic Annotation Workshop*, pages 100–108, Uppsala, Sweden.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Yuping Zhou, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the ACL 2010 Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden.

Otto Jespersen. 1965. *A Modern English Grammar on Historical Principles*. George Allen and Unwin Ltd., London, UK.

Stefan Langer. 2004. A linguistic test battery for support verb constructions. *Special issue of Linguisticae Investigationes*, 27(2):171–184.

Stefan Langer, 2005. *Semantik im Lexikon*, chapter A formal specification of support verb constructions, pages 179–202. Gunter Naar Verlag, Tübingen, Germany.

Christiane Marchello-Nizia. 1996. A diachronic survey of support verbs: the case of old French. *Langages*, 30(121):91–98.

Maria Helena Moura Neves, 1996. *Gramática do português falado VI: Desenvolvimentos*, chapter Estudo das construções com verbos-suporte em português, pages 201–231. Unicamp FAPESP, Campinas, SP, Brazil.

Ryan North. 2005. Computational measures of the acceptability of light verb constructions. Master's thesis, Toronto, Canada.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild? the mwetoolkit comes in handy. In *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Elisabete Ranchhod, 1999. *Lindley Cintra. Homenagem ao Homem, ao Mestre e ao Cidadão*, chapter Construções com Nomes Predicativos na Crónica Geral de Espanha de 1344, pages 667–682. Cosmos, Lisbon, Portugal.

Graça Rio-Torto. 2006. O Léxico: semântica e gramática das unidades lexicais. In *Estudos sobre léxico e gramática*, pages 11–34, Coimbra, Portugal. CIEG/FLUL.

Morris Salkoff. 1990. Automatic translation of support verb constructions. In *Proc. of the 13th COLING (COLING 1990)*, pages 243–246, Helsinki, Finland, Aug. ACL.

Hilda Monetto Flores Silva. 2009. Verbos-suporte ou expressões cristalizadas? *Soletras*, 9(17):175–182.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In , *Proc. of the ACL Workshop on MWEs: Integrating Processing (MWE 2004)*, pages 1–8, Barcelona, Spain, Jul. ACL.

Simone Teufel and Gregory Grefenstette. 1995. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proc. of the 7th Conf. of the EACL (EACL 1995)*, pages 98–103, Dublin, Ireland, Mar.