Syntactic Transformations in Rule-Based Parsing of Support Verb Constructions: Examples from European Portuguese

Jorge Baptista[1]
University of Algarve, Campus de Gambelas, Faro, Portugal
INESC-ID, Lisboa, Portugal
https://www.researchgate.net/profile/Jorge_Baptista
jbaptis@ualg.pt

Nuno Mamede
Universidade de Lisboa, Instituto Superior Técnico, Portugal
INESC-ID, Lisboa, Portugal
Nuno.Mamede@inesc-id.pt

Abstract

This paper reports on-going work on building a rule-based grammar for (European) Portuguese, incorporating support verb constructions (SVC). The paper focuses on parsing sentences resulting from syntactic transformations of SVC, and presents a methodology to automatically generate testing examples directly from the SVC Lexicon-Grammar matrix where their linguistic properties are represented. These examples allow both to improve the linguistic description of these constructions and to test intrinsically the system parser, spotting unforeseen issues due to previous natural language processing steps.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Natural language generation; Computing methodologies → Language resources

Keywords and phrases Support verb constructions, Rule-based parsing, syntactic transformations, language resources, European Portuguese

Digital Object Identifier 10.4230/OASIcs.SLATE.2020.11

Funding This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

Acknowledgements The authors would like to thank Sónia Reis (U.Algarve and INESC-ID Lisboa for her help in the linguistic data compilation and revision, as well as her attentive reading of initial versions of this manuscript. Naturally, any errors herein are our own responsibility alone.

1 Transformations on Support Verb Constructions: Why is this still a thing?

This paper addresses some issues involved in parsing Support Verb Constructions (henceforward *SVC*), considering not only the basic, elementary sentence forms, but also the sentences that result from the basic form having undergone some type of transformation (both some very general transformations and other not-so-general operations, but specific of these construnctions).

[1] Corresponding author

9th Symposium on Languages, Applications and Technologies (SLATE 2020).
Editors: Alberto Simões, Pedro Rangel Henriques, and Ricardo Queirós; Article No. 11; pp. 11:1–11:14
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

SVC are a large set of the elementary (or base) sentences of many languages, and consist of a *predicate noun* (*Npred*) and a *support verb* (*Vsup*), along with its subject and eventual essential complements. The concept of support verb can be traced back to Zellig S.Harris [31, p.216], though the term has been coined much later by M. Gross [25]. In a sentence such as (1):

(1) *O Pedro deu um soco ao João* "Pedro gave a punch to João"

we say that *soco* "punch" is a predicate noun and *deu* "gave" is a support verb. This sentence is a clear example of a SVC: the predicate noun *soco* "punch" is the nucleus of the elementary sentence, the element that conveys the semantic predicate, while the support verb can be considered a specialised type of auxiliary, practically devoid of meaning, and whose function is, basically, to convey the person-number and tense values, which the noun cannot express morphologically. It is the predicate noun (and not the verb!) that selects the elements that fill its argument slots; and it is the noun that selects support verb itself (and not vice-versa). It is also this particular verb-noun combination that imposes the sentence structure, including the prepositions introducing the prepositional complements (if any), as well as the syntactic properties of the construction.

Though the study of SVC is a well-established field of enquiry, dating at least from the early 1960s [31], when the linguistic status of these constructions came into the focus of theoretical debate ([17]), it has gained a renewed impetus with the recent growing interest in processing multiword expressions (MWE) [18, 46, 47] and the development of linguistic resources (especially annotated *corpora*) [37], particularly those envisioned for machine-learning approaches to MWE extraction [52].

Extensive literature has been produced on SVC, from the linguistic viewpoint, and for many languages (see [33] for an overview and references therein), and much work has been invested in the description of (European) Portuguese SVC, namely on the construction with Vsup *estar Prep* [38], *ser de*, *dar* [4, 6, 51], *fazer* [16] and others [2, 20]. More recently, extensive surveys of SVC from the Brazilian variety of Portuguese have been produced: [21] (*fazer*), [45] (*ter*), [41] (*dar*) and others [14, 43].

As multiword expressions [15, 18, 44], SVC constitute a challenge for Natural Language Processing (NLP), both in the perspective of their automatic recognition in texts [32, 47] and their integration in NLP systems [46, 40, 39]. Some corpora are also available for testing the processing of MWE, including some types of SVC [37, 42] (see [18] for an overview).

In spite of the volume of the work already produced, not much attention has been given to the challenges posed by transformations to the parsing of SVC. Not only do SVC give rise to specific transformations, such as:

Conversion [24]:

(2) *O Pedro deu um soco ao João = O João levou um soco do Pedro*
"Pedro gave a punch to (punched) João = João took a punch from Pedro"

complex NP formation [25]:

(3) *o soco que o Pedro deu ao João <...> = o soco do Pedro ao João <...>*
"the punch that Pedro gave to João = the punch of Pedro to João"

Nasp aspectual noun insertion [38]:

(4) *A empresa está em (processo de) reestruturação*
"The company is in (process of) restructuring"; and

Vop Vsup reduction and CSV restructuring under the so-called (causative) operator verbs (Vopc)[25]:

(5) *O Pedro tem medo do escuro = Aquele incidente causou-lhe medo do escuro*
"Pedro has fear of the dark (Pedro is afraid of the dark) = That incident caused him fear of the dark"

Still, SVC can also undergo very general transformations, such as [Passive], [Relative], [Symmetry] [5], and [NP restructuring] [3, 29, 34]. Even if most of these operations are already relatively well-known, their combined application to SVC render the task of parsing these complex constructions a non-trivial task. For lack of space, the reader will refer to the references above for a more detailed description of the SVC specific properties and the associated transformations. This paper main contribution resides, thus, in a method to systematically explore this complex interaction of SVC lexicon-grammar and associated transformations within the scope of building a rule-based grammar for parsing Portuguese texts.

The paper reports on an on-going project to build an integrated lexicon-grammar of Portuguese SVC, within the Lexicon-Grammar (LG) theoretical and methodological framework [25, 28, 33]. Extant linguistic descriptions date from the late 1980's til more recent work on the Brazilian variety (mid-2010s). In this paper, the focus is the European Portuguese SVC. In the development of this research, we have come to realize that some authors did not always use precisely the same definitions for many of their distributional and transformational descriptions, so we put to ourselves the task of compiling and revising all this immense bulk of data, and systematically provide a coherent and explicit description of the linguistic properties encoded in the LG. In the process of doing so, it became obvious that only the more recent work provided illustrative (either artificial or corpus-retrieved) examples for the linguistic description. The change in perspective was slow but steady, very probably having begun with [27] (French adverbial idioms). Older work (until the late 1990s) had few to no examples next to the LG resources, which were typically encoded in binary matrices. It was up to the linguist to creatively devise the adequate wording for the abstract, structural (and often theoretically motivated) description encoded in the matrices, though taking several precautions not to produce biased examples [26]. Naturally, the technological evolution brought by the personal computer and the renewed impetus of corpus-based, data-driven Linguistics also had some influence in this shift.

Example-building is not trivial, and several strategies can be combined to achieve different purposes. More recently, when describing Portuguese verbal constructions (full or distributional verbs) [7, 8, 11], and verbal idioms [9, 12, 19, 23] in view of their integration into STRING [35], a NLP pipeline system, with a rule-based parser (XIP)[1], we also felt the need to produce in a systematic way a comprehensive set examples. In these cases, first steps were taken to deal with lexically constraint transformations, that is, a limited set of transformations, specific to the verbal constructions and the verbal idioms. These transformations include pronominalisations, passive constructions (with both auxiliary verbs *ser* and *estar* "be"), symmetry[5, 10], and some types of NP restructuring [3] (see below).

The goal of automatically generating examples directly from the linguistic description in the LG served two main purposes:

- to validate the grammar rules devised for the parser, and thus serving as a testing benchmark; previous processing steps (POS-tagging and disambiguation, chunking, and dependency extraction) may fail and the error is not a fault of the piece of grammar produced for that particular phenomena under study, but it results, instead, from the pervasive ambiguity and complexity of natural language and the considerable difficulty in solving it in full;

- to facilitate the task of spotting linguistic inconsistencies or inadequacies in the LG description, thus enabling the linguist to revise, correct or complete the linguistic data in the LG resource and, eventually, aid in the development of a more precise grammar.

Both these situations will be exemplified.

Naturally, using a mechanical instead of a manual process to produce examples for the LF of SVC was soon necessary due to the complexity of the task, the many linguistic factors involved and the complex interaction between successive transformations applied to the base form. This is not to say that using a real-life, corpus-based, evaluation scenario, such as the one used in [37], could not be used for evaluating both the linguistic resources and the rule-based grammar, as that type of evaluation can be made to improve both, adding to structural description the dimension of usage. This, however, is out of the scope of this work.

The paper is organized as follows: Next, in Section 2, a brief description of the example generation process is provided, and preliminary results are presented (Section 3). The paper concludes (Section 4) with some remarks on current issues and perspectives for future work.

2 Example generation

To automatically generate examples of SVC directly from the linguistic information encoded in the SVC lexicon-grammar matrix, a Perl software was developed in-house. During the LG construction, another software, also developed in-house, validates the format and the consistency of the data and outputs error messages, allowing the correction and maintenance of the data matrix. This is done by a set of several dozens of rules. For example, if the number of arguments of a Npred is only one, then all the properties for the N_1 and N_2 argument slots must be marked "-", otherwise an error message is produced.

In the LG matrix, each line corresponds to a lexicon-grammar entry (a predicate noun); multiple word senses appear in distinct lines. Each Npred is defined according to the arity of its argument domain, and this can be either "1" (only subject, N_0), "2" (subject N_0 and first complement N_1), or "3" (subject N_0, first N_1 and second complements complement N_2). Example-generating rules are structured according to the number of arguments.

Distributional constraints (on argument slots) are used to generate the examples. These include human/non-human opposition, for instance, but can sometimes be further refined using semantic features. The semantic features were adapted from E. Bick semantic prototypes [13][2]. Besides those features, particularly relevant lexical items are explicitly stated, distinguishing lemmas and inflected/invariant forms The set of distributional constraints is then translated into a *basic string*. These also help define in a more precise way those properties. For example, for subject (N_0) distributional constraints, the following basic strings are used :

Nhum ± human noun; typically, a proper noun: *o Pedro*;

NñHum ± non-human noun; typically a concrete noun: *esta coisa* "this thing"; for consistency, other non-human features [Npc], [Nloc] and [Npred_de_N] (see below) imply that [NñHum] be marked as "-".

Nnr ± non-constraint noun; weakly constraint slot, with a <cause> semantic role; only used for subject: *isto* "this" ;

Npc ± body-part noun, represented by the semantic prototype "sem-an" in the appropriate matrix column, and by a list of nouns, adequate for a given Npred; the basic string is produced by using the first lexical item of that list; otherwise, it uses *a mão* "the hand" as a *portmanteau* word (irrespective of its adequacy);

[2] Semantic roles, based on [48, 49, 50] are indicated for each argument slot but they are not used for example generation.

Nloc ± locative noun: *este lugar* "this place";
Npred_de_N ± complex NP with a Npred head and its arguments (currently not implemented);
Vinfw ± infinitive subclause: *o Pedro fazer isso* "Pedro to_do this";
QueFconj ± finite sub-clause in the subjunctive "mood": *que o Pedro faça isto* "that Pedro does this";
QueFind ± finite sub-clause in the indicative "mood": *que o Pedro faz isto* "that Pedro does this";
O_facto_de_queF ± factive sub-clause: **o facto de o Pedro fazer isto** "the fact that Pedro does this";
Npl-obr ± obligatory plural (currently not implemented);

First (N_1) and second (N_2) complement distributional constraints are encoded in a similar way. For consistency, different proper names were used for N_1 (*João*) and N_2 (*Rui*) complements. Also, different determiners (e.g. *essa coisa* "that thing", and *aquela coisa* "the other thing") and, in the case of completives, different indefinite pronouns (*isso* and *aquilo* "that") were used to distinguish these syntactic slots. Prepositions introducing the complements ($Prep_1$ and $Prep_2$, respectively) are taken directly from the matrix, where they are explicitly provided.

Three different sentence structures are associated to Vsup Npred constructions and represented in the LG, both for the standard and the converse constructions:
CDIR ± for direct-transitive support verbs, where the Npred is the direct complement of the Vsup, e.g. *dar um soco* "give a punch";
PREDSUBJ ± for copula-like Vsup like *estar Prep* "be Prep", with a Prep introducing the Npred, e.g. *O Pedro está em crise* "Pedro is in crisis"; and
MOD ± for verbs with the Npred in a prepositional complement; e.g. *O Pedro sofre de asma* "Pedro suffers from asthma".
For each type of these three types of SVC construction, the Vsup selected by each Npred are listed; Vsup-Prep pairings in the `PREDSUBJ` and `MOD` construction are also indicated. The preposition introducing the <agent-like> complement in the converse construction is also explicitly indicated (mostly, Prep *de* and *da/por parte de*).

These structures have to do with the dependencies produced by the system's parser using the Portuguese grammar. As explained in [40], we identify the SVC by a specific dependency `support`, linking the Npred to the Vsup; a feature `_vsup-standard/converse` indicates wether this is a standard or a converse construction, which will be relevant for semantic role labelling at a later stage; e.g.,

(6) *O Pedro estabeleceu uma aliança com o João* "Pedro established an alliance with João"
`SUPPORT_VSUP-STANDARD(aliança,estabeleceu)`

A similar structural description is also used here to automatically generate the SVC examples. Hence, to generate the example sentence for a `CDIR`-type SVC, the structural elements are aligned, using the basic strings for the arguments, an inflected form of the Vsup, an eventual determiner[3] for the Npred and the prepositions it selects to introduce its eventual complements. In case multiple values appear in the same cell (e.g Prep or Vsup), or for different combinations of distributional constraints on the argument-slots (e.g. human/non-human subject), the algorithm explores all variants and combinations, producing a separate example for each.

[3] For lack of space, determiner-modifier variation has not been described here.

For generating the examples derived by transformations, a similar procedure is carried out. The [dative] pronominalization of the complement arguments, encoded next to the constituent description, is translated by a dative pronoun *-lhe* "to_him", attached to the Vsup, e.g., *O Pedro deu um soco ao João=O Pedro deu-lhe um soco* "Pedro gave him a punch".

The [NP restructuring] involving body-part nouns (Npc; only encoded for N_1), produces a complex subject NP, from two independent constituents, e.g. *O Pedro tem acne no rosto = O rosto do Pedro tem acne* "Pedro has acne on his face = Pedro's face has acne". Complex noun phrase [Complex NP] generation uses the Npred lexical item, followed by the preposition *de* "of" and the subject basic string; for 2- and 3-argument predicates, the corresponding prepositions (Prep_1 and Prep_2, respectively) are used along with the basic strings of those slots; the basic order of the arguments is maintained.

The [Symmetry] transformation consists in the coordination (*e* "and") of two arguments in a given syntactic slot, using the basic strings of those arguments; in the case of 3-argument predicates, either a subject-object or an object-object coordinated NP is produced, depending on the type of symmetry involved. Hence, for the subject-object symmetric noun *acordo* "agreement" the basic strings produce *esta pessoa e aquela pessoa* [*estão de acordo*] "This person and that person [are in agreement]"; while for the object-object symmetric noun *mistura* "mixture", the basic strings produce [O Pedro fez uma mistura] dessa coisa e aquela coisa "[Pedro did a mixture] of this thing and that thing".

The [ObligNeg] (obligatory negation) property can be seen in SVC that contain an negation element [22], e.g. *O Pedro não esteve pelos ajustes* lit:"Pedro was not by the adjustments" "not to accept or disagree with something that is proposed, presented or required", otherwise the sentence is meaningless or has another unrelated meaning. Generating this examples involves introducing a negation adverb não "not" before the Vsup.

The aspectual nouns [Nasp] insertion [38], come next. These are a type of auxiliary elements that can be inserted in the base sentence leaving the Npred as its complement. They convey an aspectual value, hence the term, and they usually render the sentence more natural. Their function in the SVC is homologous to that of auxiliary verbs (*aka.* verbal periphrasis) in full verb constructions. With Vsup *estar Prep*, the most frequente Nasp are *estado* "state", *fase* "phase", *processo* "process" and, less frequently, *vias* "verge" (7):

(7) *Esta espécie está em extinção = Esta espécie está em* ***vias*** *de extinção* "This species is in extinction (endangered) = This species is on the verge of extinction"

Certain Npred with Vsup *ter* ou *estar com*, denoting "illness/desease" select other *Nasp*, such as *ataque* "attack" and *crise* "crisis" (8):

(8) *O Pedro tem/está com asma = O Pedro está com um* ***ataque****/uma* ***crise*** *de asma*
"Pedro has/is with asthma = Pedro is with an asthma attack/crisis"

Finally, (causative) operator-verbs (Vopc) [25] insertions are described. These verbs reshape the basic SVC structure, absorving the Vsup, and altering the syntactic dependencies associated to the Npred arguments. Two structurally different constructions are considered: (i) [VOP-CDIR], when the Npred is a direct complement of the Vop:

(9) *O Pedro tem sede = Isto deu/fez sede ao Pedro*
"Pedro has thirst (is thirsty) = This gave/made thirst to Pedro (made Pedro thirsty)"

(ii) [VOP-MOD] when the Npred is a prepositional complement of the Vop:

(10) *O Pedro está com sede = Isto deixou o Pedro com sede*
"Pedro is with thirst (is thirsty) = This left Pedro with thirst (left Pedro thirsty)"

In the [Passive] constructions, not only is the sentence with auxiliary verb *ser* "be" generated, but also all the reductions that it can undergo both in the standard and in the converse constructions:

(11) *O Pedro deu um soco ao João* "Pedro gave a punch to João" [STD]

(12) [Passive] = *Um soco foi dado pelo Pedro ao João* = [Relative] *O soco que foi dado pelo Pedro ao João* = [RedRel] *O soco dado pelo Pedro ao João* = [RedVsup] *O soco do Pedro ao João*
"A punch was given by Pedro to João = The punch that was given by Pedro to João = The punch given by Pedro to João = The punch by Pedro to João"

(13) *O João apanhou um soco do Pedro* "João got a punch from Pedro"

(14) [Passive] = *Um soco foi apanhado pelo João do Pedro* = [Relative] *O soco que foi apanhado pelo João do Pedro* = [RedRel] ?*O soco apanhado pelo João do Pedro*
"A punch was caught by João from Pedro = The punch that was caught by João from Pedro = The punch caught by João from Pedro"

A specific column was added to the LG matrix representing gender-number values of the Npred, in order to ensure the correct agreement with the sentences' elements (determiner, modifier and Vsup agreement). A list of tensed forms for each Vsup was used to produce more natural sentences.

3 Results

At the time of submission, the SVC Lexicon-Grammar of European Portuguese contains approximately 7,150 entries. So far, 2,741 (38.3%) have been carefully revised. From these, 1,487 have only one argument, 1,178 have 2 arguments, and 76 have 3 arguments. For an easier inspection of the generated examples, each sentence-type is outputted to a different file. Table 1 shows the breakdown of the generated examples per sentence type and per number of the Npred arguments. A dash "–" indicates that the sentence type cannot be construed, while the note (a) corresponds to work still in progress.

First, we remark that these are just preliminary results. Still, even if only a little more than 1/3 of the LG entries have been processed, it is already evident that the number of automatically generated examples (48,421) is quite impressive. Furthermore, some transformations are still being worked out. It is likely that other, though less productive, transformations are to be added.

In order to assess the generated examples, same caution is required, keeping in mind that our goal is *not* the generation of entirely natural utterances, but their analysis. In other words, the purpose of those example sentences is to test the STRING system's [35] and its parsing module XIP [1], when dealing with SVC, and, particularly at this stage, to extract the SUPPORT dependency out of those examples. Stylistic considerations, though important in a generation system, are secondary here.

Furthermore, the size of the list of examples being so large, it is difficult to provide a direct quantitative assessment of the generated examples, so we will limit ourselves to highlight the main issues detected. For example, in these artificial sentences, constituents

Table 1 SVC automatically generated examples.

Sentence type	Arg=1	Arg=2	Arg=3
STD	10,458	9,059	795
STD-Pass	1,125	4,628	570
STD-NP	2,002	1,889	280
STD-Nasp	815	818	0
STD-ObligNeg	4	0	0
STD-VOP-CDIR	4,687	2,571	(a)
STD-VOP-MOD	1,388	1,693	(a)
STD-NP-Restr	–	1,035	(a)
STD-Dat	–	84	58
STD-Sym	–	2,567	113
CNV	–	(a)	(a)
CNV-Pas	–	1,476	306
Sub-total	20,479	25,820	2,122
Total		**48,421**	

are produced in the basic word order. This often produces formally (syntactically) correct but stylistically dubious (or even unacceptable) sentences; e.g. a subject infinitive sub-clause is more natural if moved to the end of the sentence:

(15) ?*O Pedro fazer isto está na moda* "Pedro to-do this is in fashion (=is fashionable)"
= *Está na moda o Pedro fazer isto* "[it] is in fashion Pedro to-do this"

Also, notice that in the sub-clause a zero-indefinite or an indefinite subject is preferable that the basic string *o Pedro*:

(16) *Está na moda fazer isto* "[it] is in fashion to-do this";
Está na moda as pessoas fazerem isto "[it] is in fashion people to-do this".

Distributional constraints are only approximated by the basic strings chosen for the generation process. This produces sometimes quite bizarre expressions. For example, *estar de esperanças* "be of/with hopes/expectations" or *estar no seu estado interessante* "be in her interesting state", which means "to be pregnant", can hardly accept a masculine subject like *o Pedro*. In other cases, a human-collective noun would better suit the Npred:

(17) *O Pedro fez uma inspeção a* (?*o João, ao pelotão, à empresa*)
"Pedro made an inspection to (João/the platoon/the company)"

Co-reference constraints holding between the Npred arguments and the sub-clause arguments (especially its subject) were simply ignored at this stage, in order to simplify the generation process, which produces borderline (if not altogether unacceptable) sentences (co-reference is marked by co-reference indexes in the examples below):

(18) **O Pedro*$_i$ *teve a intenção de o João*$_j$ *fazer isso*
"Pedro had the intention of John to-do this"

(19) cp. *O Pedro*$_i$ *teve a intenção de 0*$_i$ *fazer isso*
"Pedro had the intention of to-do this"

Concerning the Npred determiners, their representation in the Lexicon-Grammar is limited to those the noun selects in the base form. The rationale for this decision is that most of times, the constraints on the Npred determiners are very similar across multiple Vsup constructions of the same Npred. This, in fact, is not always so, and same generated examples are quite awkward. For example, the Npred *juramento* "oath", besides the elementary (basic) Vsup *fazer* "to do/make", also accepts, the variant *prestar* "pay". An exact match query in the `.pt` top domain of the web using Google shows that the first Vsup rarely accepts the zero determiner, while the second is significantly more frequent with this determiner.

(20) *O Pedro ?*fez/prestou juramento ao João/a esta coisa*
"Pedro made/payed oath to João/that thing"

The reverse situation occurs with the indefinite article *um* "a".

(21) *O Pedro fez/?*prestou* um *juramento ao João/a esta coisa*
"Pedro made/payed an oath to João/that thing"

In order to mimic the situations where the Npred imposes the presence of a modifier, we decided to use the basic string *um certo* "a certain". The vagueness of the determiner sometimes produce unnatural examples. The selection of an adequate adjetive can significantly improve the acceptability of the sentence:

(22) *O Pedro está com uma ?certa/forte cãimbra no pé*
"Pedro has got a certain/strong cramp in the foot (=a foot cramp)"

Another aspect that hinders the acceptability of generated examples is the fact that some Npred, though allowing number variation, are much more frequent in the plural with a given Vsup that with another one. This constraint is often associated with the determiners (and some of these combined restrictions may show high regularity). Since in the LG matrix Npred are indicated by their lemma and examples are generated directly form the LG entry, some examples, though grammatically correct, may sound awkward. For example, the Npred *borbulha* "pimple" with Vsup *ter* is much more acceptable in the plural with determiner zero, while both number values are natural with the indefinite article:

(23) *O Pedro tem borbulhas/*borbulha na cara*
"Pedro's got pimples/*pimple on the [=his] face"

(24) *O Pedro tem umas borbulhas/uma borbulha na cara*
"Pedro's got some pimples/a pimple on the [=his] face"

Obligatory Npred plural/singular forms are represented by their surface forms, irrespective of this constraint on number value being strictly morphologic, e.g., *férias* "holidays", *pêsames* "condolences"; or strictly syntactic, e.g., *braços* "arms": *O Pedro está a braços com um problema grave* "Pedro's in having to deal with a serious problem."

Another way to assess the generated examples is to parse them in STRING and check in the output whether the support dependency was correctly extracted. Table 2 shows the breakdown of error rate (false-negatives) per sentence type. The new note (b) indicates that this assessment does not applies to complex NP, as there is no Vsup in such structures.

The system takes 1h19m50s to process the generated examples' files. The overall error-rate is 0.0506 but this value varies widely depending of the sentence type. A detailed error analysis was carried out and the main issues found had to do with inadequacies in different processing stages is STRING, which prevent the SVC detection and the SUPPORT dependency extraction. Here are the most relevante situations found:

Table 2 Parsing automatically generated SVC examples: error rate (false-negatives).

Sentence type	Arg=1	Arg=2	Arg=3
STD	20/10,458=0.0019	183/9,059=0.0202	14/795=0.0176
STD-Pass	12/1,125=0.0106	1,397/4,628=0.3018	2/570=0.0035
STD-NP	(b)	(b)	(b)
STD-Nasp	2/815=0.0025	12/818=0.0146	(a)
STD-ObligNeg	0/4=0.0000	0/0=0.0000	(a)
STD-VOP-CDIR	15/4,687=0.0032	212/2,571=0.0824	(a)
STD-VOP-MOD	167/1,388=0.1203	323/1,693=0.1907	(a)
STD-NP-Restr	–	14/1,035=0.0135	(a)
STD-Dat	–	16/84=0.1904	14/58=0.2413
STD-Sym	–	22/2,567=0.0085	21/113=0.1858
CNV	–	(a)	(a)
CNV-Pas	–	10/1,476=0.0067	0/306=0.0000
Sub-total	216/20,479=0.0105	2,189/25,820=0.0847	51/2,122=0.0240
Total	**2,454/48,421=0.0506**		

(i) lacunae in the system's lexicon; e.g. *bolandas*: *andar em bolandas* "in a bustle"; the new entry was then added to the lexicon;

(ii) misspelling of the Npred lemma in the LG, particularly in the case of compound words and the use of hyphen, as an exact match with that lemma in the system's lexicon; e.g. *dor-de-cotovelo*/*dor de cotovelo* lit:"pain in the elbow" "envy/jealousy"; the entry was corrected in the lexicon-grammar;

(iii) incorrect tokenization of a string as a multiword expression (MWE), especially compound prepositions and adverbs; as tokenization of MWE has priority over simple word sequences, capturing a compound precludes the Npred identification and all subsequent processing steps; e.g. *na direção de* (Prep) "towards" vs. *O Pedro está na direção da empresa* "Pedro is at the head of the company/on the company's board"; conditions were added to the system, in order to prevent the tokenization of the string as a MWE;

(iv) incorrect statistical POS-disambiguation; several situations arose:

- **(a)** an incorrect assignment of a verb tag to the Npred, e.g. *O Pedro teve* $tosse_{N/V}$ "Pedro had a cough"; in some cases, a contextual POS-disambiguation rule could be construed, either selecting the correct POS-tag or discarding the incorrect tag; in other cases, the incorrect tag is extremely rare in the language, so we could discard it as an "exotic" homograph;
- **(b)** an incorrect assignment of a preposition or definite article tag to *a* "to/the-fs", which produces either an incorrect PP chunk, e.g. *a infância* "the childhood": *O Pedro deixou a infância <para trás>* "Pedro left the [=his] childhood behind"; or an incorrect NP chunk, e.g. *a dieta* lit: "to diet" "on a diet": *O Pedro está a dieta* "Pedro is on a diet"; these cases could not be resolved for the moment.

(v) subtle interaction of lexical features with the chunking module of the parser: with determiner *um certo* "a certain", chunking rules fail to adequately identify the NP or PP headed by the Npred, when this can have a reading as a type of measure unit, a container, a group-of-things, or human collective noun (this semantic prototypes are encoded in the nouns lexical entries. For example, in the sentence *O Pedro tinha um certo comando dessa coisa* "Pedro had a certain command of that thing", the noun *comando* designates, among other things the abstract action noun “command”. However, the same noun could also designate the human collective noun, as in *o comando de mercenários* “the command of mercenaries’; in this situations, a contextual rule removes the spurious features, preventing the chunking rule to be triggered; and, finally,

(vi) certain nouns are used by the system’s local grammars to create complex NOUN nodes, which are relevant for Named Entity Recognition (NER) [30, 36]; for example, the noun *associação* “association” is often used to build Named Entities designating an organisation. For that purpose, the system produces a NOUN node, which leads to an inadequate chunking of the sentence. This issue has not been addressed yet: `680>TOP{NP{O Pedro} VF{fez} NP{a NOUN{associação de o João com o Rui}} .}`

At the deadline for the submission of this paper, most errors from the basic standard construction have been corrected (for CSV with 1 and 3 arguments). Work on 2 arguments, CSV examples continues but it has already dramatically decreased. For the remainder files, attention must be paid to [Passive] and Vop examples, responsible for most false-negative cases. The generation process for the remaining sentence types continues. Several transformations are yet to be formalised.

4 Conclusion and future work

This paper presented a method to generate examples of SVC directly from the linguistic properties encoded in these constructions Lexicon-Grammar, built for European Portuguese. Our focus here was on the transformations allowed by SVC, both the operations that are specific of this type expressions, and other operations with a broader scope, such as Passive. Though only 1/3 of the extant SVC have been processed so far, the system already generates over 48 thousand examples. The current distribution of the generated sentences per sentence type is likely to undergo significant changes, as much of the data already processed was derived from a subset of Portuguese SVC.

Special care was taken to make sentences as simple and as natural as possible. This includes producing adequate nouns for each syntactic slot, as well as choosing the best tense and word order. For commodity, examples from each sentence type are group in distinct output files. Preliminary observations confirm not only that most examples are perfectly natural, but having them systematically spelled out helps correct the linguistic data encoded in the Lexicon-Grammar matrix.

Several transformations still await an adequate representation in the LG matrix for example generation and SUPPORT dependency extraction. These include the alternation between 3-argument and 2-argument symmetric Npred, where the longer structure has a <cause> or <agent-cause> semantic role, e.g., *O Pedro fez uma mistura dessa coisa com aquela coisa* “Pedro made the mix of this thing and that thing”, while the 2 argument drops the subject of the longer sentence, e.g. *Esta coisa fez uma mistura com aquela coisa* “This thing made the mix with that thing”. A similar process occurs with nouns designating medical procedures, with a 3-argument, agentive subject CSV, e.g., *O Pedro fez um raio-X ao peito do João* “Pedro did an X-ray to João’s chest”; and an equivalent, apparently 2-argument, patient subject, *O João fez um raio-X ao peito* “João did an X-ray to the [=his] chest”.

In the future, once this step is finished, we would like to devise methods for populating the data base with data from corpora, using the information available as heuristics for corpus exploration. Also, other sentence types have not been considered yet, for example, sentences involving clefting, negation, etc. The interaction between current transformations and new sentence types to be produced will certainly make this work useful for testing, in a systematic way, the robustness of NLP systems when detecting SVC in texts.

References

1 S. Ait-Mokhtar, J. Chanod, and C. Roux. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144, 2002.

2 M. F. Athayde. Construções com verbo-suporte (funktionsverbgefuge) do português e do alemão. *Cadernos do CIEG Centro Interuniversitário de Estudos Germanísticos*, 1, 2001.

3 Jorge Baptista. Conversão, nomes parte-do-corpo e restruturação dativa. In Ivo Castro, editor, *Actas do XII Encontro da Associação Portuguesa de Linguística*, volume 1, pages 51–59, Lisboa, 30 de setembro a 2 de outubro de 1996, Braga-Guimarães, Portugal 1997. Associação Portuguesa de Linguística, APL/Colibri.

4 Jorge Baptista. *Sermão*, *tareia* e *facada*: uma classificação das expressões conversas *dar-levar*. *Seminários de Linguística 1*, pages 5–37, 1997.

5 Jorge Baptista. Construções simétricas: argumentos e complementos. In Olga Figueiredo, Graça Rio-Torto, and F. Silva, editors, *Estudos de homenagem a Mário Vilela*, pages 353–367. Faculdade de Letras da Universidade do Porto, 2005.

6 Jorge Baptista. *Sintaxe dos Nomes Predicativos com verbo-suporte SER DE*. Fundação para a Ciência e a Tecnologia/Fundação Calouste Gulbenkian, Lisboa, 2005.

7 Jorge Baptista. Viper: A Lexicon-Grammar of European Portuguese verbs. In Jam Radimsky, editor, *Actes du 31e Colloque International sur le Lexique et la Grammaire*, pages 10–17, République Tchèque, 2012. Université de Bohême du Sud.

8 Jorge Baptista. ViPEr: uma base de dados de construções léxico-sintáticas de verbos do Português Europeu. In Fátima Silva, Isabel Falé, and Isabel Pereira, editors, *Actas do XXVIII Encontro da APL - Textos Selecionados*, pages 111–129, Lisboa, 2013. APL/Colibri.

9 Jorge Baptista, Graça Fernandes, Rui Talhadas, Francisco Dias, and Nuno Mamede. Implementing European Portuguese verbal idioms in a natural language processing system. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives / Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*, pages 102–115. Proceedings of Conference of the European Society of Phraseology (EUROPHRAS 2015), June 28-July 2, 2015, Málaga, Spain, Geneva, Switzerland 2016.

10 Jorge Baptista and Nuno Mamede. Reciprocal Echo Complements in Portuguese: Linguistic Description in view of Rule-based Parsing. In Jorge Baptista and Mario Monteleone, editors, *Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013)*, pages 33–40, Faro, Portugal, September 10–14, 2013 2013. CLG'2103, Universidade do Algarve – FCHS.

11 Jorge Baptista and Nuno Mamede. *Dicionário Gramatical de Verbos do Português Europeu*. Universidade do Algarve, December 2020.

12 Jorge Baptista, Nuno Mamede, and Ilia Markov. Integrating verbal idioms into an nlp system. In Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago Pardo, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language*, volume 8775 of *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*, pages 251–256, Berlin, 2014. 11^{th} International Conference PROPOR'2014, October 8-10, 2014, São Carlos – SP, Brazil, Springer.

13 Eckhard Bick. Noun sense tagging: Semantic prototype annotation of a Portuguese treebank. In *Proceedings of TLT*, pages 127–138, 2006.

14 Nathalia Calcia. Descrição e classificação das construções conversas no português do Brasil. Master's thesis, Universidade Federal de São Carlos, São Carlos-SP, Brasil, 2016.

15 Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. Towards best practices for Multiword Expressions in Computational Lexicons. In *Proceedings of LREC'02*, pages 1934–1940, Las Palmas, Spain, May 2002.

16 Lucília Chacoto. *O Verbo Fazer em Construções Nominais Predicativas*. PhD thesis, Universidade do Algarve, Faro, 2005.

17 Noam Chomsky. *Remarks on nominalization*. Linguistics Club, Indiana University, 1968.

18 Mathieu Constant, G. Eryigit, Joana Monti, L. van der Plas, Carlos Ramisch, Michael Rosner, and A. Todirascu. Multiword expression processing: A survey. *Computational Linguistics*, pages 837–892, 2017.

19 Gloria Corpas Pastor, Ruslan Mitkov, Maria Kunilovskaya, and María Araceli Losey León, editors. *Processing European Portuguese Verbal Idioms: From the Lexicon-Grammar to a Rule-based Parser*, Malaga (Spain), September, 25–27 2019. Tradulex.

20 Maria Francisca Mendes Queiroz-Pinto de Athayde. *A estrutura semântica das construções com verbo-suporte preposicionadas do português e do alemão*. PhD thesis, Faculdade de Letras da Universidade de Coimbra, Coimbra, 2000.

21 Cláudia Dias de Barros. *Descrição e classificaçãoo de predicados nominais com o verbo-suporte FAZER: especificidades do Português do Brasil*. PhD thesis, Universidade Federal de São Carlos, São Carlos-SP, Brasil, 2014.

22 Graça; Fernandes and Jorge Baptista. Frozen sentences with obligatory negation: Linguistic challenges for natural language processing. In Carmen Mellado-Blanco, editor, *Colocaciones y fraseología en los diccionarios*, pages 85–96. Peter Lang, Frankfurt, 2008.

23 Ana Galvão, Jorge Baptista, and Nuno Mamede. New developments on processing European Portuguese verbal idioms. In Carlos Augusto Prolo and Leandro Henrique Mendonça de Oliveira, editors, *12th Symposium in Information and Human Language Technology*, pages 229–238, Salvador, BA (Brazil), October, 15–18 2019.

24 Gaston Gross. *Les construction converses du français*. Droz, Genève, 1989.

25 Maurice Gross. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 15(63):7–52, 1981.

26 Maurice Gross. Methods and tactics in the construction of a lexicon-grammar. In *Linguistics in the Morning Calm, Selected Papers from SICOL*, pages 177–197, Seoul, 1988. Hanshin Pub. Co.

27 Maurice Gross. *Grammaire transformationnelle du français: 3 - Syntaxe de l'adverbe*. ASSTRIL, Paris, 1996.

28 Maurice Gross. Lexicon-grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge, 1996.

29 Alain Guillet and Christian Leclère. Restructuration du groupe nominal. *Langages*, 15e année(63):99–125, 1981.

30 Caroline; Hagège, Jorge; Baptista, and Nuno João Mamede. Reconhecimento de entidades mencionadas com o xip: Uma colaboração entre o inesc-l2f e a xerox. In Cristina; Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM (Aveiro, 11 de Setembro de 2008)*. Linguateca, 2009.

31 Zellig Harris. The elementary transformations. In Henry Hiz, editor, *Papers on Syntax*, pages 211–235. D. Reidel Publishing Company, 1964.

32 Adam Kilgarriff, Pavel Rychly, Vojtech Kovar, and Vit Baisa. Finding multiwords of more than two words. In *EURALEX*, Oslo, 2012.

33 Béatrice; Lamiroy. Le lexique-grammaire. In *Travaux de Linguistique*, volume 37. Duculot, 1998.

34 Christian Leclère. Sur une restructuration dative. *Language Research*, 31:179–198, 1995.

35 Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In Alberto Abad, editor, *International Conference on Computational Processing of Portuguese (PROPOR 2012) - Demo Session*, Coimbra, Portugal, April, 17–20 2012.

36 Diogo Oliveira. Extraction and classification of named entities. Master's thesis, Instituto Superior Técnico - Universidade Técnica de Lisboa Universidade Técnica de Lisboa, L^{2}F/INESC-ID, Lisboa, 2010.

37 Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, et al. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, 2018.

38 Elisabete Ranchhod. *Sintaxe dos predicados nominais com "estar"*. Instituto Nacional de Investigação Científica (INIC), 1990.

39 Amanda Rassi, Nuno Mamede, Jorge Baptista, and Oto Vale. I. Integrating support verb constructions into a parser. In *Proceedings of the Symposium in Information and Human Language Technology (STIL'2015)*, pages 57–62, 2015.

40 Amanda Rassi, Cristina Santos-Turati, Jorge Baptista, Nuno Mamede, and Oto Vale. The fuzzy boundaries of operator verb and support verb constructions with dar "give" and ter "have" in Brazilian Portuguese. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), COLING 2014*, Dublin, Ireland, August 2014. Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), COLING 2014, Dublin, August 24, 2014, COLING 2014.

41 Amanda P. Rassi. *Descrição, classificação e processamento automático das construções com o verbo* dar *em português brasileiro*. PhD thesis, Universidade Federal de São Carlos, São Carlos-SP, Brasil, 2015.

42 Amanda. P. Rassi, Jorge Baptista, and Oto Araújo Vale. Um corpus anotado de construções com verbo-suporte em português. *Gragoatá*, 39(1):207–230, June, 2015 2015.

43 Amanda. P. Rassi, N. P. Calcia, Oto A. Vale, and Jorge Baptista. Análise comparativa das construções conversas em português do brasil e português europeu. In *Abstracts from I Congresso Internacional de Estudos do Léxico e suas Interfaces (CINELI)*, page 45, Araraquara, SP (Brazil), March 2014. Congresso Internacional de Estudos do Léxico e suas Interfaces (CINELI), FCL-UNESP.

44 I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, volume 2276 of *LNAI/LNCS*, pages 1–15, Berlin, 2002. 3rd International Conference CICLing-2002, Springer.

45 Cristina Santos. *Construções com verbo-suporte* ter *no Português do Brasilrasil*. PhD thesis, Universidade Federal de São Carlos, São Carlos-SP, Brasil, 2015.

46 Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, and Ivelina Stoyanova. The PARSEME shared task on automatic identification of verbal multiword expressions. In *13th Workshop on Multiword Expressions (MWE)*, pages 31–47, 2017. `doi:10.18653/v1/W17-1704`.

47 Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, et al. PARSEME–PARSing and multiword expressions within a European multilingual network. In *Multiword expressions at length and in depth: Extended papers from the MWE workshop*, 2015.

48 Rui Talhadas. Semantic Role Labelling in European Portuguese. Master's thesis, Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Faro, Portugal, 2014.

49 Rui Talhadas, Jorge Baptista, and Nuno Mamede. Semantic roles annotation guidelines. Technical report, L2F/INESC ID Lisboa, 2013.

50 Rui Talhadas, Nuno Mamede, and Jorge Baptista. Semantic Roles for Portuguese Verbs. In Jorge Baptista and Mario Monteleone, editors, *Proceedings of the 32nd International Conference on Lexis and Grammar (CLG'2013)*, pages 127–132, Faro, Portugal, September 10–14, 2013 2013. CLG'2103, Universidade do Algarve – FCHS.

51 Aldina Vaza. Estruturas com nomes predicativos e verbo-suporte dar. Master's thesis, Faculdade de Letras da Universidade de Lisboa, Lisboa, Portugal, 1988.

52 Nicolas Zampieri, Carlos Ramisch, and Géraldine Damnati. The impact of word representations on sequential neural MWE identification. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN)*, pages 169–175, 2019. `doi:10.18653/v1/W19-5121`.