Vocatives in Portuguese: Identification and Processing*

Jorge Baptista[1] and Nuno Mamede[2]

1 **University of Algarve/FCHS & INESC-ID Lisboa/L²F, Faro, Portugal**
jbaptis@ualg.pt
2 **University of Lisboa/IST & INESC-ID Lisboa/L²F, Lisbon, Portugal**
Nuno.Mamede@l2f.inesc-id.pt

Abstract

This paper describes the most salient linguistic aspects of vocative constructions in Portuguese, with special reference to its European variety. Next, the paper presents the strategy followed for implementing this linguistic knowledge in a computational grammar of Portuguese, developed for the natural language processing chain STRING and using the XIP rule-based parser. Very precise and detailed linguistic descriptions can be implemented in this way.

1998 ACM Subject Classification I.2.7 Natural Language Processing/Text analysis

Keywords and phrases Natural Language Processing, Text analysis, Portuguese, Vocative, Parsing

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.22

1 Introduction

This paper deals with vocative constructions in Portuguese. This is the case of the initial phrases in the sentence (facultative elements in brackets): *(Ó) (meu caro) João/amigo, não faças isso!* '(Hey) (my dear) John/friend, don't do that!´ In these examples, those phrases are traditionally analysed as having the syntactic function of *vocative*, that is, phrases that are somewhat marginal to the main sentence, and that are used by the speaker to address his interlocutor. These phrases are *not* the subject of the sentence main verb, which is a dropped 2^{nd}-person-singular pronoun *tu* 'you'. The different forms that this interpellation can take are related to different sociocultural values, which may reflect, for example, in verbal inflection, for example, in the opposition between the form of treatment by *tu* 'you_2^{nd}-sg.' and *você* 'you_3^{rd}-sg.'), as in: *(Ó) Dr. João, não faça isso!* '(Hey) Dr. John, dont do that!´ However, the vocative construction may also serve to express other pragmatic values such as the expression of *affection*: *Minha coisinha fofa, não faças isso!* 'My little fluffy thing, don't do that!´ or an *insult*: *Minha grandecíssima besta, não faças isso!* 'My most-great beast, do not do this!´

In some communicative situations and certain textual genres, the use of vocative is relatively frequent, such as in dialogues, in the (formulaic) opening of epistolary texts, or at the onset of formal addresses, speeches and lectures. In the later, the speaker or the lecturer usually addresses the official entities present at the venue, using a strict protocol in their ordering and in their designation. Such lists can be quite extensive and usually

* This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

end with a generic vocative, using the formula *minhas senhoras e meus senhores* 'ladies and gentlemen'. Because of the cultural weight and the formulaic nature of such vocatives, these expressions are highly idiomatic and cannot be translated literally (v.g. **my ladies and my gentlemen*). In the previous example, the nouns in the formula are determined by a (facultative) possessive pronoun.

The communicational and institutional settings of the address or speech determine the use of specific vocative formulae or special combinations of vocatives. For example, in the Portuguese Parliament, the speaker will first address the president (chair) and then the members of the Parliament, v.g., *Senhor presidente, senhores deputados* 'Mister president, (ladies and gentlement)'. The host of a circus show will address the audience by *respeitável público* (lit: 'respectable public/audience') and will also include the children in the vocative *senhoras e senhores, meninos e meninas* 'ladies and gentlemen, boys and girls'. In a public address (PA) announcement on a supermarket, the speaker will address the clients by *estimados clientes* 'esteemed clients'. Other factors, such as language variety, also determine the choice of the adequate formulae. In Brazilian Portuguese, an epistolary vocative would naturally choose the adjetive *prezado* 'esteemed' rather than the European Portuguese *caro* 'idem' or 'dear'.

From the point of view of automatic syntactic analysis (or parsing), the precise identification and adequate linguistic analysis of this type of expressions is relevant, since: (i) they should not be analysed as fundamental constituents of the sentence in which they occur; (ii) they can serve as antecedents of anaphors in the subsequent text; (iii) they can help to determine the structure of a dialogue, namely the turn-taking or an exchange of roles between the dialogue participants, and (iv) they express different pragmatic values, as we have already seen, and their formulaic and often idiomatic nature can make them very hard to translate automatically.

This paper describe one of the modules of the Portuguese computational grammar developed for the STRING system [11] was implemented and evaluated in order to adequately process vocatives in unbound texts. The immediate motivation for this study was the need to process transcripts of public speeches and addresses delivered by various official entities in various formal contexts (the national parliament, municipal councils, etc.). This module aims, therefore, at the identification, delimitation and automatic syntactic analysis of constructions of vocative in Portuguese, with special reference to the European variant. The module is integrated into the rule-based computational grammar developed for the XIP (Xerox Incremental Parser) [1]).

This article is organized as follows: The following Section 2 presents the theoretical framework, trying to exemplify and discuss different situations, to determine a place for the vocative in the Portuguese grammar. Secondly, some parsing issues resulting from an inadequate analysis of the vocative constructs are identified and illustrated with examples taken from different syntactic parsers of Portuguese (Section 3). At the same time, a brief survey will be made into syntax dependency coding schemes developed for various languages, identifying the way the vocative is typically framed. The next section (4) presents the solution developed in the framework of a Portuguese grammar, using the declarative rules of the XIP formalism. A preliminary evaluation of this module performance is provided (Section 5). The paper concludes with some perspectives for the development of current work and future applications.

2 Theoretical framework

This section tries to determine the place of the vocative constructions in Portuguese grammar, by way of reviewing some reference work. Naturally, it is outside the scope of this paper to do a systematic survey of the phenomenon, as it seeks only to raise its most salient aspects and to discuss, if only briefly, the controversial topics that this grammatical category raises.

Traditionally, *vocative*[1] is one of the major cases of nominal declension (inflection) of several natural languages, along with other casual values (v.g. nominative, accusative, genitive, etc.) in which nouns (and other categories) can inflect. The *case* is thus seen as a morphological variation identified with the syntactic function that the affixed element performs in the sentence: *nominative* for the *subject* function, *accusative* for the function of *complement*, etc. Perhaps for this reason, the vocative has usually been integrated into the set of main syntactic functions, alongside the subject or the direct complement [7, p. 160-161]. Portuguese, however, has dropped the Latin case system (except for personal pronouns), and has no morphologic marker corresponding to the vocative case, which make identification of vocatives more difficult. For some authors [4], no particular syntactic function or dependency is defined corresponding to the vocative, though mention to such situations is unavoidable when dealing with (direct) imperatives (*idem*:p. 457-458), e.g. *Tu/Maria, empresta-me esse livro! Você/O senhor, arrume o carro* (examples taken from the authors). In this framework, a noun phrase designating the subject of imperatives is given no particular status, and it is only said "to be interpreted as a vocative and, consequently, occurs in a peripheral position in the sentence" (our translation), irrespective of its left, pre-verbal or right position in the sentence (no mention is made about the mobility of vocatives *within* the sentence).

On the same line, the authors of the more recent *Gramática do Português* [14, vol.1, p. 351 ff.] also do not mention any vocative syntactic function, but prefer to treat the matter within the description of proper nouns [14, vol.1, p. 1013 ff.]. The authors of this chapter consider the vocative as a "semantic-discursive function", though no definition of the concept seems to be provided, nor its articulation with the remainder of the grammar architecture. Still, from their (short) descriptive approach, the authors mention, among other aspects, the particular behaviour that proper nouns can present in various functions, namely the interdiction of definite article in vocatives (e.g. **(Ó) **o** João, não faças isso* '(Hey) *the* John, don't do that'). Finally, from their description, it is possible to infer that, in the authors' framework, (i) vocative should not be included in the set of syntactic relations, which hold between the constituents of the sentence/clause; and (ii) that the grammatical role/value of vocatives should be placed at the level of discourse analysis, in the broader context of language communicative functions.

On another perspective [9, p. 351 ff.][10, p. 135 ff.], though not explicitly addressed, vocatives find a natural framework within the description of *performatives* (like *say* and *ask*/*order*), which are consider as operators underlying direct discourse, namely declarative and interrogative/imperative sentences, respectively. Based on evidence of a zeroed subject in direct imperative sentences (e.g. [*You*] *wash yourself!*, where the reflex *-self* result from the repetition of the subject *you* as complement of the verb *wash*), an underlying performative is considered in these sentences (e.g. *I ask/order you that you wash yourself*), which, when zeroed, yields the interrogative sentence-type intonation .

A similar approach could be adopted for the vocative-appellative (in fact, it is even hinted at by [10, p. 139], about sentences like *You (there)!*). In this perspective, the vocative

[1] Sometimes also referred to by terms such as *appellative* or *conative*, though not necessarily in this perspective. For a general overview, refer to [2].

would constitute another case of performative operators (such as *chamar* '*address*' or '*call*'), having the speaker as its subject and the addressee as its complement, in much the same way as the other performatives. (The so-called interjection *ó* 'hey' could be treated as a specialised – but facultative – marker of the vocative in direct speech, in much the same way as an argument-marker preposition.) In this sense, utterances involving vocatives should be analysed as complex sentences, consisting of, at least, two performatives: the vocative proper, with its appellative discursive function (addressing the hearer/reader); and the second sentence (a statement, an order/request, etc.).

This analysis sorts out the problem of having an expression in a sentence that is not necessarily linked to the constituents of the remaining content of that sentence, though it can be the antecedent of anaphoric expressions (as noticed by [7, p. 161]). It also addresses the issue of vocatives showing a peripheral status, and their corresponding mobility within the sentence. Being the result of a performative (and its reduction), the vocative would not hinge (or depend) on any other sentence constituent, but rather it would be linked to the topmost sentence node in the parsing tree, in much the same way as sentence-modifying adverbs [12].

Vocatives, however, are also used to express other (pragmatic) meanings in addition to the barest appellative function, namely to produce an affective value (insult/politeness), in which the choice of the appellative (a proper name, a pronoun, a common noun invested with affective value, including profanity words) plays a major role, particularly, among other distinctions, the *tu*/*você* opposition. For lack of space, we do not pursue further in this paper the linguistic description of vocatives and the theoretical implications of the approach outlined above. These involve, for instance, the adequate analysis of so-called interjections such *ó* and *pá*, dialectal variation in the use of modifiers (*caro*/*prezado*/*estimado* 'dear'), profanity words and adjectival predication, mobility within sentence, use of politeness adverbs (e.g. *por favor*), etc.

For comparison, vocatives have been encoded in the annotated AnCora [16] *corpora* of Spanish and Catalan texts (approx. 500K words each), though its frequency is very low (13 and 7 instances, respectively). In the documentation of the *Floresta Sintá(c)tica* treebank[2], regarding the revised portion (*Bosque*), only 29 instances of vocative have been reported.

One could also refer to the proposals of de Nivre *et al.* [13][3] for a set of so-called "universal dependencies" having in view their implementation in the Stanford parser. Their proposal concerning vocatives points to a similar solution as far as the vocative dependency is concerned. However, though the definition of the vocative dependency in [13] seems to be clear enough:

> The vocative relation is used to mark dialogue participant addressed in text (common in emails and newsgroup postings). The relation links the addressee's name to its host sentence. A vocative commonly co-occurs with a null subject, [...]. If the nominal is clearly vocative in intent, the preference is to use the vocative relation.

It is not clear that the vocative element is linked to the host sentence as a whole, probably because this parser's formalism does not make use of a TOP/ROOT node, and the place of insertion of the vocative seems to vary depending of the element that fills the topmost slot in the parse tree (see Fig. 1).

[2] http://www.linguateca.pt/Floresta/BibliaFlorestal/anexo4.html
[3] http://universaldependencies.org/

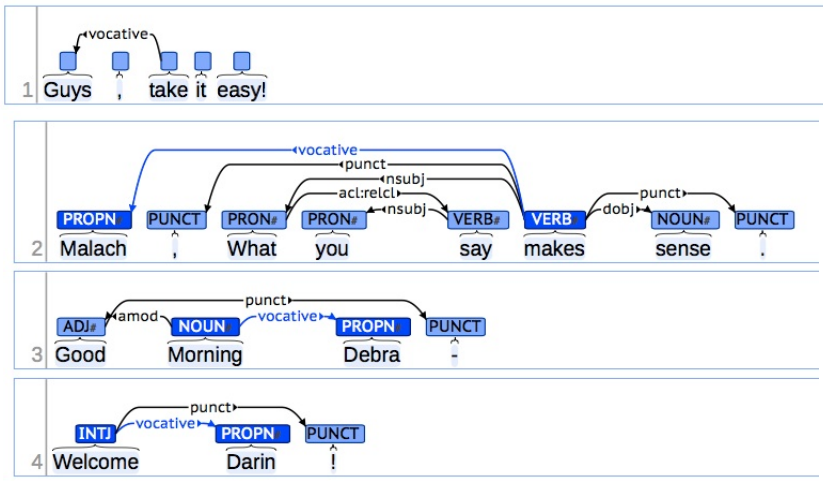

Figure 1 Examples of the *vocative* dependency in the proposal of a set of universal dependencies for the Stanford parser.

Thus, in the first example (*Guys, take it easy!*) the vocative hinges on the main verb (in the imperative); in the second sentence (*Malach, What you say makes sense*), the vocative also depends on the main verb; however, in the third example (*Good Morning, Debra*), as the parser analyzes the interjective *good morning* as an ordinary noun phrase, the vocative is made to depend on the noun *morning*; for the fourth sentence (*Welcome Darin!*), where *Welcome* is POS-tagged as an interjection, the proper name is made to depend on the only remaining element of the sentence. It should also be noted that, in spite of the definition of vocative dependency explicitly mentioning "the addressee's name", the first example involves a common noun. Thus, preference for vocative is determined for other types of noun phrases, as long as "clearly vocative in intent". Unfortunately, we were not able to find any evaluation of the Stanford parser concerning this vocative dependency, nor, to the best of our knowledge, how this proposal within a set of "universal dependencies" has been tackled for other languages.

In the next section, some parsing issues related to the processing of vocatives will be addressed.

3 Parsing issues

In this section, we look into the way that publicly available demo versions of Portuguese parsing systems treat a sample of clear-cut cases of the use of vocative. For sake of brevity, only a short commentary and not a detailed analysis can be made here. The performance of two systems was considered: (i) the LX-Suite [6], which uses the MaltParser[4]; and (ii) the Visl (Visual Interactive Syntax Learning) system[5], based on the parser *Palavras* [3]. A sample of sentences was used, exploring same variation factors, namely, the presence/absence of the interjective *ó* 'hey' in front of a proper noun, at the beginning of an imperative sentence and separated from it by comma, as well as the facultative use of affective modifiers *meu caro* 'my dear', v.g. (*Ó*) (*meu caro*) *Rui, não faças isso!* 'Hey my dear Rui, don't do that'.

[4] http://www.maltparser.org/parser
[5] http://visl.sdu.dk/visl

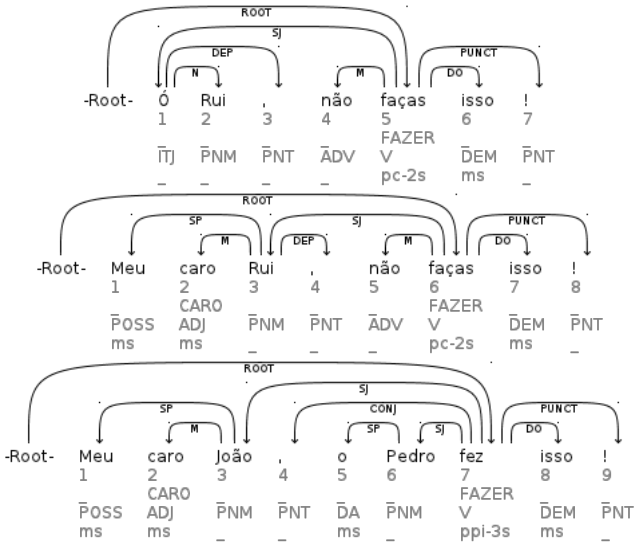

Figure 2 Sentences parsed by the Lx-Suite using the MaltParser; top: imperative sentence with *ó*; center: imperative sentence without *ó*, but with a possessive and an adjective *meu caro* modifying the noun; bottom: a declarative/exclamative with the same type of vocative as above.

As far as we could ascertain from its documentation[6], the Lx-Suite parser does not seem to include a specific dependency relation in order to capture vocatives. Therefore, the remarks below aim only at highlighting parsing issues the lack of this dependency entails. As it can be seen, in Fig. 2 (top), the proper noun *Rui* is parsed as a sort of complement of the interjection (`N` dependence: nominal modifier?, apposition?), which, in turn, is analysed as the subject (`Sj`) of the imperative. This is arguably not an adequate parse. In the second sentence (center), *Rui* is parsed as the subject of the main verb. Since these are imperative sentences, and in the absence of a vocative dependency, this formal description may not be entirely inadequate, as the vocative and the (zeroed) subject of the verb in the imperative are necessarily co-referent. However, this co-reference is not obligatory in other sentence types, and, in the absence of a vocative dependency, systems may produce inadequate analyses in such cases. Thus, when parsing a similar, but declarative (or exclamative) sentence, with a vocative non co-referent to the main verb subject (below), the results are less then adequate. In this sentence, the vocative *João* is incorrectly parsed as subject of the main verb, in much the same way as the correct subject *Pedro*.

In its turn, the Visl system parser integrates the vocative dependency (`@VOK`), as shown in Fig. 3. Still, it seems that the vocative is triggered only in the presence of the interjection *ó* (left), which is linked directly to the topmost `root` note of the parse tree (by an `@ADVL` dependency) and apparently unrelated to the noun phrase *meu caro João* 'my dear João'. Without this lexical cue (right), no vocative is extracted, and the noun phrase is linked to the verb by a `@TOP` dependency. According to its definition[7], this `@TOP` dependency seems to have been construed for another type of relation, namely, instances of topicalization, such as those provided in the examples of the symbol set manual, v.g. *A Maria, não quero convidá-la* 'Maria, [I] don't want to invite her' (object topic); *Esse rapaz, ele sabe dançar* 'That boy, he

[6] `http://nlxserv.di.fc.ul.pt/depparser/intro_en.html`
[7] `http://visl.sdu.dk/visl/pt/info/symbolset-manual.html`

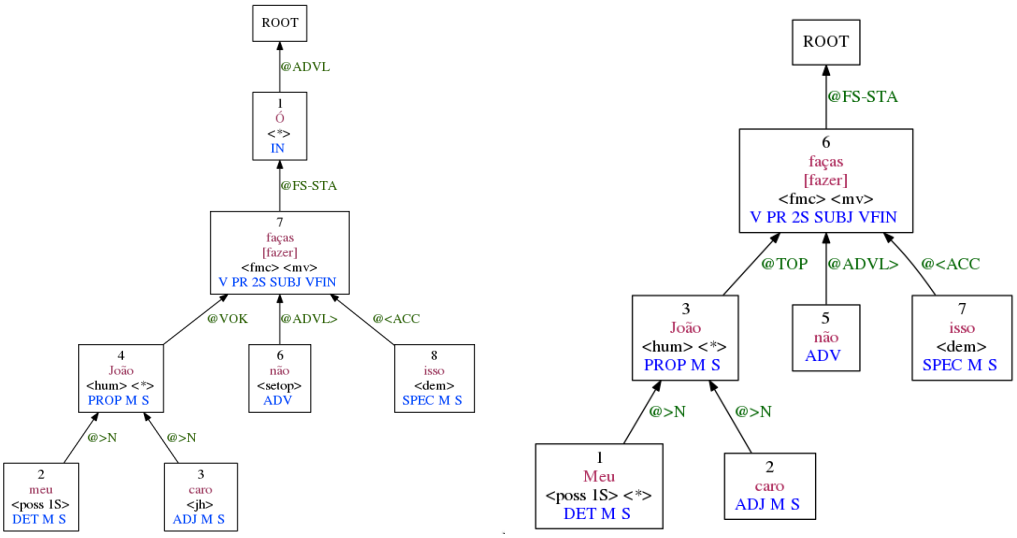

Figure 3 Sentences parsed by the VISL system, based on the PALAVRAS parser [3]. Left: imperative with *ó*: *Ó meu caro João, não faças isso* 'Hey my dear João, don't do that'; Right: the same sentence without *ó*.

knows how to dance' (subject topic). Again, as with the LX-SUITE parser, in a declarative sentence, the VISL system, also extracts two subjects (Fig. 4, right). Apparently, there is no rule preventing two (heads of) constituents to be analysed as having the same syntactic function. Still, the position of the vocative detached by comma at the end of the sentence (Fig. 4, left) is sufficient to correctly extract the vocative dependency. In this last case, with the vocative detached by comma at the end of the sentence, the LX-SUITE parser still does not produce adequate results (Fig. 5), since it deals with the vocative as a specifier (`SP`), in the same way as the definite article in front of a proper name (v.g. *O Pedro*).

Since PALAVRAS [3] is a rule-base parser, its development is not hindered by the scarcity of occurrences of a given linguistic phenomena in texts, even in large *corpora*, allowing for a very precise description of the grammar of the language. In this sense, our approach to the phenomenon of vocatives, which will also use a rule-based parser, will be similar. However, we believe that a more comprehensive linguistic description of vocative constructions is necessary, in order to obtain a larger coverage of the phenomena. On the contrary, as the language model of the LX-SUITE parser is built upon previously annotated *corpus*, the sparseness of the phenomenon may have had an impact on its performance, regarding this specific grammatical aspect.

In the next section, we present the strategy adopted by the STRING (undisclosed ref.), using the rule-based parser XIP , to parse vocatives in Portuguese.

4 Grammar module

Having in mind the linguistic phenomena briefly sketched in Section 1 and the theoretical framework outlined in Section 2, we set out to produce a new module of the dependency

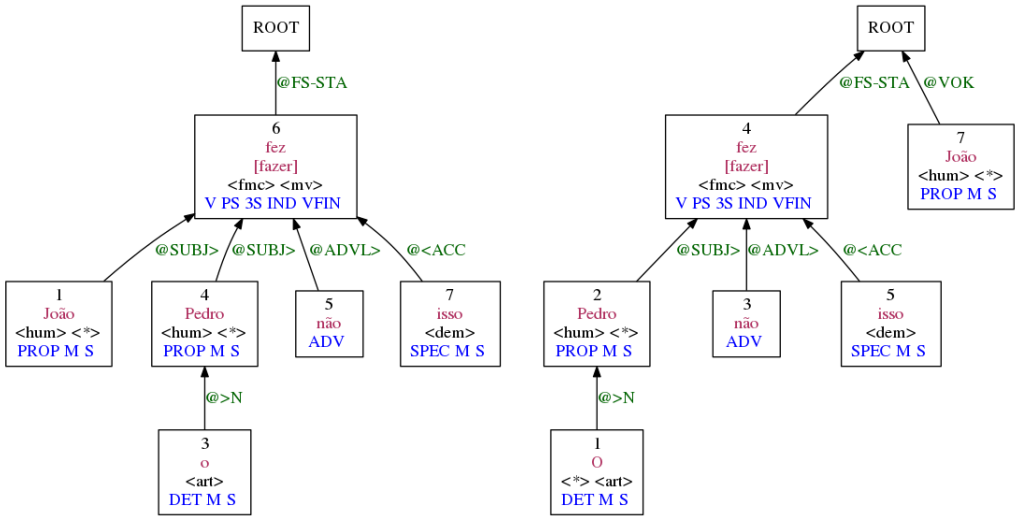

Figure 4 Declarative/exclamative sentences parsed by the VSIL system, based on the PALAVRAS parser [3]. Left: the vocative at the begining of the sentence: *João, o Pedro fez isso.* 'João, Pedro has not done that'; Right: the same sentence but with the vocative at the end: *O Pedro fez isso, João.* 'Pedro has not done that, João.'.

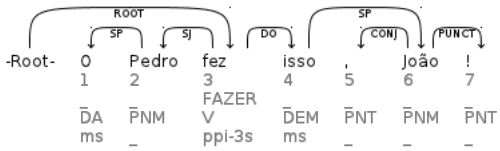

Figure 5 Declarative/exclamative sentence parsed by the LX-SUITE using the MALTPARSER: vocative detached at the end.

grammar of XIP by building a batch of artificial examples[8], inspired in instances of vocatives found in corpora, and by systematically exploring the formal variation such examples permitted, namely: (i) the presence or absence of interjection *ó* 'hey' in the vocative phrase; (ii) the position of the vocative at the beginning or at the end of the sentence, or in the middle of the sentence and detached by commas; (iii) the use of proper names, both a single noun, as in *João*, and in longer strings of nouns (and prepositions), as in *João da Silva*; (iv) different titles (*engenheiro* 'engineer') and profession nouns (*professor*), used in isolation or in combination (*professor doutor* 'doctor'), eventually abbreviated (*Eng.º* 'engineer'), as well as when combined with proper names or preceded by the respectful addressing form (*senhor* 'sir'); (v) a closed set of adjectives conventionally used in vocatives (*caro*, *estimado*, *excelentíssimo*, *prezado*, *querido*, etc. (vi) the possible presence of an affective possessive, obligatorily in the 1^{st}person-singular (v.g. *meu*), in all its gender-number values; (vii) a small number of expressions used mostly in vocatives, such as *cavalheiro*, *pá* 'man', *vossa excelência* 'your excelency', as well as certain conventional combinations with the form of two coordinated nouns: *senhoras e senhores* 'ladies and gentleman'.

The general strategy to extract the `VOCATIVE` dependency consists in identifying first the vocative phrases, based on their content and context, in order to mark them with a

[8] These sentences annotated with the `VOCATIVE` dependency can be retrieved from: `https://string.l2f.inesc-id.pt/w/images/1/16/Vocative.txt`.

`vocnp` (vocative noun phrase) feature. This feature is, then, used by a general parsing rule that extracts the `VOCATIVE` dependency. The first part of the process is handled by a *local grammar* (LG) of vocatives, specifically built for this purpose, which takes the form of declarative rules, describing a pattern that must be matched so that the `vocative` feature is added to the phrase node. For example, the following is a LG rule:

```
1> NP[vocnp=+] = ?[lemma:senhor,start], noun | ?[comma] | .
```

that adds the `vocnp` feature to a noun phrase at the begining of a sentence starting with the lema *senhor*, followed by a `noun`, and detached from the remainder of the sentence by a comma. We considered the presence of commas as a necessary context to trigger the local grammar of vocatives. This is not always the case in real texts, where the use of commas is subject to much individual variation. The `noun` is the result of a previous processing stage, the *chunking*, where some sequences of words, like strings of proper names (e.g. *João da Silva*) have been already grouped together as a single noun. The `VOCATIVE` dependency is extracted by the rule:

```
| #3[cat:0]{?*, NP#1[vocnp];PP#1[vocnp] } |
 if ( HEAD(#2,#1) & ~VOCATIVE(#3,#1) )
   VOCATIVE(#3,#2)
```

This rule reads as follows: When a noun phrase (`NP`) or a prepositional phrase (`PP`) (see below, the treatment given to `PP` phrases introduced by *ó*) marked as a vocative phrase (`vocnp`) is found, for which a `VOCATIVE` dependency as not been previously extracted, then that very dependency is extracted between the phrase head and the top node of the sentence (`[cat:0]`).

The vocative extracting rules operate at an earlier stage of the parsing process, before the other major dependencies (subject, direct complement, modifier, etc.) are extracted. Hence the rules that had been already built so far in order to extract these dependencies had to be modified by adding the condition that such constituent had *not* been previously parsed as a vocative. As one can see from the above, the bulk of the parsing is carried out by the *local-grammar* (LG) for vocatives. In the remainder of this section, we provide further details on the vocatives aimed at by the LG.

In most of its occurrences the interjective *ó* is used in combination with a noun phrase to form a vocative in an unambiguous way, This a rule-of-thumb, that explains much of the positive results from the VISL system, as shown in Section 3. In view of this regularity, we decided to chunk those combinations as a special type of prepositional phrase `PP` and mark it with the `vocnp` feature, irrespective of the content of the element appearing after the interjection, as shown in the chunking rule below:

```
> PP[vocnp=+] = interj[lemma:ó], NP.
```

These lead us then to determined other combinations and contexts of *ó* that do not constitute vocatives. For example, the (incorrect?) combination *ó quê?* 'or what?' has been treated as a compound (=multiword) interrogative interjection, whose canonical form is *ou quê?*, e.g. *Está tudo bêbado, ó quê?* 'Is everyone drunk, or what?'. In a similar way, when immediately followed by the interrogatives *quão* 'how [much]', the *ó* is being (incorrectly?) used instead of interjection *oh*. In this case, *ó* does not form a vocative, instead it contributes to the exclamative nature of the sentence, e.g. *Ó quão grande é o mundo!* 'O how large is the world!'. The same happens in the combinations *ó, sim!* and *ó, não!*, used instead of *oh, sim!* and *oh, não!*, and often without the intervening comma.

Due to its syntactic independence from the main clause, vocatives show a remarkable mobility within the sentence, though their most common position in the Portuguese *corpora* we consulted was clearly the beginning of the sentence and, less frequently, some position within the sentence. Cases of vocatives at the end of the sentence are relatively rare. Strings of vocatives found at the beginning of a speech/lecture (as in the Parliament discourses) were treated in the same way as a single vocative at the beginning of an ordinary sentence, e.g. *Sr. Presidente, senhoras e senhores deputados* 'Mister President, ladies and gentlemen (members of the Parliament)'. In this case, each vocative is linked independently to the `TOP` node of the sentence. The same applies to the initial salutation and forms of addressing the reader at the onset of a private correspondence, e.g. *Minha querida esposa* 'my dear wife', *Caríssimo amigo* 'most dear friend', even if in a separate line from the main text.

The complex `noun` chunks, formed with combinations of proper names, profession nouns and honorific titles, are identified by using, on one hand, a large lexicon of given names and surnames, nouns for professions and lists of titles, including abbreviations, and, on the other hand, a specific set of local grammars, that had been previously built for the processing of named entities [8].

A remarkably small set of adjectives is often used conventionally for vocatives. These include the relatively formal *caro* (and its Brazilian correspondent *prezado*) and the much more intimate *querido*; the superlative *caríssimo* is also used; other, less frequent adjectives, are: *belo*, *bom*, *doce*, *grande*, *ilustre*, *lindo*, *pobre*, *rico*, *sábio*, *santo*, *terno*, *triste*, *velho*, etc. To simplify the process of rule building, the most frequently occurring of these adjectives as found in the *corpus* were listed and they were given a new feature (`vocadj`). In the same way, a specific set of nouns are particularly apt to constitute vocatives: *amigo*, *amor*, *bem*, *camarada*, *colega*, *gente*, *leitor*, *menino*, *rapaz*, *senhor*, *súbdito*, *velho*, etc. The most frequent of these nouns were also listed and given a new feature: `vocnoun` (vocative noun). These lists are open, an can be extended at will. Besides this list of nouns, several subsets of nouns in the lexicon are being systematically added this `vocnoun` feature: titles, professions, family relations, etc. Notice that these adjectives can also modify proper names in a vocative: *caro João*, *querida Rita*. As mentioned above, a possessive modifier can also be combined, e.g. *meu caro amigo*, *minha querida Ana*. This possessive refers to the speaker, hence it can only be a 1^{st}-person-singular, v.g. *meu*, *minha*, *meus*, and *minhas* 'my'. Notice that this person-number can also be used in insults, along with a 3^{rd}-person-singular (*seu*, *sua*, *seus*, *suas*): '*meu/seu imbecil*'. This type of ambiguity was not addressed at this point. Below, one can see the form of two rules describing such combinations involving possessives, `vocadj` and `vocnoun`, at the beginning of the sentence:

```
1> NP[vocnp=+] = pron[poss,poss1s,start], (adj[vocadj]),
   noun[vocnoun];noun[human,proper] | ?[comma] | .
1> NP[vocnp=+] =  adj[vocadj,start],
   noun[vocnoun];noun[human,proper] | ?[comma] | .
```

Very specific and conventional combinations were left out of this subsets, and they were treated individually. For example, combinations such as *estimado cliente*, *respeitável público*, where the adjective is obligatory, and the possessive is not allowed, v.g. **meu estimado cliente*, **meu cliente*, **cliente*; **meu respeitável público*, **meu público*, **público*. Slightly less constraint is the noun *cavalheiro* 'gentleman', which can be facultatively modified by an adjective, *excelentíssimo cavalheiro* 'most excellent gentleman', but not by the possessive, **meu cavalheiro*.

We also mention the case of the formal way of addressing, *vossa excelência* 'your excellency', and the corresponding plural *vossas excelências* 'your excellencies', which is an excepcional

use of the possessive 2^{nd}-person-plural. The possessive, however, is facultative. Though not alway used as a vocative, e.g. it can be reasonably identified given a detached context (i.e. separated by the sentence by commas).

The so-called interjection *pá* is always used as a vocative, sometimes preceded by the true interjections *ó*, *oh* and *he*, e.g. *Não , pá, não é isso.* 'No, man, it's not that.'; *Ó pá, acho que devíamos seguir o conselho ... mas, eh pá, isso vive muito da altura* 'Hey man, [I] think that [we] should follow the advice... but, hey man, that depends very much of timing' (real examples taken from *corpora*).

Another form of vocative, similar in use to *pá* (very much colloquial but typical of youngsters), consist in an isolated possessive pronoun, with the same inflection restrictions as described above, in a detached context, e.g. *Vi-te no cinema, meu.* 'I saw you in the movies, man.' Curiously, though there is no linguistic reason why it should not be so, the feminine forms of this possessive have not been found in written *corpora*.

5 Evaluation

As mentioned before, only 29 instances of vocative are reported in the documentation of the *Floresta Sintá(c)tica* treebank[9], regarding the revised portion (*Bosque*). To the best of our knowledge, there is no other available *corpus*, annotated with vocative constructions in Portuguese. Therefore, a procedure had to be devised to evaluate the precision of the STRING's parser XIP in the extraction of the `VOCATIVE` dependency. This section outlines the procedure adopted to build a reasonably sized sample of examples of vocatives and describes a preliminary evaluation of this module of the Portuguese grammar built in XIP [1].

To this end, real sentences were retrieved from the CetemPúblico *corpus* [15][10] using Linguateca's interface. The sentences were extracted using commas or full stops as boundaries of the targeted expressions, trying to capture vocatives at the beginning, the end and in the middle of the sentence, that is, detached from the main sentence by commas. The targeted expressions were represented in the queries by part-of-speech alone, except for the possessive 1^{st}-person-singular (*meu* 'my') and its gender-number inflected forms. A distinction was made between proper and common nouns. Specific patterns were queried, namely, the isolated possessive, the interjections *pá* and *ó*. Once retrieved, duplicates were removed and the *corpus*' extracts were randomly sorted. Only the first 100 sentences from each pattern were kept for the evaluation.

These patterns, though constrained by the contextual delimiters with which they were defined, are broad in definition, so they can arguably be used to calculate the precision of the parser and – in an approximate way and only for those patterns, of course – its coverage/recall. At this stage of the grammar's development, it seems more pertinent to have a generic overview of the performance, as many of the shortcomings found can easily be corrected. Table 1 shows the breakdown of the retrieved patterns

As it can be seen, most frequently occurring patterns are the *Poss N* string and the medial context (between commas). The distribution of the patterns across the three contexts here defined is very uneven, in some cases only some very few instances were found.

The attentive reader will have noticed that some patterns were left out from the experiments: the isolated noun (both proper and common), and the the pattern *Adj PROP*. The first, though it can be used for vocatives (e.g. *João, não faças isso.* 'João, don't do that')

[9] `http://www.linguateca.pt/Floresta/BibliaFlorestal/anexo4.html`
[10] `http://www.linguateca.pt/acesso/corpus.php?corpus=CETEMPUBLICO`

Table 1 Patterns retrieved from the *corpus*. *Poss*= possessive, *N*= noun (common), *PROP*= proper noun,*Adj*= adjective, `[vocnoun]`= vocative noun, `[vocadj]`= vocative adjective.

Pattern	# _ ,	, _ ,	, _ #	Total
Poss	1	21	8	30
Poss Adj	1	33	7	41
Poss N	43	285	97	**425**
Poss PROP	25	66	10	101
Poss Adj N	6	43	7	56
Poss Adj PROP	5	16	5	26
Subtotal	81	**464**	134	679
pá	0	41	31	72

produced a large number of matches (over 133,000), and a cursory analysis showed that most of them are spurious cases, mostly appositive `NP`. The second pattern occurred less frequently (736 matches). It can also provide instances of vocative (e.g. *Caro João, não faças isso.* 'Dear João, don't do that'), but in most cases, they correspond to: (i) the metalinguistic operador *chamado* 'called' (and its synonyms, like *denominado* 'idem'), used in appositions (e.g. *um produto semelhante , chamado Tiger*); (ii) the adjective *antiga*, also used in appositions (e.g. [*Visita a*] *Varanasi , antiga Benares, ...* '[visit to] Varanasi, formerly known as Benares') (iii) part-of-speech ambiguous adjectives (e.g. *viva Timor Leste* '[long] live *East Timor*'); (iv) compound proper nouns (e.g. *Nova Deli*) that, unlike STRING, the *corpus* does not treated as a single token. For both patterns, some strategy must be devised to rule out these cases. This is the precisely the point of the lexical focus of this paper's approach, by defining subsets of nouns and adjectives frequently occurring in vocative constructions.

The interjection *pá* does not constitute a real issue for the parser, since its use in isolation and the exceptions described in the grammar precluded false-positives. In the same way, the interjection *ó*, which occurred 475 times in the *corpus*, did not pose much of a difficulty to the parser. For example, since the STRING's lexicon includes the proper noun *Ó*, this word is always chunked together with the other elements of the name, forming a single noun chunk, e.g. *Luís do Ó*, *Nossa Senhora do Ó* 'Our lady of Ó [=birth]'. Combinations of *ó* with adjectives are also an important cause for missing vocatives, e.g. *Ó poderosos* 'O powerfull [people]'. Some of the false-negatives correspond to strings where the vocative is not strictly at beginning and the end of the sentence, as some separator (quotes and dashes) occur, e.g. *"–Ó filho, ...* 'O [my] son'; *J.G. – Ó pá , eu digo-te* 'J.G.[speaker] – O man, I'll tell you'. These all textual-orthographic issues that surely have to be addressed, probably at a pre-processing stage, but are somehow marginal to the scope of this paper.

Regarding the patterns with possessives, though the majority of the instances of vocatives have been correctly identified and the corresponding dependency extracted (approximately 75% precision), some remarks are in order. A major source for false-negatives are lexical *lacunae*: *Hastings* as proper name wass missing from the lexicon, thus, in some contexts some vocatives with this name were missed. Another reason is the incomplete (semantic) description of nouns forming homogenous subsets of the lexicon. This process of systematically extending (by way of rules) the feature `vocnoun` in still underway. This explains why several vocatives involving nouns designating family relations and professions were missed, e.g. *minha filha* 'my daughter', *meu general* 'my general', *meu caro psiquiatra* 'my dear psychiatrist'. Other cases of `vocnoun` not previously considered were also found: *amado*, *bem-amado* 'loved-one', *amor* 'love', *nené* 'baby', *jóia* 'jewel', *santo* 'saint', etc.

Other false-negatives relate to the ambiguous expression *meu deus* 'my god' (upper/lower-case variation involved), which can correspond to either an interjection (*Ai, meu Deus, eu quero ir para a beira dele* 'Oh, my God, I want to go near him') – which is the most frequent case – or a vocative proper (*Obrigado, meu Deus, ... clamava o pastor* 'Thank you, my God, ... the shepherd cried') – corresponding to a real address to the deity. Because of this ambiguity, this expression has been left out of the scope of the vocatives' grammar.

A special attention should be paid to names involving *insults* (usually with profanity words), which systematically enter vocative constructions, e.g. *meu grande paneleiro* 'my big fagot', *meu sacana* 'you bastard', *minha cabra* 'you bitch'. As mentioned before, vocative constructions can be used for insulting the addressee, though the specific lexicon involved has not been described in the STRING yet (see [5]). The lexicon-grammar of insults is the topic for another study.

6 Conclusion and future work

This paper aimed at providing a coherent theoretical framework to support the formal description of vocative constructions within a computational grammar of Portuguese. From a (necessarily brief) overview of different linguistic-grammatical perspectives on the status of vocatives, the paper has shown some of the issues vocatives raise for an adequate parsing. Theses issues regard, mostly, (i) the place on insertion (or attachment) of the vocative phrases within the sentence parse or to the sentence's topmost node (if the system adopts such formalism); and (ii) the inadequate parsing of other sentence constituents, foremost the subject. In our perspective, vocatives are to be treated as an independent syntactic function, distinct from sentence-internal constituents like the subject; and they hinge on (zeroed) metalinguistic operators involved in the communicative functions of language, hence, they should be linked to the sentence as a whole. A (very brief) comparison of two Portuguese parsing systems showed: (i) one that apparently does not consider a specific dependency for vocatives, with all the inadequacies in the resulting parses, whenever a vocative is involved; being a statistical parser, based on a language model, the sparsity of the phenomena may be the source for the vocatives not having been taken into account. (ii) another system that, while considering a vocative dependency, raises several issues on the *locus* the vocative phrase should be attached to, besides the coverage of different linguistic aspects of this phenomena.

We set out to produce a module for the processing of vocatives within the computational grammar of Portuguese built for XIP [1], by systematically exploring the formal variation found in an initial overview of examples taken from *corpus*. The strategy here adopted relies on preexisting modules for identifying and chunking proper names, including titles and profession nouns, and on a large-sized lexicon. The processing of vocatives involves two major steps, illustrated in detail by the paper: First, with a specifically built *local-grammar*, declarative rules identify vocative phrases and tag them with a `vocnp` feature, which is then used by the dependency rules to extract the `VOCATIVE` dependency. For a preliminary evaluation of the parser, a set of relatively flexible patterns were extracted from the CetemPúblico *corpus*, and results were commented in detail. Current results are promising but there is still room for improvement, especially in the fine-grained description of lexical features, the intersection of vocatives and the grammar of insults, and in the removal of spurious dependencies still left in the parse. This will constitute our next steps towards a comprehensive and efficient grammar for parsing vocatives in Portuguese.

References

1. Salah Ait-Mokhtar, Jean Pierre Chanod, and Claude Roux. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(3):121–144, 2002.
2. James M. Anderson. Case grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 58–65. Pergamon, 1996.
3. Eckard Bick. *The Parsing System "PALAVRAS". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Arhus University Press, 2000.
4. Ana Maria Brito, Inês Duate, and Gabriela Matos. Estrutura das frases simples e tipos de frases. In Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, and Isabel Hub Faria, editors, *Gramática da Língua Portuguesa*, pages 432–506. Caminho, 3rd edition, 2003.
5. Paula Carvalho. *Análise e representação de construções adjectivais para processamento automático de texto. Adjectivos intransitivos humanos*. PhD thesis, Faculdade de Letras, Universidade de Lisboa, 2007.
6. Authors: Francisco Costa and António Branco. LX-Gram: A deep linguistic processing grammar for portuguese. In *Computational Processing of the Portuguese Language (PROPOR)*, pages 86–89, 2010.
7. Celso Cunha and Luís Lindley-Cintra. *Nova Gramática do Português Contempoâneo*. João Sá da Costa, 1986.
8. 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: O Segundo HAREM*. Linguateca, 2009.
9. Zellig Sabettai Harris. *A Grammar of English on Mathematical Principles*. Jouhn Wiley & Sons, Wiley-Interscience Pub., 1982.
10. Zellig Sabettai Harris. *A Theory of Language and Information. A Mathematical Approach*. Clarendon Press, 1991.
11. Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. STRING - a hybrid statistical and rule-based natural language processing chain for Portuguese. In *Intl. Conf. Computational Processing of Portuguese (PROPOR)*, 2012.
12. Christian Molinier and Françoise Levrier. *Grammaire des adverbes: description des formes en -ment*. Droz, 2000.
13. Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666, 2016.
14. Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura, Amália Mendes, Graça Vicente, and Rita Veloso. *Gramática do Português*. Fundação Calouste Gulbenkian, 2013.
15. Diana Santos and Paulo Rocha. Evaluating CETEMPúblico: A free resource for portuguese. In *39 Annual Meeting of the Association for Computational Linguistics*, pages 442–449, 2001.
16. Mariona Taulé, M. Antònia Martí, and Marta Recasens. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *6th International Conference on Language Resources and Evaluation (LREC)*, pages 96–101, 2008.