Towards a Syntactic Lexicon of Brazilian Portuguese Adjectives

Ryan Saldanha Martinez
Universidade Federal de São Carlos, Brazil
`ryan.saldanha.martinez@gmail.com`

Jorge Baptista
Universidade do Algarve, Portugal
INESC-ID Lisboa, Portugal
`jbaptis@ualg.pt`

Oto Araújo Vale
Universidade Federal de São Carlos, Brazil
`otovale@ufscar.br`

Abstract

This paper aims to present an ongoing large-scale classification of Brazilian Portuguese adjectives. The 2,000 most frequent adjective lemmas in a reference corpus, corresponding to 87.94% of the occurrences of adjectives, were classified into predicative and non-predicative. The former were further classified based on argument number (one or two) and type (noun phrase or clause), which led to six different classes of predicative adjectives plus two subclasses. The results suggest that the most representative class is non-predicative adjectives, followed by intransitive adjectives with noun phrase and clausal subjects, respectively.

1 Introduction

Analyzing the syntactic properties of lexical items in inventories offers valuable insights into sentence construction, the syntactic similarities among items, the correlation (or lack thereof) between these similarities and their meaning, and the effective differentiation of homonyms based on their syntax. For example, certain adjectives have both non-predicative (1a)-(1b) and predicative (1c) uses (Rio-Torto, 2006; Veloso and Raposo, 2013):

(1) a. *João tem um problema cardíaco.*
         'João has a cardiac problem'

     b. **João tem um problema que é/está cardíaco.*
         'João has a problem that is cardiac'

     c. *João é/*está cardíaco.*
         'João is cardiac' (= has a cardiac problem)

Other, semantically similar, adjectives should be considered only as non-predicative, even if this requires some further syntactic demonstration. Although sentences such as (2a-2b), with a copular verb, are acceptable, the adjective *ortopédico* 'orthopedic' should be considered as non-predicatve, since classifier nouns like *tipo* 'type' or *natureza* 'nature' can be reconstructed for these sentences, as shown in (2c).

(2) a. *Esse sapato é ortopédico.*
         'This shoe is orthopedic'

     b. *O problema do João é ortopédico.*
         'João's problem is orthopedic'

     c. *Esse sapato é de tipo ortopédico.*
         'This shoe is of an orthopedic type'

     d. *O problema do João é de natureza ortopédica.*
         'João's problem is orthopedic in nature'

None of this is true of sentence (1c), which cannot be paraphrased with classifier nouns *tipo* or *natureza*:

(3) a. **João é de tipo cardíaco*
         'João is of a cardiac type'

     b. **João é de natureza cardíaca*
         'João is cardiac in nature'

Thus, *cardíaco* can be both a predicative and non-predicative adjective, whereas *ortopédico* can only be non-predicative. An inventory of syntactic properties can be used to get insight on the patterns (or the lack thereof) behind facts such as these.

Syntactic lexicons of adjectives have been developed for several languages, such as French (Picabia, 1978), Korean (Jee-Sun, 1996), Greek (Valetopoulos, 2003), and Italian (Messina, 2019). For European Portuguese there are three works concentrating on subclasses of adjectives: adjectives taking complement clauses (Casteleiro, 1981), adjectives with the suffix *-vel* '-able' (Freire, 1995), and intransitive adjectives with human subjects (Carvalho, 2007). Additionally, a considerable amount of nouns accepting support verb *ser de* 'to be of' have an adjectival counterpart with similar syntactic properties (Baptista, 2005).

There is considerable contemporary work dealing with the syntax of Brazilian Portuguese adjectives through different perspectives. The topics under discussion in recent studies include semantic factors determining pre- or post-nominal adjective position (Prim, 2010), the syntax behind agreement and the lack thereof between nouns and predicative adjectives in certain constructions (Rodrigues and Foltran, 2013), the syntax and semantics of adjectival intensifiers (Foltran and Nóbrega, 2016), the derivation of adjectives ending in *-vel* (-ble) from verbs and nouns (Jovem and Silva, 2017), and the relation between formal and semantic properties of adjectives and their cognitive, discursive, and pragmatic counterparts (Romerito Silva and Ferreira Cabral Oliveira, 2022), among others. However, these approaches do not attempt at creating large-scale syntactic lexicons.

These related works need to be complemented by further data for two reasons: (i) some types of adjectives have not been dealt with at all so far; these include non-predicative adjectives, such as *arterial*, derived from nouns with argument status on further operators (e.g. *obstrução arterial* 'arterial obstruction' = *obstrução das artérias* 'obstruction of the arteries'); those operating on non-human subjects, e.g. *compacto* 'compact'; and adjectives selecting two nominal arguments, i.e. establishing a relation between two noun phrases, e.g. *leal* 'loyal'; (ii) as far as could be ascertained, there is no syntactic lexicon for Brazilian Portuguese adjectives (besides traditional valency dictionaries, such as (Fernandes, 1948; Borba, 2002)); and, while descriptions of European Portuguese might be considerably similar, they often differ in detail.

In addition to their utility for linguistic research, syntactic lexicons such as these may be used as a resource for syntactic annotation and correction in corpus annotation tasks in different formalisms.

This paper presents an ongoing effort to fill these gaps by building a classification of a significant number of frequently used Brazilian Portuguese predicative adjectives.

2 Method

2.1 Lexical Selection

To guarantee that the lexicon focuses on relevant units, a list of the 2,000 most frequent lemmas of adjectives in the Brazilian partition of the PtTenTen2020 corpus (Kilgarriff et al., 2014b,a;

Wagner Filho et al., 2018) was extracted using SketchEngine[1].

This corpus is mostly composed of internet texts. The 2,000 adjectives cover 87.94% of all occurrences of adjectives in this corpus, which, for this ongoing project, seemed a reasonable lexical coverage.

Some of these items were excluded due to not being considered adjectives (mainly possessive pronouns, ordinal numbers, and some mislabeled given names). These amounted to 244 types of the original list. Thus, 1,756 adjectives were included in this study.

Adjectives ending in *-vel* that are related to transitive verbs allowing for the Passive transformation were also excluded, since these can be regularly derived from the corresponding verbal construction (Leeman and Meleuc, 1990), as in examples (4a)-(4c):

(4) a. *Essas montanhas são escaláveis.*
        'These mountains are climbable'

     b. *Essas montanhas podem ser escaladas.*
        'These mountains can be climbed'

     c. *Alguém escala essas montanhas.*
        'Someone climbs these mountains'

In contrast, certain adjectives, while also ending in *-vel*, present no such correspondence, as in the example (5), below:

(5) *É provável que João faça isso*
     'It is likely/probable that João does that'

which cannot be derived from any construction of the verb *provar* 'taste, prove': *João provou a sopa* 'João tasted the soup'; *João provou que tinha razão* 'João proved that he was right' Thus, only autonomous adjectives ending in *-vel*, such as *provável*, were included in the classification.

Some past participles were considered a type of adjective (Gross, 1996a) when accepting predicative constructions with *ser* and *estar* 'to be' but are not trivially derivable from a verbal counterpart, as in (6a)-(6b), below.

(6) a. *João está aberto a fazer isso.*
        'João is open to do that.'

     b. **Alguém abriu João a fazer isso.*
        'Someone opened João to do that.'

These items still lack a systematic description in Brazilian Portuguese.

2.2 Classification Criteria

Following Lexicon-Grammar (Gross, 1975, 1981, 1996b) as a theoretical and methodological framework, the classification was based on three criteria: (i) whether or not these adjectives were *predicative*, i.e., accept the post-copulative verb context; (ii) the *number* of arguments (one ore two) selected by the predicative adjective; and (iii) the *type* (nominal or sentential) of selected arguments. This process is illustrated with the classification key in Figure 1. The resulting classification is shown in Table 1.

Predicative adjectives are those which accept constructions with copular verbs in non-contrastive contexts, as in example (7).

(7)   *João é eficaz em fazer isso.*
      'João is efficient in doing that.'

In turn, non-predicative adjectives might surface in sentences with copular verbs, but only in conjoined contrastive sentences (Casteleiro, 1981), as in *Essa pesquisa é científica, não mercadológica* 'This research is scientific, not (a) market (one)', and are often associated with a classifier noun (Gross, 1988a), like *tipo* 'type' or *natureza* 'nature'.

(8)   *Essa pesquisa é de tipo científico/natureza científica*
      'This research is of a scientific type/nature.'

Naturally, *científico* has both a predicative and a non-predicative reading. The predicative reading, meaning 'following the scientific method', accepts both human and non-human subjects.

(9)   *O João / esta pesquisa foi (muito) científica*
      'João / this research was (very) scientific'.

Some non-predicative adjectives are restricted to pre-nominal position (10a)-(10b), while others are restricted to post-nominal position (11a)-(11b):

(10) a.   *João viveu altas aventuras no Rio.*
           'João had great adventures in Rio.'

      b.   **João viveu aventuras altas no Rio.*
           'João had adventures great in Rio.'

(11) a.   *João fez um procedimento cirúrgico.*
           'João had a surgical procedure.'

      b.   **João fez um cirúrgico procedimento.*
           'John had a surgical procedure.'

Non-predicative adjectives are marked as ANP and classified no further, at this time.

Predicative adjectives are firstly classified based on their number of arguments. Some of them accept a single argument (the subject), but still present multiple constructions, depending on the structural and distributional constraints on that argument slot. For example, the nominal (12a)-(12b) or sentential (12c) nature of the subject of *falso* 'deceiving, counterfeit, false', as well as the human (12a) or non-human (12b) nature of the subject noun phrase, are criteria used to distinguish three lexical entries for this adjective (the codes in brackets indicate their lexical-syntactic class):

(12) a.   *João é falso* [AN0h]
           'João is deceiving.'

      b.   *Esse objeto é falso.* [AN0n]
           'This object is counterfeit.'

      c.   *Que o João tenha feito isso é falso.* [AQ0]
           'It is false that João did that.'

Other adjectives take also a complement, introduced by a preposition:

(13) a.   *João é natural de São Paulo.* [AN2]
           'João is from São Paulo.'

      b.   *João está pronto para fazer isso* [ANQ]
           'João is ready to do that.'

      c.   *Que o João tenha feito isso é preocupante para a Maria.* [AQN]
           'That João did this is worrying for Maria.'

      d.   *Que o João tenha feito isso é sugestivo de que a Maria fez aquilo.* [AQ2]
           'That João did this is suggestive that Mary did that.'

The last classification criterion is whether an argument can be a complement clause (Q) or only a noun phrase (N). In the case of single argument constructions, examples (12a)-(12c) illustrate those distinctions. Similarly, for two-argument adjectives, sentence (13a) exemplifies constructions whose
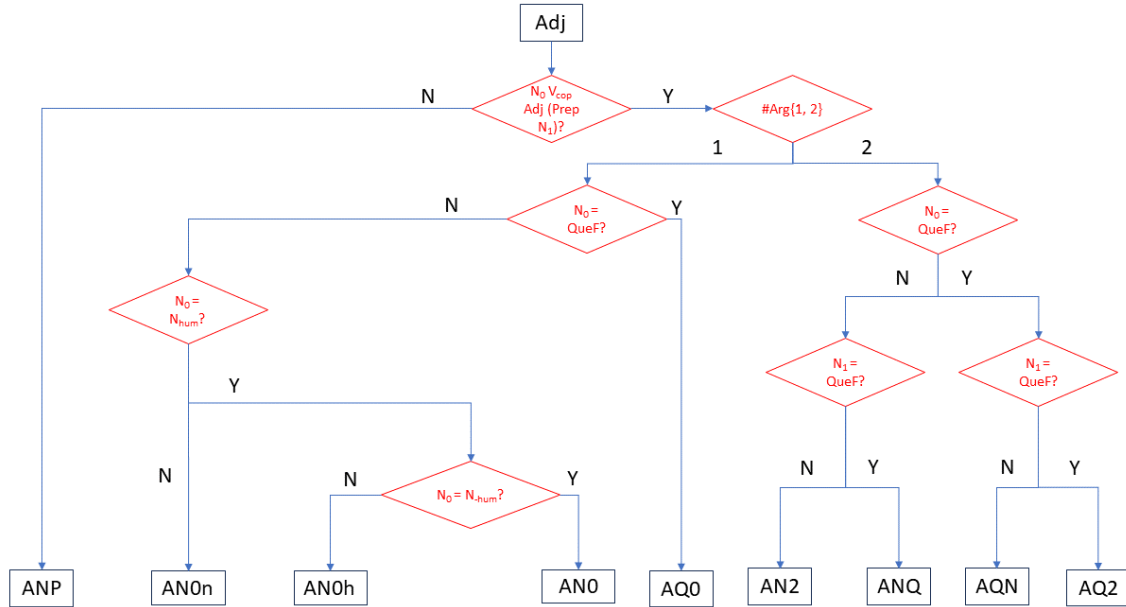
Figure 1: Adjective classification key. Adj = Adjective; #Arg{1, 2} = Number of arguments (1 or 2); N_0 = Subject; N_1 = Object; N_{hum} = Human Noun; N_{-hum} = Non-human Noun; V_{cop} = Copular verb; QueF = complement clause; N = No; Y = Yes; the conventional codes in last line indicate the syntactic classes.

Class	Definition	Example	n	%
ANP	*N_0 V_{cop} Adj (Prep N_1)	*problema cardíaco*: *O problema do João é cardíaco* 'cardiac problem' 'João's problem is cardiac'	827	35.1
AN0	N_0 V_{cop} Adj	*O João é falso* 'João is false'	762	32.3
AQ0	F_0 V_{cop} Adj	*Que o João tenha feito isso é falso* 'It is false that João has done this'	377	16.0
AN2	N_0 V_{cop} Adj Prep N_1	*O João é natural de São Paulo* 'João is from São Paulo'	163	6.9
ANQ	N_0 V_{cop} Adj Prep F_1	*O João está pronto para fazer isso* 'João is ready to do this'	94	4.0
AQN	F_0 V_{cop} Adj Prep N_1	*Que o João tenha feito isso é preocupante para a Maria* 'That João did this is worrying for Maria'	83	3.5
AQ2	F_0 V_{cop} Adj Prep F_1	*Que o João tenha feito isso é sugestivo de que a Maria fez aquilo* 'That João has done this is suggestive that Maria did that'	47	2.0
A_{comp}	N_0 ser Adj do que N_1	*Esse prédio é maior do que aquele* 'This building is bigger than that one'	4	0.2
Total			**2,357**	

Table 1: Number and percentage of items in each adjective class

subject and complement can be only N. The construction of *pronto* (13b) provides an example of a type N subject construction and a type Q complement, while the example of *preocupante* (13c) shows the inverse distribution. Finally, the construction of *sugestivo* (13d) has both Q subject and object. Since the same adjective can sometimes have different meanings depending on the human and non-human type subjects, AN0x adjectives were also sub-classified into AN0h, for exclusively human subject, as in (12a) and AN0n, for exclusively non-human subject, as in (12b). Due to the number of entries found so far, two-argument constructions have not yet undergone this sub-classification process. A small set of *comparative* adjectives, namely *melhor* 'better', *pior*

'worse', *maior* 'bigger, higher', and *menor* 'smaller, lower' were grouped in a separate class, which is not included in the flowchart, as they require a more complex syntactic analysis, involving a comparative conjunction, *do que* 'than' (14)

(14)   *Este prédio é maior/menor* do que *aquele.* :
       'This building is bigger / smaller than that one'

Adjectives such as *superior* 'id.', *inferior* 'id.' , though also comparative in meaning, do not present such properties and fall within the general classification (15):

(15)   *Este orçamento é superior/inferior àquele*
       'This budget is higher/lower than that one'.

2.3   Acceptability Judgements

To verify whether adjectives matched the aforementioned properties, native speakers' competence/intuition was primarily resorted to (Gross, 1976, 1988b). On occasion, concordances were also extracted from the Brazilian partition of the PtTenTen2020 corpus accessible through SketchEngine, to verify whether a dubious form would show up. Thus, the application of the classification criteria combines introspection and corpora (Laporte, 2007, 2015).

3   Preliminary Results

The classification of the most frequent 2,000 forms labeled as adjectives in the Brazilian part of PtTen-Ten2020 led to 2,357 entries, each with an illustrative example. These adjectives cover 87.94% of all forms labeled as adjective in that subcorpus.

The data so far show a disproportionate distribution of the proposed classes. The class ANP (non predicative) has the most adjectives; this result is not unexpected, since ANP is a particularly overarching class involving several types of adjectives, which will require further analysis.

The second most frequent class of adjectives is AN0 (without complement), which includes 78 AN0h, 82 AN0n and 602 AN0x. Adjectives with sub-clausal arguments, when combined, cover nearly one fourth of the data; the most frequent of these is AQ0, which is also the third most frequent class in this data, followed by ANQ and AQN with similar numbers; and, then, AQ2 (two sub-clausal

arguments). Finally, AN2 (non-completive, two-argument constructions) is so far the fourth most frequent class.

4   Conclusion and Next Steps

This paper presented an ongoing effort to build a syntactic lexicon of Brazilian Portuguese adjectives. About 2,000 lemmas were assigned lexical-syntactic classes taking into account whether these adjectives were predicative or not, as well as their number and type of arguments. This list represents a coverage of 87.94% of the adjective tokens in the reference corpus (ptTenTen2020). The next steps of this project are, firstly, to carry on with this classification to achieve a higher coverage of the corpus aiming at +95% of all adjectives' instances. The data will be made available to the scientific community once a satisfactory coverage is achieved.

Secondly, a detailed description of the syntactic properties of these items will be provided. These include: (i) choice of the adjective's support verb (or copula) *ser* or *estar* 'to be' and their aspectual variants (Gross, 1996a); (ii) preposition introducing the complement; (iii) subjunctive/indicative mood of the sub-clauses; (iv) correferential constraints between a nominal argument and the subject of the sub-clause; (v) syntactic transformations, such as the equivalence (paraphrastic) relation to nominal (*falso* 'false' *falsidade* 'falsehood') or verbal constructions (*preocupante* 'worrying' *preocupar* 'to worry') (Harris, 1964, 1991), among others, or the sub-clause restructuring:

(16)   a.   *Que o João tenha feito isso foi muito construtivo* (*da sua parte*)
            'That João has done this was very constructive of him'

       b.   *João foi construtivo em fazer isso*
            'João was very constructive in doing that'

       c.   (*João*) *ter feito isso foi muito construtivo da parte dele*
            '(João) having done that was very constructive of him'

The initial focus will be on the description of adjectives accepting clausal arguments (classes AQ0, AQN, ANQ, and AQ2) and to account for contrasts between the Brazilian and European Portuguese varieties (Casteleiro, 1981).

Acknowledgements

References

Jorge Baptista. 2005. *Sintaxe dos predicados nominais com* ser de. Fundação Calouste Gulbenkian & Fundação para a Ciência e a Tecnologia, Lisboa.

Francisco da Silva Borba. 2002. *Dicionário de usos do português do Brasil*. (Editora Ática).

Paula Cristina Quaresma da Fonseca Carvalho. 2007. *Análise e representação de construções adjectivais para processamento automático de texto: adjectivos intransitivos humanos*. Ph.D. thesis, Universidade de Lisboa.

João Malaca Casteleiro. 1981. *Sintaxe transformacional do adjectivo – regência das construções completivas*. INIC, Lisbon.

Francisco Fernandes. 1948. *Dicionário de Regimes de Substantivos e Adjetivos*, 28th edition. Globo, São Paulo, Brasil.

Maria José Foltran and Vítor Augusto Nóbrega. 2016. Adjetivos intensificadores no português brasileiro: propriedades, distribuição e reflexos morfológicos. *Alfa: Revista de Linguística (São José do Rio Preto)*, 60:319–340.

Helena Maria Serras Reis Silva Freire. 1995. Determinação e formalização das propriedades sintácticas de adjectivos terminados em-vel. Master's thesis, Universidade de Lisboa.

Gaston Gross. 1988a. Degré de figement des noms composés. *Langages*, 90:57–72.

Maurice Gross. 1975. *Méthodes en syntaxe: régime des constructions complétives*. Herman, Paris.

Maurice Gross. 1976. Présentation. In Jean-Paul Boons, Alain Guillet, and Christian Leclère, editors, *La structure des phrases simples en français: constructions intransitives*. Droz.

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

Maurice Gross. 1988b. Methods and tactics in the construction of a lexicon-grammar. *Linguistics in the morning calm*, 2:177–197.

Maurice Gross. 1996a. Les verbes supports d'adjectifs et le passif. *Langages*, pages 8–18.

Maurice Gross. 1996b. Lexicon-Grammar. In *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon.

Zellig Sabbettai Harris. 1964. Transformations in Linguistic Structure. *Proceedings of the American Philosophical Society*, 108(5):418–422.

Zellig Sabbettai Harris. 1991. *Theory of Language and Information. A Mathematical Approach*. Clarendon Press, Oxford.

Nam Jee-Sun. 1996. *Classification syntaxique des constructions adjectivales en coréen*. John Benjamins Publishing Company.

Manuella Soares Jovem and José Romerito Silva. 2017. Rede construcional dos adjetivos formados por *-vel* no português. *Revista Odisseia*, 2(1):3–18.

A. Kilgarriff, M. Jakubíček, J. Pomikalek, T.B. Sardinha, and P. Whitelock. 2014a. Pttenten: A corpus for portuguese lexicography. In T.B. Sardinha and T. de Lurdes São Bento Ferreira, editors, *Working with Portuguese Corpora*, pages 111–128. Bloomsbury Academic.

Adam Kilgarriff, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha, and Pete Whitelock. 2014b. *PtTen-Ten: A Corpus for Portuguese Lexicography*, page 111–128. Bloomsbury Publishing.

Eric Laporte. 2007. Exemples attestés et exemples construits dans la pratique du lexique-grammaire. In *Observations et manipulations en linguistique: entre concurrence et complémentarité*, volume 16, pages 11–32. Peeters.

Éric Laporte. 2015. The science of linguistics. *Inference: International Review of Science*, 1(2):1.

Danielle Leeman and Serge Meleuc. 1990. Verbes en tables et adjectifs en *-able*. *Langue française*, 87:30–51.

Simona Messina. 2019. The predicative adjective and its propositional arguments: A lexicon-grammar classification. *Lingvisticæ Investigationes*, 42(2):234–261.

Lélia Picabia. 1978. *Les constructions adjectivales en français: systématique transformationnelle*, volume 11. Librairie Droz.

Cristina de Souza Prim. 2010. A sintaxe de adjetivos nas posições pré- e pós-nominal. Master's thesis, Universidade Federal de Santa Catarina.

Graça Rio-Torto. 2006. Para uma gramática do adjetivo. *Alfa: Revista de Linguística*, 50(2):103–129.

Patrícia Rodrigues and Maria José Foltran. 2013. Construções de *small clauses* complexas em português brasileiro. *Estudos Linguísticos (São Paulo. 1978)*, 42(1):497–511.

José Romerito Silva and Ana Catarina Ferreira Cabral Oliveira. 2022. O adjetivo no português brasileiro contemporâneo. *Revista de Estudos da Linguagem*, 30(2):1056–1102.

Freiderikos Valetopoulos. 2003. *Les adjectifs prédicatifs en grec et en français: de l'analyse syntaxique à l'élaboration des classes sémantiques*. Ph.D. thesis, Université Paris 13.

Rita Veloso and Eduardo B Paiva Raposo. 2013. Adjetivo e sintagma adjetival. In *Gramática do português*, volume 2, pages 1359–1493.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: a new open resource for Brazilian Portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.