




Linguistic resources for paraphrase generation in portuguese: a lexicon-grammar approach

Anabela Barreiro² · Cristina Mota²  ·
Jorge Baptista^{1,2} · Lucília Chacoto^{1,3} ·
Paula Carvalho²

Accepted: 3 August 2021 / Published online: 24 January 2022
© Springer Nature B.V. 2021

Abstract This paper presents a new linguistic resource for the generation of paraphrases in Portuguese, based on the lexicon-grammar framework. The resource components include: (i) a lexicon-grammar based dictionary of 2100 predicate nouns co-occurring with the support verb *ser de* ‘be of’, such as in *ser de uma ajuda inestimável* ‘be of invaluable help’; (ii) a lexicon-grammar based dictionary of 6000 predicate nouns co-occurring with the support verb *fazer* ‘do’ or ‘make’, such as in *fazer uma comparação* ‘make a comparison’; and (iii) a lexicon-grammar based dictionary of about 5000 human intransitive adjectives co-occurring with the copula verbs *ser* and/or *estar* ‘be’, such as in *ser simpático* ‘be kind’ or *estar entusiasmado* ‘be enthusiastic’. A set of local grammars explore the properties described in linguistic resources, enabling a variety of text transformation tasks for paraphrasing applications. The paper highlights the different complementary and synergistic components and integration efforts, and presents some preliminary evaluation results on the inclusion of such resources in the eSPERTo paraphrase generation system.

✉ Cristina Mota
cristina.mota@inesc-id.pt

Anabela Barreiro
anabela.barreiro@inesc-id.pt

Jorge Baptista
jbaptis@ualg.pt

Lucília Chacoto
lchacoto@ualg.pt

Paula Carvalho
pcc@inesc-id.pt

¹ Universidade do Algarve, Faro, Portugal

² INESC-ID Lisboa, Lisbon, Portugal

³ IELT, Lisbon, Portugal

Keywords Language Generation Resources · Paraphrasing · Lexicon-Grammar · Paraphrase Generator · Portuguese

1 Introduction

Several linguistic experiments have shown the benefits of using expert-crafted, large-coverage dictionaries and grammars in software applications that can parse sentences and produce paraphrases. Port4NooJ is a publicly available library of linguistic resources that contains lexica and grammars, cohesively integrated into a larger resource used in paraphrasing, among other natural language processing applications, including translation from and into Portuguese. Port4NooJ formalizes the standard vocabulary of the Portuguese language with a description at the lexical, morphological, syntactic and semantic levels. The availability of these resources enables to generate paraphrases of support verb constructions (e.g., paraphrasing the verbal construction with the equivalent nominal construction and vice-versa, or alternation between the support verb and other stylistic or aspectual verbs), active and passive constructions, adverbial compounds, relatives and participles, among a wide variety of paraphrasing capabilities.

The kick-start Port4NooJ linguistic resources derive from the OpenLogos English–Portuguese bilingual dictionary. OpenLogos is an open source derivative of the commercial Logos system underpinned on the Logos Model (Barreiro, 2011; Scott, 2003, 2018). OpenLogos is downloadable from the DFKI website¹ and a demo of the system is also available for testing at the INESC-ID Lisboa website.²

The OpenLogos bilingual resources were converted into the NooJ format and enhanced with new properties including derivational, morpho-syntactic and semantic relations, which allow for the generation of different types of paraphrases for Portuguese (Barreiro, 2009). The paraphrase generator is based on the NooJ’s linguistic environment (Silberztein, 2015, 2016).

NooJ has been created as a support platform to develop linguistic resources and process texts in French (Silberztein, 1993). Among other components, these resources include a large coverage dictionary containing standard vocabulary substantiated by the lexicon-grammar theory. Over the years, NooJ has become an interoperable electronic multilingual environment currently formalizing and serving over 20 languages, and public resources have been made available for each one of these languages, allowing the development of lexica and grammars conceived for different purposes, among others, to establish relationships between semantically related elements, for morphological, syntactic and semantic analysis, disambiguation, identification of multiword units, paraphrasing and translation. One of the most interesting and distinctive characteristics of NooJ is that it parses discontinuous multiword units effectively, making it a powerful tool for language generation. Port4NooJ, as well as NooJ resources for each language module, is based on lexicon-grammar.

¹ <http://logos-os.dfki.de/>.

² <http://www.hlt.inesc-id.pt/openlogos/demo.html>.

Lexicon-grammar is a theoretical and methodological framework developed by Gross (1975, 1982) and based on the transformational operator grammar by Harris (1952, 1965, 1968, 1991). The principle underlying the lexicon-grammar is that the unit of meaning is not the word, but the elementary sentence, which is composed of a predicate and its arguments. Within the lexicon-grammar framework, researchers have been studying systematically different types of linguistic phenomena, especially in the Romance languages, often with an explicit computational processing purpose in mind. Lexicon-grammar tables describe and formalize phenomena like verb, noun, and adjective predicate constructions, including multiword units (support verb constructions, phrasal verbs, idioms, etc.). The lexicon-grammar theory is particularly suitable for the recognition and generation of paraphrastic knowledge, since it is based on the Harrisian notion of transformation, which are construed as equivalence relations between sentences having the same content-words (or morphemes) and the same amount of distributional information. Also, the systematic description of predicates in different languages allows for multilingual studies, particularly translation (Barreiro, 2009).

In our approach, the Port4NooJ dictionaries describe a set of syntactic-semantic constructions, and the set of lexically-dependent transformations they allow, making it possible to establish linguistically motivated relations between equivalent utterances associated with the same predicates. Port4NooJ includes linguistic resources, such as: (i) a large coverage dictionary with over 40,000 Portuguese lemma entries and respective English correspondents (also known as ‘transfers’), which generates a dictionary of inflected forms with over 1 million entries; (ii) morphological rules to formalize and document Portuguese inflectional and derivational descriptions; (iii) local grammars for named entity recognition, pattern recognition and disambiguation, inflection and generation of multiword units; (iv) transformational grammars for paraphrasing and translation. In the core of the Port4NooJ bilingual Portuguese–English dictionary, the syntactic-semantic information assigned to each entry validates the linguistic relation between the terms in both languages making it a unique resource for machine translation. The addition of transfers for other languages are easily implemented, since they already exist in the Logos system and are publicly available in OpenLogos (Barreiro et al., 2014).

This paper focuses on the description of three new components of Port4NooJ v3.1, incorporated to strengthen and enhance eSPERTO’s paraphrastic capabilities. These new components include lexicon-grammar based dictionaries of predicate nouns co-occurring with the support verbs *ser de* ‘be of’ and *fazer* ‘do’ or ‘make’, and of human intransitive adjectives, such as *entusiasmado* (‘enthusiastic’); and sets of local grammars that explore the syntactic, semantic and transformational properties of these nouns and adjectives for paraphrasing.

All Port4NooJ components interact among them. Although Port4NooJ resources require refinement and enlargement, there is no other public resource for Portuguese that incorporates such a wide range of linguistic knowledge. These resources are being explored for paraphrasing tasks in the scope of the project eSPERTO,³ described in Sect. 2. For its potential and unique characteristics, the availability of

³ <http://esperto.hlt.inesc-id.pt/>.

its resources appears to be useful for the Portuguese language processing community. The complete set of linguistic resources can be downloaded from the NooJ website.⁴

2 The eSPERTo project

eSPERTo is an acronym for ‘System of Paraphrasing for Editing and Revision of Text’ (in Portuguese, *Sistema de Parafraseamento para Edição e Revisão de Texto*). The main objective of the eSPERTo project⁵ is the development of a web-based paraphrasing system, i.e. a system that recognizes and generates equivalent forms of expression. In Portuguese, the word *esperto* also means ‘smart’. The project name was chosen because it retains a central idea that symbolizes and represents what is intended of the paraphrasing system developed within the project, i.e. that it is linguistically intelligent, sophisticated, and context-sensitive. Therefore, eSPERTo is a ‘smart system’ in the sense that it contains semantic ‘understanding’, not only lexical or syntactic knowledge. Among other functions, the platform includes text-editing mechanisms, which provide a variety of alternatives for each expression, allowing the user to choose among several suggestions that can be immediately applied to text.

The development of eSPERTo is twofold, with on the one hand the development of a context-sensitive and linguistically enhanced paraphrase generator that recognizes both syntactic-semantic units, and frozen and semi-frozen expressions, and transforms them into semantically equivalent constructions or expressions, and on the other hand, the development of a new hybrid technique that combines statistics and linguistic knowledge for identifying and generating new and increasingly more complex paraphrases.

Currently, eSPERTo is integrated in an interactive web-based application that provides paraphrastic suggestions (or alternatives) intended to illustrate their potential impact into derivative applications. As this tool evolves, it is envisaged that its resources will be used in text production and revision, and Portuguese language learning. The utility of eSPERTo’s paraphrasing capabilities has been explored in two other application scenarios described in Mota et al. (2016): (i) in a question-answering system to increase the linguistic knowledge of an intelligent conversational virtual agent, and (ii) in a summarization tool to assist the paraphrasing task. However, the new resources herein described have not been integrated and evaluated in any of these applications yet. Another proposed application involves the construction of a dataset of paraphrastic contrasts among the distinct varieties of the Portuguese language, an indispensable resource for variety adaptation (or conversion), i.e., for dealing with cultural, linguistic and stylistic differences between varieties, for which we have already demonstrated in some experimental studies (cf. Barreiro & Mota, 2018, Barreiro et al., 2018, and Rebelo-Arnold et al., 2018). The undertaken efforts are in line with the work of Janssen et al. (2018) to create the Pluricentric Corpus of the Portuguese

⁴ <http://www.nooj-association.org/>.

⁵ <http://esperto.hlt.inesc-id.pt/>.

eSPERTo - System for Paraphrasing in Editing and Revision of Text

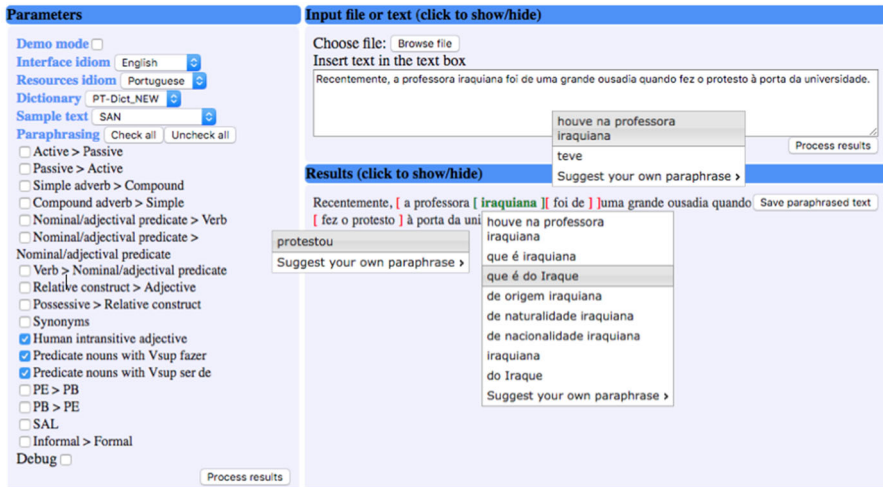


Fig. 1 eSPERTo's paraphrastic suggestions for a simple sentence involving predicate nouns with the support verbs *ser de* 'be of' and *fazer* 'do' or 'make' and the human intransitive adjective *iraquiana* 'Iraqi'

Language (CPLP Corpus) which aims to provide comparable corpora for the national varieties of the countries where Portuguese is an official language. Santos (2014, 2015b) also discuss the creation of an international standard of Portuguese but on a much broader and visionary way where the differences lose significance—they should not even be marked—and the similarities strengthen the idea that there is only one Portuguese, the International Portuguese (π).

Figure 1 illustrates the application of Port4NooJ resources, dictionaries and transformational grammars, to generate paraphrases that can be selected from eSPERTo. The examples show the three distinct linguistic phenomena discussed in this paper: predicate nouns with the support verbs *fazer* and *ser de*, and human intransitive adjectives. For the sentence *Recentemente, a professora iraquiana foi de uma grande ousadia quando fez o protesto à porta da universidade*. 'Recently, the Iraqi teacher was very bold when she made the protest at the university door.', eSPERTo provides several suggestions that can contribute to paraphrase this sentence in several possible ways. For example, the main sentence, *a professora iraquiana foi de uma grande ousadia* can be paraphrased as *houve na professora iraquiana uma grande ousadia* 'there was great boldness/daring by the Iraqi teacher' or as *a professora iraquiana teve uma grande ousadia* 'the Iraqi teacher had a great (dose of) boldness/daring', by applying the transformational properties of the predicate noun co-occurring with the support verb *ser de* converting it into its semantically equivalent verbs *houve* 'there was' or *teve* 'had'. Furthermore, the support verb construction *fez o protesto* 'made the protest' can be replaced by the verb *protestou* 'protested'. Also, the human intransitive adjective *iraquiana* 'Iraqi' can be replaced by any of the following equivalents: *que é iraquiana* 'who is Iraqi';

que é do Iraque ‘who is from Iraq’; *de origem iraquiana* ‘from Iraqi origin’; *de naturalidade iraquiana* ‘whose country of birth is Iraq’; *nascida no Iraque*, ‘born in Iraq’; *de nacionalidade iraquiana* ‘from Iraqi nationality’; *do Iraque* ‘from Iraq’. All the suggestions illustrated in the figure have in common the sharing of properties between semantically related words or the establishment of links to paradigmatic alternatives enabling the meaning of the sentence to remain unchanged. This is just one simple example of the positive effects that a paraphrasing system can have in applications such as language teaching, providing the students with the skills and knowledge to expand their vocabulary, learn grammar, i.e., syntactic functions, be more creative by having access to a wider range of writing possibilities and develop better improvising and interpreting skills. Interactive paraphrasing tools of this type are very motivating and pedagogical, because they permit an enormous interaction on the part of the students and users in general, a feature that is increasingly more popular in e-learning, but also very valuable in modern classroom learning and teaching methods in computer-based environments.

3 Related work

The idea of establishing relationships among semantically-related words of different parts-of-speech has been addressed in the field of linguistics since its early days, having its scientific roots in structural linguistics (Gross, 1975, 1981; Harris, 1952, 1965, 1981). From the application standpoint, the first commercial machine translation systems were presented with challenges at the morpho-syntactic and semantic level, which developers had to face at an early stage in order to produce correct translations. Requirements of machine translation models, such as the Logos Model (Scott, 2018), illustrate how early commercial machine translation systems tackled natural language generation, namely paraphrasing challenges, playing an important role in implementing transformations of the type described in this paper. For example, the need to transform predicate nouns (‘process nouns’ in Logos terminology) into verbs and vice-versa was already felt in the 1970s during the development of the original English–Vietnamese machine translation system, where the basis for dictionaries with so-called ‘alternate word classes’ has been established. As sustained in Scott (2018), the Vietnamese language does not easily support expressions like *He made a sudden move to the left*, preferring the more direct sentence *He moved suddenly to the left*. Logos developers had to be able to transform the English parts of speech to their alternative forms in the very early stages. Many other languages behave like Vietnamese, and paraphrasing is indispensable in (machine) translation and any language generation task.

Inspired in the Logos Model, but developed within the lexicon-grammar framework (Gross, 1975) and supported by NooJ, Barreiro (2009) established a methodology to create links among verbs, predicate nouns, and predicate adjectives, and between adjectives and adverbs, and so on, which is suitable for the generation of paraphrases. Our initial implementation efforts started with the linguistic phenomenon of support verb constructions. These constructions have been studied systematically for a long time in lexicon-grammar both from a monolingual and a bilingual

Furthermore, Rassi et al. (2014, 2015) describe the first attempts to integrate and use information formalized in a lexicon-grammar of Brazilian Portuguese predicate nouns with support verb *dar* ‘give’ in a natural language processing system, namely into STRING⁶ (Mamede et al., 2012), a hybrid statistical and rule-based pipeline for natural language processing of Portuguese. By the same time, Baptista et al. (2014, 2015) discuss lexical and parsing issues of integrating a lexicon-grammar of verbal idioms, such as *bater a bota* ‘kick the bucket’. More than 2000 rules were created semi-automatically for ten formal classes of verbal idioms.

The lexicon-grammar methodology is a labor intensive endeavour that will not produce results fast: a linguist needs to build a lexicon grammar first and then needs to develop the companion grammars. Converting the lexicon-grammar in a dictionary that the grammars can use, can be more or less automated as we previously showed. The transformational nature of the lexicon-grammar makes it ideal to generate paraphrases given an expression that the grammar recognizes, but using it to detect new paraphrases that the system does not know already, relating two different sequences in a text, is not straightforward.

It contrasts with the current corpora driven approaches that take advantage of machine learning algorithms and statistical analysis to rapidly detect paraphrastic pairs (Grycner & Weikum, 2016; Mayhew et al., 2020; Paşca & Dienes, 2005; Shinyama & Sekine, 2003). On the other hand, the corpora driven approaches usually detect paraphrastic relations among sequences of words that occur in those corpora whereas, as mentioned before, with our approach for a certain sequence found in a text, it suggests different ways of rewriting that sequence based on the morpho-syntactic and semantic properties of that sequence. Pershina et al. (2015) propose a new method to identify whether two idiomatic expressions are paraphrases in a corpus of SMS and tweets. Regarding Portuguese in particular, Souza and Sanches (2019) use sentence embeddings to detect whether two sequences of arbitrary sizes are paraphrases of one another, and Gamallo and Pereira-Fariña (2019) describe various unsupervised methods based on distributional models and dependency parsing to detect textual semantic similarity.

4 Lexicon-grammar resources for paraphrasing

This section describes the three main new components of Port4NooJ v3.1. used in paraphrasing, in particular, the lexicon-grammars of nominal predicate constructions with the support verbs *fazer* and *ser de* (cf. Sects. 4.1 and 4.2), and the lexicon-grammar of Portuguese human intransitive adjectives (cf. Sect. 4.3). These resources represent a consistent and sound scientific basis that allow the establishment of paraphrastic equivalences in Portuguese. The paraphrases obtained via the lexicon-grammar resources are possible due to a combination of these with Port4NooJ dictionaries and grammars. From this point forward, we present the main features of each lexicon-grammar and then the transformational properties encoded therein, which are the source for the paraphrastic equivalences found for these

⁶ <https://string.hlt.inesc-id.pt/>.

constructions. Some of these transformations are specific of certain support verb constructions (or even of certain lexicon-grammar classes), while others are common to different types of support verb constructions, and so they were described together.

All paraphrases achieved have in common a relationship between predicates, which can be either verbal or nominal. These predicates are often used interchangeably without any significant difference in meaning. Nominal predicates constituted by nouns and adjectives have argumental selection properties like verbs. These nouns and adjectives require complements (e.g., *ele tem desejo de ir à praia* ‘he has desire to go to the beach’ = *ele está deseioso de ir à praia* ‘he is eager/desiring to go to the beach’), being classified as transitive nouns and adjectives. Non-predicate nouns and adjectives that do not require complements are classified as intransitive, such as intransitive verbs. For example, *ele tem (uma grande) loucura* ‘he suffers from craziness’ = *ele é de uma grande loucura* ‘he is of a big craziness’ = *ele está louco* ‘he is crazy’. In these cases, the predicate corresponds to a noun or to an adjective, which impose restrictions both on the number and type of arguments they select and on the nature of the support verb with which they co-occur.

4.1 Nominal predicate constructions with *fazer* ‘do’/‘make’

Chacoto (2005) studied, classified and formalized the structural, distributional, and transformational properties of nearly 3000 predicate nouns which occur with the support verb *fazer* ‘do’/‘make’,⁷ like in *O Pedro fez o desenho de uma casa* ‘Peter made the drawing of a house’.

Chacoto’s systematic survey was carried out by perusing seven dictionaries and a corpus of around 5 million words (Part 20 of the 180 million word corpus CETEMPúblico.⁸ In total, 17 lexical and syntactic subclasses have been identified, where the predicate nouns in the lexicon are, in general, everyday vocabulary (simple and compound nouns), with the exception of a group of predicate nouns of the sports and medical domains. Table 1 illustrates the breakdown of these predicate nouns by classes and their corresponding basic syntactic structure. The main taxonomic criteria were: (i) the equivalence (or not) of the predicate noun with a verb or adjective construction—classes of autonomous predicate nouns are indicated by C; (ii) intransitive constructions, i.e., predicate nouns without any complement, and constructions with one or two prepositional complements; (iii) the preposition introducing the complement, main prepositions being, *a* ‘at’/‘to’, and *de* ‘of’ (other prepositions collapsed under notation *Prep*; (iv) distributional constraints on the subject and prepositional complement phrases’ head noun; (v) symmetry,

⁷ This is one of the most frequent verbs in European Portuguese, both in written texts and in the spoken language. Sentences with support verb constructions are often more frequent than sentences with the equivalent verbal constructions. This is corroborated by Barreiro (2009), who showed that from a search on all sentences of the COMPARA parallel corpus (Frankenberg-Garcia & Santos, 2003; Santos & Inácio, 2006) where the infinitive form of *fazer* occurs with a noun or with a left modifier and a noun, 47% of the times the occurrence is a support verb construction.

⁸ <http://www.linguateca.pt/aceso/corpus.php?corpus=CETEMPUBLICO>.

Table 1 Distribution of nominal predicates with support verb *fazer* by class attribute after integration into Port4NooJ

Class	Count	%	Structure
FN	347	5.8	$N_0 \text{ fazer } N$
FNAN	542	9.0	$N_0 \text{ fazer } N \text{ a } N_1$
FNDNh	467	7.8	$N_0 \text{ fazer } N \text{ de } (N_{\text{hum}})_1$
FNDN-hl	2127	35.5	$N_0 \text{ fazer } N \text{ de } N_1 \text{ <?>}$
FNDNa	1079	18.0	$N_0 \text{ fazer } N \text{ de } N_1 \text{ <?>}$
FNP	526	8.8	$N_0 \text{ fazer } N \text{ Prep } N_1$
FNDNAN	295	4.9	$N_0 \text{ fazer } N \text{ de } N_1 \text{ a } N_2$
FNDNPN	271	4.5	$N_0 \text{ fazer } N \text{ de } N_1 \text{ Prep } N_2$
FNDNPNSI	104	1.7	$N_0 \text{ fazer } N \text{ de } N_1 \text{ Prep } N_2$ [symmetric]
FNSI	99	1.7	$N_0 \text{ fazer } N \text{ Prep } N_1$ [symmetry]
FCSI	49	0.8	$N_0 \text{ fazer } C \text{ Prep } N_1$ [symmetric]
FND	74	1.2	$N_0 \text{ fazer } N \text{ <sport>}$
FNQ	11	0.2	$N_0 \text{ fazer } N \text{ Prep } \text{QueF}_1$
FCQ	9	0.2	$N_0 \text{ fazer } C \text{ Prep } \text{QueF}_1$
Total	6000		

either between the complements, or between the subject and a complement. Table 2 illustrates various paraphrases that we formalized involving the support verb construction *fazer*.

4.2 Nominal predicate constructions with *ser de* ‘be of’

Baptista (2000, 2005b) studied, classified and formalized into a lexicon-grammar the structural, distributional, and transformational properties of 2085 predicate nouns occurring in constructions with the support verb *ser de* ‘be of’ in European Portuguese, such as in *O Pedro foi de uma ajuda inestimável* ‘Peter was of an invaluable help’.

Table 3 presents the breakdown of these nouns by classes and their corresponding basic syntactic structure. Like the nouns with *fazer* ‘do’/‘make’, these classes were construed on the basis of the number or arguments (one or two) selected by the predicate noun, the syntactic (sentential/nominal) constraints and the distributional (semantic) selection constraints on the nominal argument slots (human/non-human). Two special classes were established, one for the nouns selecting a body-part noun as their subject, such as *A voz de Eva é de uma maviosidade impressionante* ‘Eva’s voice is of an impressive maviosity’, and another for the symmetrical constructions, such as *A é de uma equivalência perfeita alcom B* ‘A is of a perfect equivalence to/with B’. Figure 3 illustrates a local grammar to transform the sentence *O Pedro é de um certo altruísmo* ‘Pedro is of a certain altruism’ into two identical paraphrases: *O Pedro tem um certo altruísmo* ‘Pedro has a certain altruism’, and *Há no Pedro um certo altruísmo* ‘There is in Pedro a certain altruism’ and vice-versa. (Table 4) illustrates these paraphrases as well as others also involving the support verb construction *ser de*.

Table 2 Paraphrases of support verb constructions with *fazer*

	Paraphrases
(Complex) NP	
[Active S]	<i>O Pedro fez uma viagem a Londres</i> <i>Pedro went on a visit to London</i>
[Rel NG]	<i>A viagem que o Pedro fez a Londres</i> <i>The visit that Pedro made to London</i>
[Active NG]	<i>A viagem do Pedro a Londres</i> <i>Pedro's visit to London</i>
[Active S]	<i>O júri fez a contagem dos votos</i> <i>The jury made the vote count</i>
[Passive S]	<i>A contagem dos votos foi feita pelo júri</i> <i>The vote count was made by the jury</i>
[Rel Passive NG]	<i>A contagem dos votos que foi feita pelo júri</i> <i>The vote count that was made by the jury</i>
[Passive NG]	<i>A contagem dos votos por (parte de) o júri</i> <i>The vote count by the jury</i>
Dative	
[Dative S]	<i>O Pedro fez uma festa no rosto da Maria</i> <i>Pedro gave a caress in Maria's face</i>
[Restructuring]	<i>O Pedro fez uma festa à Maria (no rosto + θ)</i> <i>Pedro gave a caress to Maria (in the face)</i>
Symmetry	
[N0 Vsup/V Prep N1]	<i>O Pedro fez uma aliança/aliou-se (com + a) a Maria</i> <i>Pedro joined/allied himself with/to Maria</i>
[N1 Vsup/V Prep N0]	<i>A Maria fez uma aliança/aliou-se (com + a) o Pedro</i> <i>Maria joined/allied himself with/to Pedro</i>
[N0 Vsup N N1]	<i>O Pedro fez um acordo com a Maria</i> <i>Pedro made an agreement with Maria</i>
[N1 Vsup N N0]	<i>A Maria fez um acordo com o Pedro</i> <i>Maria made an agreement with Pedro</i>
Conversion	
[N0 fazer N Prep N1]	<i>O Paulo fez um telefonema à Maria</i> <i>Paulo made a phone call to Maria</i>
[N1 Vsup-conv Prep N0]	<i>A Maria recebeu um telefonema do Paulo</i> <i>Maria received a phone call from Paulo</i>
Vsup Variant	
[Vsup=fazer]	<i>O espião fez a codificação da mensagem</i> <i>The spy did the codification of the message</i>
[Vasp=iniciar]	<i>O espião iniciou a codificação da mensagem</i> <i>The spy initiated the codification of the message</i>
[Vsup=fazer]	<i>O Pedro fez uma fraude</i> <i>lit. Pedro made a fraud</i>

Table 2 continued

	Paraphrases
[Vstyle=cometer]	<i>O Pedro cometeu uma fraude</i> <i>Pedro committed a fraud</i>
[Vstyle=fazer]	<i>O Pedro faz natação</i> <i>Pedro does swimming</i>
[Vstyle=praticar]	<i>O Pedro pratica natção</i> <i>Pedro practices swimming</i>
Support Verb	
[fazer]	<i>O Pedro faz uma ideia sobre a situação política nacional</i> <i>Pedro makes an idea of the national political situation</i>
[ter]	<i>O Pedro tem uma ideia sobre a situação política nacional</i> <i>Pedro has an idea of the national political situation</i>
[fazer]	<i>O professor fez uma palestra</i> <i>The professor made a speech</i>
[dar]	<i>O professor deu uma palestra</i> <i>The professor gave a speech</i>
[fazer]	<i>O Pedro fez uma expedição ao Pólo Norte</i> <i>Pedro made an expedition to North Pole</i>
[estar] + Prep [em]	<i>O Pedro esteve numa expedição ao Pólo Norte</i> <i>Pedro was in an expedition to North Pole</i>

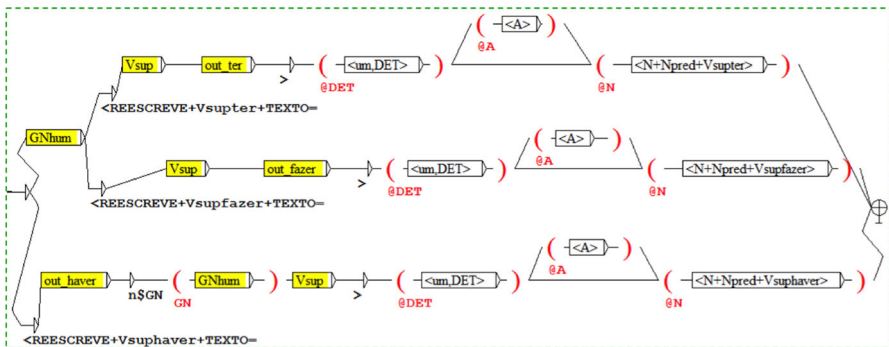


Fig. 3 Grammar to paraphrase constructions with *ser de*

4.3 Human intransitive adjectives

Carvalho (2007) studied, formalized, and classified the distributional properties of 4250 human intransitive adjectives in European Portuguese based on the lexicon-grammar methodology. Human intransitive adjectives can be defined as predicates

Table 3 Distribution of nominal predicates with support verb *ser de* by class attribute

Class	Count	%	Structure
SdH1	388	19	(<i>Nhum</i>) ₀ <i>ser de N</i>
SdH2	54	3	(<i>Nhum</i>) ₀ <i>ser de N Prep N₁</i>
SdNH1	363	17	(<i>Nnhum</i>) ₀ <i>ser de N</i>
SdNH2	30	1	(<i>Nnhum</i>) ₀ <i>ser de N Prep N₁</i>
SdNPC	30	1	(<i>Npc de Nhum</i>) ₀ <i>ser de N</i>
SdQ0	820	39	<i>QueF₀ ser de N</i>
SdQ1	308	15	<i>QueF₀ ser de N Prep N₁</i>
SdQ2	37	2	<i>N₀ ser de N Prep QueF₁</i>
SdSIM	55	3	<i>N₀ ser de N Prep N₁</i> [symmetry]
Total	2085		

co-occurring with a human subject and not requiring any complements, for example *O Pedro é corajoso* ‘Pedro is courageous/brave’. These adjectives, henceforth mentioned as *human adjectives*, were classified into 15 classes (Table 5), including adjectives denoting diseases, affiliations, nationalities, and other properties, typically used to qualify human nouns. These were sub-classified into 12 categories, based on: (i) the auxiliary verb with which they co-occur, either *ser*, *estar* ‘be’ or both, such as, *O Pedro é/está arrogante* ‘Pedro is arrogant’; *O Pedro *é/está bronzado* ‘Pedro is tanned’; *O Pedro é/está alucinado* ‘Pedro is hallucinated’; (ii) the possibility of accepting the indefinite construction [IC], that is, of being preceded by an indefinite article *um* ‘a/an’, e.g. *O Pedro é (um) arrogante* ‘Pedro is (an) arrogant’; (iii) the syntactic and semantic nature of the subject selected by each adjective, which can correspond to a human noun (*Nhum*), a complex noun phrase involving an appropriate noun (*NappdeNhum*), or to a finite or non-finite clause. Table 6 illustrates various paraphrases that we formalized involving human intransitive adjectives.

5 From transformational properties to paraphrases

The transformations described in this section are the source for several types of paraphrase. While some transformations’ distribution is very broad and of general use (i.e. clefting, length permutation of complements, etc.), those referred to here are lexically constrained, in the sense that they depend on the predicative element (noun or adjective) of the construction they apply to, even if their distribution is quite wide across the lexicon grammar. In this way, these transformations are highly conditioned by the lexical items involved, and cannot be predicted even given the sentence structure. On the contrary, they have been encoded in the lexicon-grammar of the language, as it was made in the case of the lexicon-grammars presented here. Also, some transformations are specific of a particular sentence structure and support verb, while others pertain to more than one support verb construction type. Tables 2, 3, and 6 summarize the most significant transformations described and illustrated in Sects. 5.1 to 4.3.

Table 4 Paraphrases of support verb constructions with *ser de*

Negative prefix	Paraphrases
[in-N]	<i>O Pedro é de uma certa intolerância à lactose</i> <i>Pedro is of a certain intolerance to lactose</i>
[falta de N] ([lack of N])	<i>O Pedro é de uma certa falta de tolerância à lactose</i> <i>Pedro is of a certain lack of tolerance to lactose</i>
NP restructuring	
[NP Restr]	<i>As formas desse de desenho são de uma certa assimetria</i> <i>The forms of that drawing are of a certain asymmetry</i>
[Restr GN0]	<i>Esse desenho é de uma certa assimetria nas suas formas</i> <i>That drawing is of a certain asymmetry in its forms</i>
[Restr Nop QueF] (como o Pedro age)	<i>O modo de o Pedro agir é de uma certa teimosia</i> <i>Pedro's way of acting is of a certain stubbornness</i> <i>O Pedro é de uma certa teimosia no modo de agir</i> <i>Pedro is of a certain stubbornness in his way of acting</i>
[Rest Nop QueF]	<i>O Pedro falar é de uma grande coragem</i> <i>Pedro speaking is of a great courage</i> <i>O Pedro é de uma grande coragem em falar</i> <i>Pedro is of a great courage in speaking</i>
Symmetry	
[Symmetric-Coord]	<i>N0 é de uma certa simetria com N1</i> <i>N0 is of a certain symmetry with N1</i> <i>N0 e N1 são de uma certa simetria (um com o outro)</i> <i>N0 and N1 are of a certain symmetry (with each other)</i>
Classifier Noun	
[Nclass=pessoa]	<i>O Pedro é um indivíduo de um certo altruísmo</i> <i>Pedro is an individual of a certain altruism</i> <i>O Pedro é de um certo altruísmo</i> <i>Pedro is of a certain altruism</i>
VSup Replacement	
[ser de=ter]	<i>O Pedro é de um certo altruísmo</i> <i>Pedro is of a certain altruism</i> <i>O Pedro tem um certo altruísmo</i> <i>Pedro has a certain altruism</i>
[ser de=haveer]	<i>O Pedro é de um certo altruísmo</i> <i>Pedro is of a certain altruism</i> <i>Há (da parte do)/no Pedro um certo altruísmo</i> <i>There is a certain altruism in Pedro</i>
[ser de=fazer]	<i>O Pedro é de uma certa gentileza (para com N1)</i> <i>Pedro is of a certain kindness (with N1)</i> <i>O Pedro faz uma/a gentileza (para com N1=lhe)</i> <i>Pedro is kind (with N1=him)</i>
Nominalization	

Table 4 continued

Negative prefix	Paraphrases
[ser de Npred=ser Adj]	<i>O Pedro é de uma certa cortesia (para com N1)</i> <i>Pedro is of a certain courtesy (toward N1)</i> <i>O Pedro é cortês (para com N1)</i> <i>Pedro is courteous (toward N1)</i>
Object pronoun [a N1 = Dat1]	<i>O Pedro é de uma grande obediência aos pais</i> <i>Pedro is of great obedience to his parents</i> <i>O Pedro é-lhes de uma grande obediência</i> <i>Pedro is of great obedience to them</i>
Possessive pronoun [de N1 = Poss1]	<i>(Fazer isso) é do interesse do Pedro</i> <i>(To do this) is Pedro's interest</i> <i>(Fazer isso) é do seu interesse</i> <i>(Doing so) is in his best interest</i>

Table 5 Distribution of adjectives by table attribute after integration into Port4NooJ

Class	Count	%	Structure
SAHP1	818	16	<i>(Nhum/Napp de Nhum/QueF)₀ ser Adj [-IC]</i>
SAHP2	498	10	<i>(Nhum/Napp de Nhum)₀ ser Adj [-IC]</i>
SAHP3	435	8	<i>(Nhum)₀ ser Adj [-ICC]</i>
SAHC1	441	9	<i>(Nhum/Napp de Nhum/QueF)₀ ser Adj [+IC]</i>
SAHC2	238	5	<i>(Nhum/Napp de Nhum)₀ ser Adj [+IC]</i>
SAHC3	561	11	<i>(Nhum)₀ ser Adj [+IC]</i>
EAHP2	137	3	<i>(Nhum)₀ estar Adj [-IC]</i>
EAHP3	317	6	<i>(Nhum/Napp de Nhum)₀ estar Adj [-IC]</i>
SEAHP2	209	4	<i>(Nhum/Napp de Nhum)₀ ser/estar Adj [-IC]</i>
SEAHP3	158	3	<i>(Nhum)₀ ser/estar Adj [-IC]</i>
SEAHC2	62	1	<i>(Nhum/Napp de Nhum)₀ ser/estar Adj [+IC]</i>
SEAHC3	72	1	<i>(Nhum)₀ ser/estar Adj [+IC]</i>
SAN	676	13	<i>(Nhum)₀ ser Adj [-IC]</i>
SAF	326	6	<i>(Nhum)₀ ser Adj [±IC]</i>
SEAD	203	4	<i>(Nhum/±Napp de Nhum)₀ ser/estar Adj [±IC]</i>
Total	5151		

Table 6 Paraphrases of constructions involving human intransitive adjectives

Morphologically-related predicates	Paraphrases
[Pred-Adj]	<i>O Pedro está zangado</i> <i>Pedro is angry</i>
[Pred-Verb]	<i>O Pedro zangou-se</i> <i>Pedro got (himself) angry</i>
[Pred-N]	<i>O Pedro envolveu-se numa zanga</i> <i>Pedro got involved in a fight</i>
Predicate Adj's Copula Verbs	
[Pred-Adj=estar]	<i>O Pedro está perdido</i> <i>Pedro is lost</i>
[Pred-Adj=andar]	<i>O Pedro anda perdido</i> <i>Pedro goes lost</i>
Nationality and Affiliation Adj	
[Origin-Adj]	<i>Os rapazes são de origem portuguesa</i> <i>The boys are of Portuguese origin/roots</i>
[Nationality-Adj]	<i>Os rapazes são portugueses</i> <i>The boys are Portuguese</i>
[Country-Adj]	<i>Os rapazes são de Portugal</i> <i>The boys are from Portugal</i>
[Membership-Adj]	<i>Os rapazes são benfiquistas</i> <i>The boys are Benfica fans</i>
[Adept-Adj]	<i>Os rapazes são do Sport Lisboa e Benfica</i> <i>The boys are fans of Sport Lisboa e Benfica</i>
Cross-construction	
[Cross-Adj]	<i>O idiota do rapaz está sempre em apuros</i> <i>The idiot of the boy is always in trouble</i>
[Adj-SVC]	<i>O rapaz é (um) idiota</i> <i>The boy is (an) idiot</i>
Appropriate Noun	
[N-App-Adj]	<i>Ele foi moderado nos seus comentários</i> <i>He was moderated in his comments</i>
[Pred-Adj]	<i>Os seus comentários foram moderados</i> <i>His comments were moderated</i>
[N-App= θ]	<i>Ele foi moderado</i> <i>He was moderated</i>
Generic Noun Phrases	
[N-Adj]	<i>Ele é um indivíduo estúpido</i> <i>He is a foolish individual</i>
[N θ -Det-Adj]	<i>Ele é um estúpido</i> <i>He is a fool</i>
[N θ -Adj]	<i>Ele é estúpido</i> <i>He is fool</i>

5.1 Reduction of support verb construction to complex noun phrase

An important transformational process pertaining to the predicate nouns with *fazer* ‘make’/‘do’ concerns the reduction of the support verb and the formation of a complex noun phrase, whose head is the predicate noun, along with all its arguments as complements, within the context of a relative clause (Gross, 1981). This operation is often called *reduction of support verb* [RedVsup] or *reduction of relative*.

Take an active sentence [Active S] such as *O Pedro fez uma viagem* ‘Pedro made a trip’, which can be paraphrased by a relative clause, like *a viagem que o Pedro fez* ‘the trip that Pedro made’. In this noun phrase, the relative pronoun and the support verb can undergo further reduction, leaving the subject *Pedro* as a complement of the predicate noun, linked by preposition *de* ‘of’, e.g. *a viagem do Pedro* ‘Pedro’s trip’. We call this an active complex noun phrase [Active NP], and this transformation is a way of reducing to the barest form the elements composing the semantic predicate expressed by the predicate noun and its arguments. The NP may then be inserted as an argument of a further predicate: [*A viagem do Pedro*] foi muito agradável ‘Pedro’s trip was very pleasant’. In fact, any (prepositional) complement the predicate noun may have is normally kept in the noun phrase, along with the preposition that introduces it, as in the derivation: *O Presidente fez um discurso ao país* ‘The president made a speech to the country’ = *o discurso do Presidente ao país* ‘The President’s speech to the country’.

Moreover, from an active sentence such as *Alguém fez a contagem dos votos* ‘Someone did the counting of the votes’ (i.e. ‘the vote count’), the corresponding passive sentence [Passive S] can be derived, v.g. *A contagem dos votos foi feita por alguém* ‘The vote count was made by someone’. This, in turn, can also become a relative clause modifying the predicate noun, *A contagem dos votos que foi feita por alguém* ‘The vote count that was made by someone’. Passive clauses being adjective-like, the relative pronoun and the auxiliary verb *ser* ‘be’ are often reduced and the past participle of the support verb becomes a modifier of the predicate noun: *A contagem dos votos feita por alguém* ‘The vote count made by someone’. But, unlike ordinary (distributional, full) verbs, the support verb can undergo further reduction, yielding a passive complex noun phrase [Passive NP]: *A contagem dos votos por alguém* ‘The vote count by someone’.

These formal operations are all very general, and affect a large subset of the lexicon-grammar, thus, making it possible to produce a large number of paraphrases from the base sentence form. The verb support reduction and formation of the complex noun phrase crucially depend on the support verb basic sentence structure. It is a very common operation with the support verb *fazer* (Chacoto, 2005) and *ter* (Santos, 2015a), less so with *dar* (Rassi, 2015) [see also Rassi et al. (2012, 2013), for an overview on the description of Brazilian Portuguese predicate nouns’ lexicon-grammars for these support verb constructions]. The [RedVsup] operation does not apply altogether to the support verb *estar* (Ranchhod, 1990) and to the support verb *ser de* (Baptista, 2005b). In these last two construction types, the support verb can, in fact, be zeroed, but another structure is formed, not a noun phrase (see Sect. 5.10).

5.2 Dative restructuring

Dative restructuring (Leclère, 1995) is a special case of the more general operation named *noun phrase restructuring* (Guillet & Leclère, 1981), which is a very productive phenomenon, affecting not only support verb and predicate noun constructions, but also verbal and adjectival structures. This transformation splits a complex noun phrase of the form $[N_a \text{ de } N_b]$ into two constituents $[N_a][a N_b]$, that is, the noun complement *de* N_b is morphed into a dative complement aN_b , becoming more closely linked to the support verb. Usually, a meronymy relation exists between N_a and N_b . Though in the Harrisian framework transformations are equivalence relations, here we assume that the complex NP= $[N_a \text{ de } N_a]$ is the base form, whose head noun (N_a) is selected by the predicate noun. This N_a is the noun designating the ‘part’ element of the meronymy relation. In the case of ‘body part’ nouns (N_{bp}), these always imply an ‘inalienable possession’ relation with another entity (N_b), interpreted as the ‘whole’ that the N_a belongs to. For example, in the sentence *O Pedro fez uma festinha em[_a cara da Joana]* ‘Pedro did a caress in the face of Joana’, the complex noun phrase is restructured into two complements, e.g., *O Pedro fez uma festinha em[_a cara] de[_a Joana]*.⁹

5.3 Symmetry restructuring

Symmetric predicates are defined as having two arguments that play the same semantic role in relation to that predicate; in other words, the semantic relation between the arguments and the predicate they depend on can be construed as ‘symmetric’. Because of that, the arguments can both change their syntactic slots, e.g. *O Pedro fez um acordo com o João* = *O João fez um acordo com o Pedro* ‘Pedro made a deal with João’ = ‘João made a deal with Pedro’; or be coordinated in the same syntactic slot, e.g. *O Pedro fez um acordo com o João* = *O Pedro e o João fizeram um acordo (um com o outro)* ‘Pedro made a deal with João’ = ‘Pedro and João made a deal (with each other)’, without change of the global meaning of the sentence. Furthermore, in the case of the coordinated structure above, the so-called *echo complement* (*um com o outro* ‘with each other’) is facultative, which is not the case with mere reciprocal constructions. In this sense, symmetrical predicates are ‘intrinsically reciprocal’.

Symmetry is a pervasive property of all predicative categories (verbs, adjectives and nouns) and all types of support verb constructions have one or more symmetric lexicon-grammatical classes [see Baptista (2005a) for an overview in Portuguese]. There are at least two types of symmetry: (i) *subject-complement* symmetry, as in *O Pedro namora com a Joana* ‘Pedro is dating [with] Joana’ → *O Pedro e a Joana namoram (um com o outro)* ‘Pedro and Joana are dating ([with] each other)’; and (ii) symmetry *between complements*, such as in *O Pedro misturou o azul com o*

⁹ The underscore indicates that two lexical units (preposition and definite article), normally contracted, were split here for clarity purposes.

¹⁰ This is also a paraphrase of *O Pedro fez uma festinha à Joana na cara* ‘Pedro did a caress to Joana in the face’.

amarelo ‘Pedro mixed the blue with the yellow’ → *O Pedro misturou o azul e o amarelo* ‘Pedro mixed the blue and the yellow’. As far as paraphrase processing is concerned, symmetric constructions imply recognizing/generating both the sentences with distinct symmetric arguments expressed and the sentences where they have been coordinated, independently of the syntactic slot those arguments occupy. Furthermore, as coordinated noun phrases can give rise to plural forms, deriving the detached NPs from such plurals is not a trivial task. For example, from a sentence like *O Pedro misturou bem as cores* ‘Pedro mixed well the colors’ (i.e., ‘Pedro mixed the colors well’), one can derive the theoretically valid source sentence *O Pedro misturou uma cor com outra cor* ‘Pedro mixed well one color with another color’, however, referential information is insufficient in the plural noun phrase to allow for a safe paraphrasing of that sentence (more than two colors could have been mixed, for instance).

5.4 Conversion

Predicate noun constructions with two arguments often allow *Conversion* (Gross, 1989), an operation where the arguments are rearranged around the predicate noun, the later is construed with another support verb, and the preposition introducing the complement also changes; e.g. *O Pedro fez um telefonema ao João* ‘Pedro made a phone call to João’ → *O João recebeu um telefonema do Pedro* ‘João received a phone call from Pedro’. This operation does not change the global meaning of the sentences, namely the semantic roles of the arguments of the predicate noun; in both these sentences, *Pedro* is the AGENT and *João* the PATIENT/GOAL of the predicate noun *telefonema* ‘phone call’. Rather, conversion can be used to present the process denoted by the predicate noun in a two-fold way and provide a different salience for its arguments; namely, it puts the non-AGENT argument in the salient syntactic position of subject, while diminishing the AGENT argument by framing it in a prepositional phrase complement. In this way, conversion is similar or corresponds to the active-passive opposition, found in verbal constructions, though remaining a specific transformation of nominal predicates.¹¹ Active-like sentences, where the semantic role of AGENT is aligned with the subject syntactic slot are called *standard*, while the passive-like, with a PATIENT/GOAL subject, are called *converse* constructions.

All major support verb constructions allow for some type of conversion. Most prominently, constructions with support verb *dar* ‘give’ are particularly apt to express the active-like orientation of the predicate. Two major subsets have been found (Baptista, 1997): (i) those involving the standard-converse pair of support verbs *dar-receber* ‘give-receive’, e.g. *O Pedro deu um forte apoio ao João* ‘Pedro gave a strong support to João’ = *O João recebeu/*levou/*apanhou um forte apoio do Pedro*¹² ‘João received/took/got a strong support from Pedro’; and (ii) those involving the verbs *dar-levar/apanhar* ‘give-receive’, e.g. *O Pedro deu um forte soco ao João* ‘Pedro gave a strong punch to João’ = *O João ?recebeu/levou/apanhou um forte soco do Pedro* ‘João received a strong punch from Pedro’. This is also the case with predicate nouns with *ser de* ‘be of’, e.g. *O Pedro é de uma confiança total no*

¹¹ Many support verb construction can undergo passive as well.

João ‘Pedro is of a complete trust on João’ = *O João tem a confiança total do Pedro* ‘João has the complete trust of Pedro’.

By providing different salience to arguments of the same predicate, along with different surface ordering of the sentence main elements, conversion is a very important discursive paraphrastic mechanism.

5.5 Support verb variants

As the predicate noun can be considered the center of the semantic predicate, support verbs often present lexical variants, that is, other support verbs can be used to construe the same noun. The selection of these variants¹³ is made by the noun, which keeps the same overall meaning, and the same distributional constraints on their arguments. Generic (also called elementary/generic support verbs, with broad lexical distribution (i.e. selected by a large number of predicate nouns) include *fazer* ‘do’/‘make’, *ter* ‘have’, *dar* ‘give’, *estar Prep* ‘be Prep’, *ser* ‘be’ and *ser de* ‘be of’.

The set of variants was also deemed an important feature to distinguish semantically homogeneous subsets of the lexicon of predicate nouns, such as those designating violent actions (Baptista, 2004). Most of these nouns accept conversion (see Sect. 5.4), using the verbs *dar* ‘give’ and *pregar* ‘throw’ in the standard (active-like) construction: *O Pedro deu/pegou um murro ao João* ‘Pedro gave/throw a punch to João’ = ‘Pedro punched João’; while selecting the *apanhar* ‘get’ e *levar* ‘take’ in the converse construction *O João apanhou levou um murro do Pedro* ‘João got/ received a punch from Pedro’.

In some cases, either the support verb or the set of its variants is so predicate-specific, that it can be used to delimit semantically coherent subsets of the lexicon. This is the case of *cometer* ‘commit’, used as a variant of the support verb *fazer* ‘do’/‘make’, which is primarily used for predicates designating crimes or other unlawful deeds, e.g., *O Pedro cometeu um crime/roubo/pecado horrível* ‘Pedro committed a horrible crime/theft/sin’. It is also the case of the support verb *praticar* ‘practice’, used for predicates designating sports, e.g., *O Pedro pratica atletismo/futebol/karatê/natação/yoga* ‘Pedro practices athletics/football/karate/swimming/yoga’.

Within the variants of an elementary/generic support verb, two main types can be identified: (i) *stylistic* variants, such as those shown above for *fazer* ‘do’/‘make’ and *dar* ‘give’; and (ii) *aspectual* variants, that is, support verbs who introduce an aspectual nuance when contrasted with the elementary support verb. Examples of the later are, for the support verb *fazer* ‘do’/‘make’, *iniciar* ‘begin’/‘initiate’, *continuar/prosseguir* ‘continue’/‘proceed with’ or *terminar/concluir* ‘end’/‘conclude’, e.g., *O Pedro fez/iniciou/continuou/prosseguiu com/concluiu/terminou a transferência bancária* ‘Pedro made/begun/ initiated/continued/proceeded with/ ended/concluded the banking transfer’. For *ter* ‘have’, one can find *ganhar* ‘gain’,

¹² The asterisk ‘*’ signals the sentence unacceptability, while the question mark indicates doubtful acceptability.

¹³ Even though variants are also support verbs, they may feature syntactic properties of their own, so a detailed description is in order.

manter ‘keep’ or *perder* ‘lose’ as aspectual variants, e.g., *O Pedro ganhou/manteve/perdeu a esperança* ‘Pedro gained/grew/kept/lost hope’.

Finally, an important feature, specific to predicate noun constructions and also directly bearing on the study of paraphrasing mechanisms within the language, consists of the fact that many predicate nouns can be construed with more than one elementary support verb (Gross, 1981, p. 33 ff.), while maintaining the same core of meaning and (largely) the same distributional constraints.

Hence, for example, the noun *festinha* ‘caress’ can select both support verbs *fazer* ‘do’/‘make’ and *dar* ‘give’, e.g., *O Pedro fez/deu uma festinha à Joana* ‘Pedro did/gave a caress to Joana’. Some of these often paired constructions include *ser de/ter*, e.g., *O Pedro é de/tem uma grande coragem* ‘Pedro is of/has a great courage’; *estar com/ter*, e.g., *O Pedro está com/tem muito frio* ‘Pedro is with/has much cold (=Pedro is cold)’. These independent (but semantically equivalent) constructions may show specific syntactic properties (Ranchhod, 1990, pp. 168–172), and so they require independent description. Though they can map into each other in a relatively regular way, these pairings are not systematic, and must be established for each noun.

One of the major findings of the study of predicate nouns with *ser de* ‘be of’ was the absence of stylistic and aspectual variants. Even the inchoative auxiliary verb *tornar-se* ‘become’, often a substitute for *ser* ‘be’ in other contexts, is seldom observed. On the contrary, these predicate nouns may have equivalent constructions with other elementary support verbs, mostly *ter* ‘have’, e.g., *O Hugo é de/tem uma grande coragem* ‘Hugo is of/has a great courage’; or *haver* ‘there be’, e.g., *Há no Hugo uma grande coragem* ‘There is a great courage in Hugo’; and, more rarely, with *fazer* ‘do’/‘make’, e.g., *O Hugo é de uma certa gentileza para com a Eva* ‘Hugo is of a certain kindness to/towards Eva’, which can be considered a paraphrase¹⁴ of *O Hugo fez uma gentileza para com a Eva* ‘Hugo did a certain kindness to Eva’.

In constructions headed by adjectives, the copula verb supporting the adjective may also be replaced by other auxiliary copula verb, resulting in paraphrase generation. For example, the predicate adjective *perdido* ‘lost’, can be supported both by an elementary copula verb, in this case *estar* ‘be’ [Pred-Adj=estar], and specific aspectual verbs denoting an equivalent meaning, namely *andar* ‘go’ [Pred-Adj=andar]. So, *estar perdido* ‘be lost’ and *andar perdido* ‘go lost’ are paraphrases. The ambiguity exists in both sentences, where *perdido* ‘lost’ can be used to express a physical or psychological state or condition.

5.6 Nominalizations

An important source for the analysis of paraphrastic relations among sentences is Nominalization, that is, equivalence relation between sentences with a predicate

¹⁴ To be precise, the *ser de* construction expresses not only a human quality, but it also characterizes the attitude or a gesture from the subject towards the human complement, e.g. *A atitude/o gesto do Pedro foi de uma certa gentileza* ‘Pedro’s attitude/gesture is of a certain kindness’. On the other hand, the sentence with *fazer* is not strictly semantically equivalent, as the paraphrase involves a regular meaning difference, where the expression of a human quality is, at least, not so obvious, and only the second interpretation of the *ser de* construction is kept. They can be treated as **approximate** paraphrases and the difference is systematic.

noun and a support verb, on one hand, and a verb or an adjective (and its auxiliary verb), on the other hand. The study of nominalizations within the transformational framework can be traced back to Harris (1964). In his famous example, the author proposes an elementary transformation to establish the paraphrastic equivalence between sentences like *O Pedro estuda eclipses* 'Pedro studies eclipses' = *O Pedro faz estudos sobre eclipses* 'Pedro makes studies on eclipses'.

For the most part, the predicate nouns with support verb *ser de* correspond to the nominalization of adjectival constructions, e.g., *O Pedro foi de uma grande crueldade para com o João* 'Pedro was of a great cruelty towards João' = *O Pedro foi muito cruel para com o João* 'Pedro was very cruel to João'. More rarely, a verbal construction can be found: *O Pedro foi de uma grande compaixão para com o João* 'Pedro was of (=had) a great compassion for João' = *O Pedro compadeceu-se do João* 'Pedro took pity on João'.

It should be noticed, however, that the lexical-morphological relation between a predicate noun and a verb or between a noun and an adjective is necessary but insufficient to establish a nominalization with a transformational status. Not only the meaning of the sentences being related must be the same, but the distributional constraints of the predicate noun on its argument domain must be similar. Establishing such paraphrastic status requires highly granular, and systematic linguistic description. For example, consider the verb *contar* 'to count' and its multiple word senses ('to count', 'to rely on', etc.). One of its senses has a nominalization with the noun *conta* 'count', which can be used in the singular or in the (bare) plural form: *O Pedro contou quantos livros tinha* 'Pedro counted how many books he had' = *O Pedro fez uma conta/contas de quantos livros tinha* 'Pedro made the count/counts of how many books he had'; but another verb sense, in spite of the obvious morphological relation, can not be associated with the same nominalization where only the bare plural form exists *O Pedro contou com o João para esta tarefa* 'Pedro counted on João for this task' = *O Pedro fez *a conta/contas com o João para esta tarefa* 'Pedro made the count/counts on João for this task'. Hence, for each predicate noun, the corresponding nominalizations (verbs and adjectives) must be (and have been) provided in the lexicon-grammar table.

Concerning adjectival constructions, several 3-tuples involving adjective, noun and verb, morphologically and syntactically related constructions were found to be related by nominalization. For example, the predicate adjective *isolado* 'isolated', and the predicate noun *isolamento* 'isolation' are morphologically related to the predicate verb *isolar-se* 'isolate (oneself)'. The predicate adjective is supported by the copula verb *estar* 'be', e.g., in *O Pedro está isolado* 'Pedro is isolated', or its aspectual variants like *continuar* 'remain', e.g., *O Pedro continua isolado* 'Pedro remains isolated'; the predicate noun can be supported also by the verb *estar* 'be', e.g. *O Pedro está em isolamento* 'Pedro is in isolation'. However, it should be noticed that some adjectives do not have a morphologically related verb and/or noun (e.g. *Ele está cabisbaixo* 'He is crestfallen'), while others can only be paraphrased by a predicate noun construction, e.g. *Ele está faminto* = *Ele tem fome* 'He is starving' = 'He is hungry'.

5.7 Negation prefixes and other negation devices

The insertion of negation/negative prefixes (mainly *des-* and *in-*, and rarely *a-*) on predicate nouns may involve differences in the meaning and in the syntactic proprieties of lexicon-grammar entries. Strictly speaking, and regardless of the morphological (historical) process involved, whenever significant differences are found, the prefixed and the base forms should be treated as independent lexicon-grammatical entries.

The nouns accepting these prefixes were collapsed as a single entry, e.g., *Este fenómeno é de uma certa (in-)vulgaridade* ‘This phenomenon is of a certain (in)vulgarity’; while nouns for which no transformational (syntactic-semantic) relation could be established were kept as distinct entries, e.g., *O Pedro é de uma certa (*in-)vulgaridade para com o João* ‘Pedro was of a certain (*in)vulgarity towards João’ (= ‘Pedro was offensive/had rude manners towards João’). The paraphrastic relation between the surface (base) constructions and the derived noun can thus be exploited, as this paper suggests, in several applications.

A complementary but analytical form of negation involves the expression *falta de* ‘lack of’ (about 140 nouns in the *ser de* ‘be of’ constructions); it also applies to predicate nouns with support verb *ter* ‘have’; most of these predicates express human qualities/attributes, e.g., *O Pedro é de/tem uma (falta de) memória prodigiosa* ‘Pedro is of/has a prodigious (lack of) memory’. Naturally, this only occurs with nouns with a *positive* polarity, like *coragem* ‘courage’; predicate nouns with a *negative* polarity, like *cobardia* ‘cowardice’, do not accept this construction; e.g. *O Pedro é de/tem uma certa (*falta de) cobardia* ‘Pedro is of/has some (*lack of) cowardice’. This relation provides a formal base for this semantic distinction, not always easy to determine, and it can be seen as an alternative mechanism to prefixation, being particularly apt for nouns that, for lexical, morphotactic, phonotactic or historical reasons, do not accept the negative prefixes. Negation with *falta de* ‘lack of’ has also been systematically encoded in the lexicon-grammar.

5.8 Appropriate nouns and noun phrase restructuring

In many cases, a complex noun phrase (NP) is found as the subject of the sentence with *ser de* ‘be of’. The head of such subject NP is also a predicate noun and it may be present along with its arguments, particularly its semantic/notional ‘subject’ argument, in the form of a prepositional phrase (PP) introduced by the preposition *de* ‘of’: [*A disposição destes objetos*] *é de uma certa assimetria* ‘The placement of these objects is of a certain asymmetry’. A NP restructuring operation (Guillet & Leclère, 1981) is then found, which splits the noun phrase into two distinct constituents: (i) the head of the PP becomes the sentence subject, while (ii) the predicate noun (formerly the head of the subject NP) is moved to a new PP, usually at the end of the sentence; thus, this operation relates more closely the semantic ‘subject’ of the predicate noun to the sentence’s main predicate, while downgrading the predicate noun to a peripheral position in the sentence: [*Estes objetos*] *são de uma certa assimetria [na sua disposição]* ‘These objects are of a certain asymmetry in their placement’. In the new PP, the reference of the possessive determiner (*sua*

‘their’, in the example) is constrained, and it has to refer to the subject of the predicate noun; obviously, this possessive cannot be derived from a free PP, non-correferent to the subject. The semantic roles of the elements are kept exactly the same in spite of the formal changes the sentence undergoes. An *appropriate* relation, in the sense of Harris (1976, pp. 113–115), usually exists between the predicate noun in the subject position and the sentence’s main predicate, which explains this transformation, as well as the possibility of zeroing the restructured PP without losing information. *Estes objetos são de uma certa assimetria (na sua disposição)* ‘These objects are of a certain asymmetry (in their placement)’. The predicate noun is, then, considered to be in an *appropriate position*, i.e., it is the most likely noun to appear as an argument of the main predicate, thus contributing little or nothing to the information conveyed by the sentence, hence it may be zeroed. By establishing the paraphrastic relation, the lexicon-grammar keeps track of the operator-argument semantic relations between the NP head (a predicate noun) and its argument(s).

Human intransitive adjectives used in appropriate noun constructions are also a rich source of paraphrastic units. For example, the sentence *Os comentários do Pedro foram moderados* ‘Pedro’s comments were moderated’ includes an appropriate noun *comentários*. This noun can be found both in the subject position *os seus comentários* ‘his comments’ [N-App-Adj] or after the adjective, through this restructuring transformational process *O Pedro foi moderado nos seus comentários* ‘Pedro was moderated in his comments’ [Pred-Adj]. In both constructions, the appropriate noun can be reduced, and the subject be represented only by the proper noun: *O Pedro foi moderado* ‘Pedro was moderated’, or corresponding pronoun: *Ele foi moderado* ‘He was moderated’ [N-App= θ].

5.9 Manner sub-clauses restructuring

An interesting distributional constraint regards the subject NPs with manner operator nouns (Gross, 1975) *forma*, *maneira* and *modo* ‘manner’/‘way’ (in Brazilian Portuguese, there is also the operator noun *jeito* ‘idem’). These operators can be construed with an infinitive sub-clause complement introduced by the preposition *de* ‘of’: *A formala maneiralo modo de o Pedro fazer isso é de uma arrogância impressionante* ‘The way of Pedro doing this is of an impressive arrogance’; or a *pseudo-relative*, *finite* clause, introduced by the so-called interrogative adverb *como* ‘how’: *a formala maneira/ o modo como o Pedro faz isso é de uma arrogância impressionante* ‘The way how Pedro does this is of an impressive arrogance’. When the predicate noun accepts these operator nouns, these sentences are used to qualify the way a process takes place or the manner in which an action is performed, rather than expressing the attributes of a person or an object.

A similar NP restructuring as seen above (Sect. 5.8) operates on these manner constructions, splitting the complex NP, extracting the subject of the subordinate clause to the subject of the predicate noun and leaving the operator noun as a PP manner complement: *O Pedro é de uma arrogância impressionante em_a formala maneiralo modo de fazer isso* ‘Pedro is of an impressive arrogance in the way of

doing that' = *O Pedro é de uma arrogância impressionante em_a formala maneiral o modo como faz isso* 'Pedro is of an impressive arrogance in the way [he] does that'. Notice, again, that the zeroed subject of the subordinate clause resulting from the NP restructuring is obligatorily co-referent of the 'extracted' subject of the predicate noun.

5.10 Reduction to noun adjunct

Finally, another interesting property of *ser de* 'be of' nominal constructions arises from the fact that they do not yield complex NPs, like, for example, the predicate nouns built with *fazer* 'do'/'make' or *ter* 'have'. On the contrary, being for the most part nominalizations of adjectival predicates, and being introduced by preposition *de* 'of', these predicate nouns can function as an adnominal adjunct of a common noun: *O Pedro é um rapaz/homem/indivíduo/tipo de uma certa arrogância* 'Pedro is a boy/man/individual/guy of a certain arrogance', most likely obtained from the reduction of complex sentence with a relative clause: *O Pedro é um rapaz/homem/indivíduo/tipo # que é de uma certa arrogância* 'Pedro is a boy/man/individual/guy that is of a certain arrogance'.

5.11 Transformations of nationality, affiliation, and membership adjectives

Adjectives that describe nationality, affiliation and membership relations are also a source of paraphrases, as illustrated in the the following examples: *de origem portuguesa* ('of Portuguese origin/roots') = *portugueses* 'Portuguese'=*de Portugal* 'from Portugal'; *benfiquistas* 'Benfica fans'=*do Sport Lisboa e Benfica* 'fans of Sport Lisboa e Benfica'. We have already illustrated eSPERTO paraphrasing of these types of adjectives in Fig. 1. All these adjectives can modify a specific set of appropriate nouns (e.g. origin, nationality, place of birth, ethnicity or (human) race), which can always be lexically reduced (e.g. *Os indivíduos de origem portuguesa* = *Os portugueses* 'Individuals of Portuguese origin' = 'The Portuguese'). Moreover, these adjectives can also fill the head of a noun phrase, by removing the generic human noun they modify (*Os indivíduos portugueses* = *Os portugueses* 'The Portuguese individuals' = 'The Portuguese'). In addition, these adjectives can be paraphrased by a nominal complement where the morphologically related noun is introduced by the preposition *de* (e.g. *Os adeptos benfiquistas* = *Os adeptos do Benfica*, literally, 'The Benficana adepts' = 'The supporters/fans of Benfica').

5.12 Cross-constructions

Some intransitive adjectives, particularly the ones denoting negative properties, such as *idiota* ('idiot'), may occur in cross-constructions, where the adjective fills the head of a noun phrase. In this case, cross constructions can be paraphrased by both a regular copula construction, such as *O idiota do Pedro* ('The idiot of Pedro') = *O Pedro é idiota* ('Pedro is idiot'), and an indefinite construction *O idiota do Pedro* ('The idiot of Pedro') = *O Pedro é um idiota* ('Pedro is an idiot'), with

identical meaning. As described below, in indefinite constructions, the adjective is preceded by an indefinite article.

5.13 Indefinite constructions

Some human intransitive adjectives can superficially fill the head of a noun phrase. This analysis is based on the fact that it is always possible to reconstruct the human noun to which the adjective is related to. For example, the sentences *és um indivíduo estúpido* ‘you are a stupid person’ and *és um estúpido* ‘you are (a) stupid (one)’ are paraphrases. In this case, the human generic noun is elided and, by consequence, the adjective appears after the indefinite article. As previously mentioned, the adjective can also be used in an ordinary predicative construction where the determiner (indefinite article) is not present. So, *ele é estúpido* ‘he is fool’ may function as a paraphrase of both the sentence where the noun phrase contains the generic human *indivíduo* ‘person’/‘guy’ and the sentence where it was elided.

6 Integration of the lexicon-grammar resources

Our first experiments integrating lexicon-grammars into Port4NooJ envisaging improving its paraphrasing capabilities started arbitrarily with integrating the lexicon-grammar of human intransitive adjectives, followed by integrating the lexicon grammar of predicate nouns that occur with the support verb *fazer* and then with support verb *ser de*. This integration is a two step process. In the first step, the lexicon-grammar tables are converted semi-automatically into a NooJ standalone dictionary, and, in the second step, paraphrasing grammars are created manually given the transformational properties formalized in those grammars.

As referred in Mota et al. (2015), which describes the detailed process of integrating the lexicon-grammar tables of human intransitive adjectives into Port4NooJ, the new standalone dictionary of human intransitive adjectives includes 5151 entries, corresponding to 4138 different adjectives. Given that only 26% of the adjectives formalized in the lexicon-grammar tables existed initially in Port4NooJ, the number of different adjectives in Port4NooJ increased about 50%. Table 5 shows the distribution of adjectives in the new standalone lemma dictionary by table attribute sorted by the most frequent attribute in the dictionary.

Some tables include information about the nouns and/or verbs morphological and semantically related to those adjectives. The derivation between the adjective and the noun or verb was automatically assigned from the existing derivation paradigms in part of Port4NooJ. In cases where the derivation did not exist, new derivational descriptions (1202) were created.

A dictionary of toponyms with 676 entries was derived from the adjectival entries marked with the attribute +Table=SAN, which marks geographical adjectives that are derived from country names or other toponyms. Each toponym includes the attribute =Adj, which is assigned the adjective derived from that toponym, and also the attribute +TopDET, which is assigned the determiner that most likely occurs before the toponym.

This dictionary does not distinguish the types of toponym but that information can be inferred from the corresponding adjective properties. For example, *tunisino* ‘Tunisian’ has the attribute +NclassPnacionalidade, but *nórdico* ‘Nordic’ does not, which means that the first *Tunísia* ‘Tunisia’ is a country, but *Norte da Europa* ‘North Europe’ is not:

27% of SAN adjectives are indeed derived from country names. A first set of grammars was constructed to extend eSPERTo’s paraphrastic knowledge. These grammars recognize and generate paraphrases of (i) constructions involving patronymic adjectives, (ii) characterizing indefinite constructions, (iii) the possibility of alternating the copula verbs *ser* and *estar* with other aspectual variants, and (iv) cross constructions.

The integration of the lexicon-grammar tables formalizing the properties of nominal predicates into Port4NooJ is also still currently underway, but some of the initial integration challenges can be found in Mota et al. (2018). Out of the 17 lexicon-grammar tables formalized by Chacoto (2005), 14 have been already integrated in Port4NooJ. For now, the three tables that formalize medical technical terms were left out for a second stage. The first attempt to convert the tables into a standalone Port4NooJ dictionary resulted in 5998 nominal entries, corresponding to 1610 different noun lemmas. Most entries already existed in Port4NooJ (63%) which corresponds to about a 5% increase in nominal entries in Port4NooJ from 11,719 different noun lemmas in the previous version of Port4NooJ (i.e., before removing entries corresponding to nominal predicates that are nominalizations). Table 1 shows the distribution of nominal predicates in the lemma dictionary by table attribute sorted by the most frequent table assigned.

With regards to the integration of the predicate nouns that occur with the support verb *ser de*, one of the major challenges was related to the overlap with the noun entries in Port4NooJ (50% of the predicate nouns already existed in Port4NooJ), a problem that is somewhat already being addressed since we started integrating the lexicon-grammar with support verb *fazer*. Consolidating information from an old entry and the lexicon-grammar table is far from a perfect solution and it needs thorough revision. The other challenge is that 55% of cases where the predicate nouns are linked to equivalent adjectival constructions, the adjectives are homographs of a human intransitive adjectives (HIA) already formalized in the lexicon-grammar of HIA. This is a new problem, as adjectives equivalent to predicate nouns are being treated by derivation. We are unsure about how to harmonize those derived entries with the entries and have not started to tackle the problem.

The integration of the predicate nouns with support verb *ser de* resulted in a standalone dictionary with 2132 predicate nouns corresponding to 1376 different noun lemmas. (Mota et al., 2019) Additionally, 797 entries await revision of inflectional codes of derived adjectives or have format problems that must be fixed before adding them to the final dictionary.

Although we have preliminary versions of the standalone dictionaries of predicate nouns that occur with *fazer* and *ser de* that still needs to be reviewed, we

initiated the process of using the information in the lexicon-grammar tables of these constructions to generate paraphrases. Firstly, Port4NooJ grammars already paraphrase constructions involving support verbs (for example, paraphrasing the verbal construction with the nominal construction and vice-versa, or alternation between the support verb and other stylistic or aspectual verbs). Grammars are being updated with attributes from the new tables and are also being extended to take into account the generation of paraphrases involving this type of constructions. Secondly, the grammars created to generate paraphrases of an active into a passive construction and vice-versa were only paraphrasing sentences involving a verbal predicate. We are modifying these grammars to enable them to generate paraphrases of nominal predicates in sentences and (complex) noun phrases. With the new set of resources, we also started developing new grammars to generate paraphrases of equivalent constructions based on specific properties formalized in the lexicon-grammar tables described in this paper. Among others, we point out the grammars to paraphrase:

- the negative construction using the negative prefix or other forms of negation—this grammar has the particularity that uses two properties in the lexicon-grammar in combination (*PfxNeg* and *Negfaltade*) to make sure the correct paraphrastic relation is established. However, these properties are independent and it might be a limitation of the NooJ formalism. This specificity is currently under revision and tests;
- a noun phrase with a possessive, which are unidirectional, i.e., we identify the noun phrase and generate the possessive, but not the opposite. In order to identify the possessive and generate the appropriate noun phrase one needs a larger context and a more complex analysis of the text, assuming all the information is present to be able to rebuild the noun phrase.

7 Preliminary evaluation of the lexicon-grammar resources

As previously mentioned (cf. Sect. 2), in the scope of the project eSPERTo, we starting to study variety adaptation (or conversion) of paraphrastic units between Portuguese varieties. The corpus that we are using is the e-PACT corpus (Barreiro & Mota, 2017) which is comprised of 2669 pairs of aligned sentences of Portuguese from Portugal and Portuguese from Brazil. All those aligned sentences were extracted from the COMPARA Portuguese-English bilingual corpus (Santos & Inácio, 2006). The extracted pairs resulted from querying COMPARA with the adjectives marked as human intransitive; words marked with support verbs (VSUP), which are added to adjectives and nouns that accept a construction with a support verb; and other predicate nouns. After excluding words that are so frequent they would make COMPARA return random results, the query list was comprised of 186,150 words.

In the e-PACT corpus, 30% of the sentence pairs (802) were annotated contrasting semantically identical multiwords, phrases, and expressions in

Portuguese from Portugal and Portuguese from Brazil, resulted in the Gold CLUE4Paraphrasing, a set of 26,101 paraphrastic unit pairs.

To evaluate the lexicon-grammar resources within eSPERTo, we are using the 801 sentences of Portuguese from Portugal the Gold CLUE4Paraphrasing. The size of the corpus is comparable to the corpus built by Cohn et al. (2008) to develop and evaluate paraphrase systems for English. It offers the opportunity to compare between paraphrases in Portuguese from Portugal and Portuguese from Brazil. The efforts of evaluating the results produced by eSPERTo can be coordinated so that they will enable, on the one hand, the enrichment of information in the Gold Clue4Paraphrasing, and, on the other hand, the information in the corpus can be used to check recall.

As a first step to establish a baseline for the future and establish a methodology for evaluation, we split the first 100 sentences of the 30% e-PACT corpus subset and distribute them among the authors of this paper (author evaluated a set of 20 paraphrases). On those 100 sentences, 90 expressions were identified, generating 155 paraphrases. On average, each identified expression generated 1.74 paraphrases. The identified expressions had 1 paraphrase (54 cases), 2 paraphrases (30 cases), or 6 paraphrases (6 cases). From the 155 paraphrases generated, only 39 expressions (25%) were considered in context as correct within the context in which they occur.

If one considers cases where the expression can not be rephrased with any of the paraphrastic suggestions offered by eSPERTo, the most typical errors found include:

- expressions with different semantics, such as **ninguém ainda [reparou] nele = ninguém ainda [fez reparação] nele* ('nobody noticed him' = nobody made a repair on him'). In the sentence corpus, the verb *reparar* 'notice' does not accept the paraphrase *fazer uma reparação* 'make a repair';
- a different part-of-speech was identified as a verb, because the grammars are applied without disambiguation being run first, so, in *para cada [teste]* 'for each test', the noun *teste* 'test' is going to be incorrectly identified as a form of the verb 'test' and replaced with the equivalent nominal construction with support verb 'do' or 'make' which is correct, but not in that context;
- the sequence of auxiliary verb and main verb was not identified as an expression and consequently the auxiliary verb was paraphrased, such as in [*passou*] *despercebido* 'it went unnoticed' where the auxiliary verb *passar* 'go' is incorrectly identified as the predicate and replaced by the equivalent nominal construction with the support verb *fazer*;
- the subject of the predicate adjective does not correspond to a human noun, as required in the grammar describing such constructions, like in the sentence *A sugestão fora trocista* 'The suggestion was mocking' which results from the fact that currently no checking on the head of the noun phrase is being done to verify if it is of a human type. intransitive adjectives is not currently checking whether the head of the noun phrase is a human noun.

Even when the paraphrases are semantically valid, the following errors may still occur:

- the predicate noun should be in the plural form in the paraphrase *não está acostumado a viajar de avião/não está acostumado a fazer viagem de avião* ‘(he) is not used to travel by plane’/ ‘(he) is not used to make the trip by plane’ should be *não está acostumado a fazer viagens de avião* ‘(he) is not used to make trips by plane’;
- the determiner and the preposition are missing when rephrasing the verb with the nominal construction with the support verb *fazer* ‘do’ or ‘make’, such as in *Um colega declarou/Um colega fizera declaração* ‘a colleague declared’ = *‘a colleague made declaration’ instead of *Um colega fizera a declaração a/de*.

Most errors found are due to underdeveloped grammars (i.e., in a early stage) that require more sophisticated iterations to be reliably applicable to texts. Many of those errors can easily be handled locally on the paraphrasing grammar, adding properties such as [+Hum], or adding the determiner and the preposition. Others depend on disambiguation or on the POS-tagged corpus might have to be tackled at a later stage.

The next steps to be taken to improve our resources will be to fine tune our grammars and dictionaries according to the results that we obtained after applying the resources to the corpus for the first time, and evaluate them on the same sentences again. It is also important to increase our corpus with 461 sentences more, splitting them among the 5 authors of this paper (on a total of 70% of the 801 sentences). The process should be repeated, the resources improved and evaluated, and then tested on the remaining 240 sentences.

Finally, those 240 sentences will all be evaluated by each author to assess inter-annotator agreement. We can than opt for randomly choosing two of them to calculate the agreement statistic proposed by Cohn et al. (2008) or to calculate the agreement among more than two annotators as discussed by Artstein & Poesio (2008).

8 Conclusions and future work

Port4NooJ includes resources that incorporate a wide range of linguistic knowledge, e.g., lexicon-grammar tables created to interact with expert-crafted large-coverage dictionaries and local grammars, and other linguistic resources, such as dictionaries of multiword units, a matter which warrants more attention in future papers. In comparison to other available public resources, Port4NooJ offers two strong competitive advantages, its bilingual nature and the integration of syntactic-semantic knowledge associated to each entry, characteristics that make possible its use in complex natural language processing tasks, such as in the generation of paraphrases and translations (or multilingual paraphrases).

This paper described the new components of the Port4NooJ module. These components are fairly autonomous, but further integration is required to fully take advantage of the several combined modules. We also included the process and results of integrating lexicon-grammar tables that formalize the properties of predicate noun constructions with the support verbs *fazer* ‘do’ or ‘make’, *ser de*

'be', and the human intransitive adjectives. This ongoing process is only complete after consolidation of linguistic information in Port4NooJ dictionaries and grammars.

We will continue the integration work by creating all the necessary grammars to process the constructions formalized in the lexicon-grammar tables at stake. After completion of this integration task, we intend to revise and evaluate the new resources in distinct applications. In the near future, we plan to continue integrating and adapting additional lexicon-grammars, like the constructions with *ser* 'be' formalized by Ranchhod (1990), as these are also a rich resource for paraphrase generation. After including the lexicon-grammar tables available for Portuguese from Portugal, we will include existing ones for Portuguese from Brazil. We also plan to expand our tests to terminology, initially to health terms that co-occur with support verb constructions, such as *fazer uma endoscopia* 'do an endoscopy', we wish to generate paraphrases of technical versus non-expert expressions.

Given the fruitful implementation results obtained so far, further peer collaboration aims at creating synergies between individual researches, and enhancing their potential to address more complex challenges in paraphrasing.

The development of Port4NooJ represents an attempt to address scarceness of resources with the level of linguistic detail that is suitable for applications, such as paraphrasing (as stressed in this paper), and translation. In spite of our integration efforts and commitment in achieving an improved collective quality set of resources, there are still challenges to face and problems to address, such as revision and enhancement of the resources, correctly identifying Portuguese varieties, generating and identifying paraphrases of technical versus non-expert expressions. At last, a thorough evaluation of the resources when used in context is a much required task.

With the new resources, we envisage to improve our paraphrasing capabilities by integrating additional lexicon-grammar tables that allow the generation of language strings. With the full integration of publicly available resources, Port4NooJ will become a more complete, complex and dynamic library of linguistic resources for the description of Portuguese. It will allow to enlarge and improve the quality of the resources behind natural language processing applications such as eSPERTo, but be useful in several tasks and applications that require paraphrase identification and generation.

Acknowledgements This research work was supported by Fundação para a Ciência e Tecnologia (FCT), under projects EXPL/MHC-LIN/2260/2013, UIDB/50021/2020,, and UTAP EXPL/EEL-ESS/0031/2014, and post-doctoral grant SFRH/BPD/91446/2012. The authors would like to thank Max Silberztein for his continued support with NooJ since the development of the first version of Port4NooJ.

References

- Artstein, R., & Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>.
- Baptista, J. (1997). Sermão, tarefa e facada: uma classificação das expressões conversas dar-levar. *Seminários de Linguística*, 1, 5–37.

- Baptista, J. (2000). *Sintaxe dos predicados nominais construídos com o verbo-suporte ser de*. Tese de doutoramento, Universidade do Algarve.
- Baptista, J. (2004). Instrument nouns and fusion. Predicative nouns designating violent actions, *Linguisticae Investigationes Supplementa vol Lexique, Syntaxe et Lexique-Grammaire (Syntax, Lexis and Lexicon-Grammar)*. Hommage à Maurice Gross (pp. 31–40). John Benjamins Publishing Co.
- Baptista, J. (2005a). Construções simétricas: argumentos e complementos. In O. Figueiredo, G. Rio-Torto, & F. Silva (Eds.), *Estudos de homenagem a Mário Vilela* (pp. 353–367). London: Faculdade de Letras da Universidade do Porto.
- Baptista, J. (2005b). *Sintaxe dos predicados nominais com 'ser de'*. Lisbon: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia.
- Baptista, J., Fernandes, G., Talhadas, R., Dias, F., & Mamede, N. (2015). Implementing European Portuguese Verbal Idioms in a Natural Language Processing System. In *Proceedings of conference of the European Society of Phraseology (EuroPhras 2015)*, Málaga, Spain (pp. 102–115).
- Baptista, J., Mamedem, N., & Markov, I. (2014). Integrating a Lexicon-Grammar of Verbal Idioms in a Portuguese NLP System, PARSEME General Meeting, Athens, 10–11 March 2014 (poster session).
- Barreiro, A. (2009). *Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation*. PhD thesis, Universidade do Porto.
- Barreiro, A. (2011). Spider: A system for paraphrasing in document editing and revision—applicability in machine translation pre-editing. In A. Gelbukh (Ed.), *Proceedings of 12th international conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, Tokyo, Japan, 20–26 February 2011 (pp. 365–376), Part II. Springer.
- Barreiro, A., Batista, F., Ribeiro, R., Moniz, H., & Trancoso, I. (2014). OpenLogos semantico-syntactic knowledge-rich bilingual dictionaries. In NCC Chair, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA).
- Barreiro, A., & Mota, C. (2017). e-PACT: eSPERTo paraphrase aligned corpus of EN-EP/BP translations. *Tradução em Revista*, 1(22), 87–102.
- Barreiro, A., & Mota, C. (2018). Paraphrastic variance between European and Brazilian Portuguese. In M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, S. Malmasi, & A. Ali (Eds.), *Proceedings of the fifth workshop on NLP for similar languages, varieties and dialects (VarDial) (COLING 2018)*. Association for Computational Linguistics.
- Barreiro, A., Rebelo-Arnold, I., Mota, C., Garcez, I., & Baptista, J. (2018, forthcoming). Automatic paraphrasing and normalization of Portuguese informal into formal language. In A. Barreiro, J. Baptista, P. Quaresma & R. Vieira (Eds.), *Proceedings of the first workshop on linguistic tools and resources for paraphrasing in Portuguese (POP@PROPOR 2018)*. Springer.
- Carvalho, P. (2007). *Análise e Representação de Construções Adjectivais para Processamento Automático de Texto. Adjectivos Intransitivos Humanos*. PhD thesis, Universidade de Lisboa.
- Casteleiro, J. M. (1981). *Sintaxe transformacional do adjetivo*. INIC.
- Chacoto, L. (2005). *O Verbo Fazer em Construções Nominais Predicativas*. PhD thesis, Universidade do Algarve.
- Cohn, T., Callison-Burch, C., & Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4), 597–614. <https://doi.org/10.1162/coli.08-003-R1-07-044>.
- D'Agostino, E., & Elia, A. (1998). Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. In AA VV, *Ai limiti del linguaggio* (pp. 287–310). Laterza.
- Frankenberg-García, A., & Santos, D. (2003). Introducing COMPARA: The Portuguese-English parallel corpus. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 71–87). St. Jerome.
- Gamallo, P., & Pereira-Fariña, M. (2019). Explorando métodos non-supervisados para calcular a similitude semântica textual. *Linguamática*, 10(2), 63–68. <https://doi.org/10.21814/lm.10.2.275>.
- Gross, G. (1989). *Les construction converses du français*. Droz.
- Gross, M. (1975). *Méthodes en syntaxe: régime des constructions complétives*. Actualités scientifiques et industrielles. Hermann.
- Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, 15(63), 7–52.

- Gross, M. (1982). Une classification des phrases «figées» du français. *Revue québécoise de linguistique*, 11(2), 151–185.
- Grycner, A., & Weikum, G. (2016). POLY: Mining relational paraphrases from multilingual sentences. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2183–2192). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1236>. <https://www.aclweb.org/anthology/D16-1236>.
- Guillet, A., & Leclère, C. (1981). Restructuration du groupe nominal. *Langages*, 1(63), 99–125.
- Harris, Z. S. (1952). Discourse analysis. *Language*, 1(28), 1–30.
- Harris, Z. S. (1964). Papers on Syntax, D. Reidel Publishing Company, The elementary transformations, (pp. 211–235).
- Harris, Z. S. (1968). *Mathematical structures of language*. Interscience tracts in pure and applied mathematics, Interscience Publishers.
- Harris, Z. S. (1976). *Notes du Cours de Syntaxe*. Seuil.
- Harris, Z. S. (1981). *The elementary transformations* (pp. 211–235). Springer.
- Harris, Z. S. (1991). *A theory of language and information: A mathematical approach*. Clarendon Press.
- Harris, Z. S. (1965). Transformational theory. *Language*, 41(3), 363–401.
- Janssen, M., Kuhn, T. Z., Ferreira, J. P., & Correia, M. (2018). The CPLP corpus: A pluricentric corpus for the common portuguese spelling dictionary (VOC). In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings of the XVIII EURALEX international congress: Lexicography in global contexts* (pp. 835–840). Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia.
- Laporte, E., & Voyatzi, S. (2008). An electronic dictionary of French multiword adverbs. In *Language resources and evaluation conference. Workshop towards a shared task for multiword expressions* (pp. 31–34).
- Leclère, C. (1995). Sur une restructuration dative. *Language Research*, 1(31), 179–198.
- Machonis, P. (2010). English phrasal verbs: from lexicon-grammar to natural language processing. *Southern Journal of Linguistics*, 34(1), 21–48.
- Mamede, N., Baptista, J., Diniz, C., & Cabarrão, V. (2012). STRING: A hybrid statistical and rule-based natural language processing chain for Portuguese. In *International conference on computational processing of Portuguese (PROPOR 2012)*, Coimbra, Portugal, vol Demo Session
- Mayhew, S., Bicknell, K., Brust, C., McDowell, B., Monroe, W., & Settles, B. (2020). Simultaneous translation and paraphrase for language education. In *Proceedings of the fourth workshop on neural generation and translation* (pp. 232–243). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.ngt-1.28>. <https://www.aclweb.org/anthology/2020.ngt-1.28>.
- Mota, C., Baptista, J., & Barreiro, A. (2019). The lexicon-grammar of predicate nouns with ser de in Port4NooJ. In I. M. Mirto, M. Monteleone, & M. Silberstein (Eds.), *Formalizing natural languages with NooJ 2018 and its natural language processing applications* (pp. 124–137). Springer. https://doi.org/10.1007/978-3-030-10868-7_12.
- Mota, C., Barreiro, A., Raposo, F., Ribeiro, R., Curto, S., & Coheur, L. (2016). eSPERTo's paraphrastic knowledge applied to question-answering and summarization. In L. Barone, M. Monteleone, & M. Silberstein (Eds.), *Automatic processing of natural-language electronic texts with NooJ: 10th International Conference (NooJ 2016)*, České Budějovice, Czech Republic, 9–11 June 2016 (pp. 208–220). Revised Selected Papers. Springer.
- Mota, C., Carvalho, P., Raposo, F., & Barreiro, A. (2015). Generating paraphrases of human intransitive adjective constructions with Port4NooJ. In T. Okrut, Y. Hetsevich, M. Silberstein, & H. Stanislavenska (Eds.), *Automatic processing of natural language electronic texts with NooJ—Selected papers of the 9th international conference* (pp. 107–122). Communications in Computer and Information Science. Springer.
- Mota, C., Chacoto, L., & Barreiro, A. (2018). Integrating the lexicon-grammar of predicate nouns with support verb fazer into Port4NooJ. In S. Mbarki, M. Mourchid & M. Silberstein (Eds.), *Formalizing natural languages with NooJ and its natural language processing applications* (pp. 29–39). Springer.
- Paşca, M., & Dienes, P. (2005). Aligning needles in a haystack: Paraphrase acquisition across the web. In *Second international joint conference on natural language processing: Full papers*. https://doi.org/10.1007/11562214_11. <https://www.aclweb.org/anthology/I05-1011>.
- Pershina, M., He, Y., & Grishman, R. (2015). Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the first workshop on linking computational models of lexical, sentential and*

- discourse-level semantics* (pp. 76–82). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W15-2709>. <https://www.aclweb.org/anthology/W15-2709>.
- Ranchhod, E. (1983). On the support verbs *ser* and *estar* in portuguese. *Linguisticae Investigationes*, 7(2), 317–353.
- Ranchhod, E. (1990). *Sintaxe dos predicados nominais com estar*. Linguística, INIC.
- Rassi, A., Mamede, N., Baptista, J., & OV, I. (2015). Integrating support verb constructions into a parser. In: Proceedings of the Symposium in Information and Human Language Technology (STIL'2015), pp. 57–62
- Rassi, A., Santos-Turati, C., Baptista, J., Mamede, N., & Vale, O. (2014). 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* (pp. 92–101). Springer.
- Rassi, A. P. (2015). *Descrição, classificação e processamento automático das construções com o verbo dar em português brasileiro*. PhD thesis, Universidade Federal de São Carlos, São Carlos-SP.
- Rassi, A. P., Barros, C. D., & Santos-Turati, M. C. A. (2012). *Correlações sintático-semânticas entre as construções com os verbos-suporte 'dar', 'ter' e 'fazer'* (pp. 193–206). Dialogar é preciso: Linguística para o processamento de línguas.
- Rassi, A. P., Barros, C. D., & Santos-Turati, M. C. A. (2013). Tipologia sintática das construções com os verbos-suporte *dar*, *fazer* e *ter*. In *Proceedings of III workshop on Portuguese description* (pp. 36–43), Fortaleza, Ceará.
- Rebelo-Arnold, I., Barreiro, A., & Quaresma, P. (2018). EP–BP paraphrastic alignments of verbal constructions involving the clitic pronoun *lhe*. In A. Barreiro, J. Baptista, P. Quaresma, & R. Vieira (Eds.), *Proceedings of the first workshop on linguistic tools and resources for paraphrasing in Portuguese (POP) (PROPOR 2018)*. Springer.
- Salkoff, M. (1990). Automatic translation of support verb constructions. In *Proceedings of the 13th conference on computational linguistics (COLING '90)* (Vol. 3, , pp. 243–246). ACL.
- Salkoff, M. (1999). *A French-English grammar: A contrastive grammar on translational principles*. Linguisticae investigationes. John Benjamins.
- Santos, D. (2014). Como estudar variantes do português e, ao mesmo tempo, construir um português internacional? *Presentation at Contact, Variation and Change: Corpora development and analysis of Iberoromance language varieties workhop*. <http://www.linguateca.pt/Diana/download/VariantesPIGSCP.pdf>.
- Santos, C. (2015a). *Construções com verbo-suporte ter no português do brasilasil*. PhD thesis, Universidade Federal de São Carlos, São Carlos-SP.
- Santos, D. (2015b). Portuguese language identity in the world: adventures and misadventures of an international language. In E. Khachaturyan (Ed.), *Language–Nation–Identity: The questione della lingua in an Italian and non-Italian context* (pp. 31–54). Cambridge Scholars Publishing.
- Santos, D., & Inácio, S. (2006). Annotating COMPARA, a grammar-aware parallel corpus. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odjik, & D. Tapias (Eds.), *Proceedings of the 5th international conference on language resources and evaluation (LREC 2006)* (pp. 1216–1221).
- Scott, B. (2003). The logos model: An historical perspective. *Machine Translation*, 18(1), 1–72.
- Scott, B. (2018). *Translation, brains and the computer: A neurolinguistic solution to ambiguity and complexity in machine translation. machine translation: technologies and applications*. Springer.
- Shinyama, Y., & Sekine, S. (2003). Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on paraphrasing* (pp. 65–71). Association for Computational Linguistics. <https://doi.org/10.3115/1118984.1118993>. <https://www.aclweb.org/anthology/W03-1609>.
- Silberztein, M. (1993). Les groupes nominaux productifs et les noms composés lexicalisés. *Linguisticae Investigationes*, 17(2), 405–425.
- Silberztein, M. (2015). *La formalisation des langues: l'approche de NooJ*. ISTE.
- Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. Wiley.
- Souza, M., & Sanches, L. M. P. (2019). Detecção de paráfrases na lingua portuguesa usando sentence embeddings. *Linguamática*, 10(2), 31–44. <https://doi.org/10.21814/lm.10.2.286>.
- Vietri, S. (2004). *Lessico-grammatica dell'italiano: metodi, descrizioni, applicazioni*. PhD thesis, UTET.
- Vietri, S. (2010). The formalization of Italian lexicon-grammar tables in a nooj pair dictionary/grammar. In J. Kutí, M. Silberztein, & T. Váradi (Eds.), *Applications of finite-state language processing:*

Selected papers from the NooJ 2008 International conference (pp. 138–147). Cambridge Scholars Publishing.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.