# Generating Paraphrases of Human Intransitive Adjective Constructions with Port4NooJ

Cristina Mota[1(✉)], Paula Carvalho[1,2], Francisco Raposo[1],
and Anabela Barreiro[1]

[1] INESC-ID Lisboa, Lisbon, Portugal
{cmota,francisco.afonso.raposo}@ist.utl.pt, {abarreiro,pcc}@inesc-id.pt
[2] Universidade Europeia – Laureate International Universites, Lisbon, Portugal

**Abstract.** This paper details the integration into Port4NooJ of 15 lexicon-grammar tables describing the distributional properties of 4,248 human intransitive adjectives. The properties described in these tables enable the recognition and generation of adjectival constructions where the adjective has a predicative function. These properties also establish semantic relationships between adjective, noun and verb predicates, allowing new paraphrasing capabilities that were described in NooJ grammars. The new dictionary of human intransitive adjectives created by merging the information on those tables with the Port4NooJ homograph adjectives is comprised of 5,177 entries. The enhanced Port4NooJ is being used in eSPERTo, a NooJ-based paraphrase generation platform.

## 1 Introduction

Port4NooJ (Barreiro 2010) is a set of resources that allows the generation of paraphrases for Portuguese, feeding the linguistic engine of the eSPERTo paraphrasing system[1] based on NooJ technology (Silberztein 2015). The term paraphrase is commonly used to refer the relation between two or more constructions that are morpho-syntactically and/or semantically related (e.g. *to make a presentation (of ) = to present*). In most cases, this relation is established between constructions corresponding to the same syntactic unit. However, the transformations described in lexicon-grammar tables also allow establishing relations between different syntactic units (e.g. *These teachers are Portuguese = The Portuguese teachers [...]*). Hence, we extend the term paraphrase to the semantico-syntactic relation between two or more sentences and/or their constituents.

This paper presents an enhanced Port4NooJ that includes 15 lexicon-grammar tables describing the distributional properties of 4,248 human intransitive adjectives formalized by Carvalho (2007). Among other properties, these linguistic resources provide information on: (i) the syntactic and semantic nature of the subject modified by each adjective, which can correspond to a human noun, a complex noun phrase involving an appropriate noun, or to a finite or non-finite clause; (ii) the copulative verbs (and aspectual variants) selected by

---

[1] https://esperto.l2f.inesc-id.pt/.

each adjective; (iii) the constraints related to the quantification of adjectives by an adverb or a degree morpheme; (iv) the position of adjectives in adnominal context (pre- or post-nominal position); (v) the possibility of certain adjectives being optionally followed by an infinitive clause, with causal interpretation, or by a human noun phrase introduced by the preposition *para* (*for*). In addition to general properties, these resources also describe particular constructions in which human intransitive adjectives may occur, as in: (vi) generic and cross-constructions, where the adjective fills the head of a noun phrase; (vii) charac-terizing indefinite constructions, where the adjective occurs after an indefinite article; (viii) exclamative sentences expressing insult. Moreover, these lexicon-grammar tables specify the morpho-syntactically related predicative nouns and verbs, whenever they exist, as well as the appropriate nouns that can appear in specific adjective constructions.

The properties described in the lexicon-grammar tables add new paraphras-ing capabilities to eSPERTo. Port4Nooj initial paraphrases involved transforma-tions of support verb constructions (or their stylistic or aspectual variants) into single verbs. Later on, new paraphrasing capabilities were added to Port4NooJ, namely transformations of phrasal verbs into equivalent expressions, compound adverbs into single adverbs, relatives into participial adjectives, relatives into possessives, relatives into compound nouns, and agentive passives into actives. Section 4 presents examples of these paraphrases, but a more detailed descrip-tion of Port4NooJ first paraphrasing capabilities can be found in Barreiro (2009) and Barreiro (2011).

The use of the linguistic knowledge described in the integrated tables allows mapping of several other types of paraphrasing constructions resulting in a semantic relationship between adjective, noun and verb predicates. The lexicon-grammar tables enable eSPERTo to paraphrase (i) adjective, noun and verb morphologically related constructions (*está zangado* (*is angry*) = *zangou-se* (*got (self) angry*) = *esteve envolvido numa zanga* (*was involved in anger*)); (ii) adjec-tive constructions supported by different copulative verbs (*estar perdido* (*to be lost*) = *andar perdido* (*walk around lost*)); (iii) constructions involving nation-ality and other membership relations (*de origem portuguesa* (*of Portuguese ori-gin/roots*) = *portugueses* (*Portuguese*) = *de Portugal* (*from Portugal*); *ben-fiquista* (*Benfica fan*) = *do Sport Lisboa e Benfica* (*a fan of Sport Lisboa e Benfica*)); (iv) cross-constructions (*o idiota do rapaz* (*the idiot of the boy*) = *o rapaz é um idiota* (*the boy is an idiot*)); appropriate noun constructions (*foi moderado nos seus comentários* (*he was moderated in his comments*) = *os seus comentários foram moderados* (*his comments were moderated*) = *foi moderado* (*he was moderated*), (v) generic noun phrases (*é um indivíduo estúpido* (*he is a fool*) = *é um estúpido* (*he is a fool*) = *é estúpido* (*he is fool*)), among others.

## 2   Related Work

Our work follows previous attempts to integrate lexicon-grammar tables in nat-ural language processing systems.

Machonis ([2010](#)) presents experiments in NooJ with lexicon-grammar tables of transitive and neutral phrasal verbs in English to identify these types of constructions in texts with and without insertion. The author corroborates that, on the one hand, NooJ is a very powerful tool for parsing compound expressions that involve insertion, such as English phrasal verbs, and, on the other hand, a lexicon-grammar of phrasal verbs can help distinguish prepositional usage from true particles in natural language processing.

Vietri ([2010](#)) describes the integration, also into NooJ, of one of her 13 lexicon-grammar tables, table EpinC, which contains 900 Italian idioms with the structure $N_0$ *essere in C*, where *essere* (*to be*) is a support verb, and *in C* is a frozen or semi-frozen prepositional phrase starting with the preposition *in*. The author shows that this integration allowed to refine the linguistic analysis of this type of sequences.

Baptista et al. ([2014](#)) discuss lexical and parsing issues of integrating a lexicon-grammar of Portuguese verbal idioms into STRING, a hybrid statistical and rule-based pipeline for natural language processing of Portuguese. More than 2,000 rules were created semi-automatically for ten formal classes of verbal idioms. The system precision was estimated after processing a large Portuguese corpus of news texts.

## 3   The eSPERTo Project

This research work was developed in the scope of the eSPERTo[2] project. The main objective of this project is the development of a context-sensitive and linguistically enhanced paraphrase generator that recognizes semantico-syntactic, multiwords and other phrasal units, and transforms them into semantically equivalent phrases, expressions, or sentences. This semantically-driven paraphrasing system uses a new hybrid technique that combines statistics and local grammars to acquire linguistic knowledge applied in the identification and generation of new and increasingly more complex paraphrases. Currently, eSPERTo is integrated in an interactive application that helps Portuguese language learners in producing and revising their texts. The utility of eSPERTo's paraphrasing capabilities are now being explored in two other application scenarios: (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. Figure 1 shows eSPERTo's current interactive Web interface designed to help Portuguese language learners in producing and revising their texts. Among other functionalities, 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. For the sentence illustrated in Fig. 1: *O homem americano apresentou o*

---

[2] In Portuguese, "esperto" means "smart", but here it is also an acronym for "System of Paraphrasing for Editing and Revision of Text" (*"Sistema de Parafraseamento para Edição e Revisão de Texto"*). eSPERTo is a "smart system" in the sense that it contains semantic "understanding" in its linguistic knowledge base.
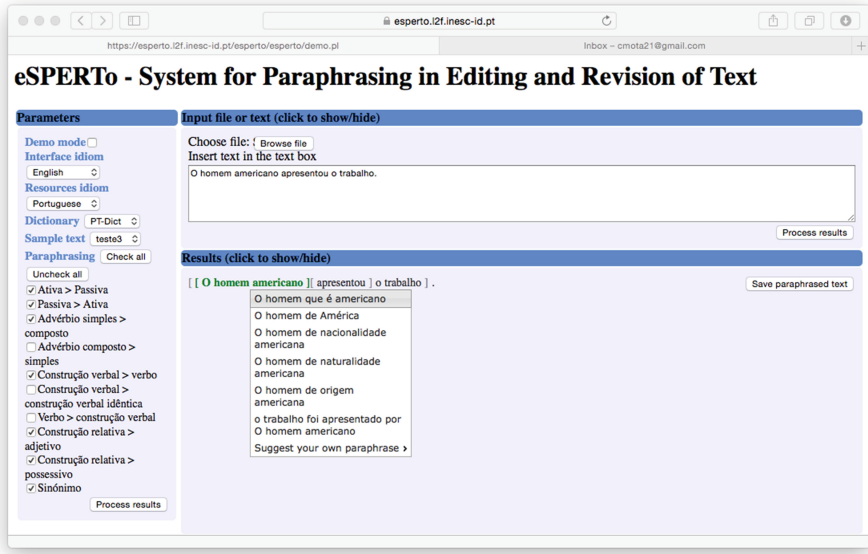
**Fig. 1.** Online use of eSPERTo in text editing and revision

*trabalho* (*The American man presented the work*), eSPERTo suggests its equivalent passive paraphrase: *O trabalho foi apresentado pelo homem americano* (*The work was presented by the American man*). For the noun phrase *o homem americano* (*the American man*), eSPERTo suggests paraphrases such as: *o homem que é americano* (*the man who is American*), *o homem de nacionalidade americana* (*the man with American nationality*), *o homem de naturalidade americana* (*the man with American origin*), *o homem de origem americana* (*the man with American origin*). The user can then select any of the paraphrases listed, or provide his/her own paraphrase.[3]

## 4    Port4NooJ and Its First Paraphrases

Port4NooJ is the Portuguese linguistic module of NooJ. The module can be downloaded from the NooJ website[4] or from the Linguateca's resources repository[5]. The initial Port4NooJ resources derive from OpenLogos. OpenLogos is an open source derivative of the commercial Logos system downloadable from the DFKI website[6] and available for testing at INESC-ID[7]. The Logos system

---

[3] For more information about the eSPERTo paraphrasing system and interface, see Mota et al. (2015 forthcoming).

[4] http://www.nooj-association.org.

[5] http://www.linguateca.pt/Repositorio/Port4Nooj/.

[6] http://logos-os.dfki.de/.

[7] http://www.l2f.inesc-id.pt/openlogos/demo.html.

was built on the Logos Model (Scott 2003), (Barreiro et al. 2011). In order to create Port4NooJ, the OpenLogos English-Portuguese bilingual dictionary was converted into NooJ format and its language pair order was inverted. Besides the large coverage electronic dictionary with English transfers, the Port4NooJ module contains two other important components: (i) the rules which formalize and document Portuguese inflectional and derivational descriptions, and (ii) different types of grammar, namely morphological[8], disambiguation, semantico-syntactic, multiword expressions, and translation and paraphrasing grammars. The different components of Port4NooJ interact among them and are used to process texts. Several processing functions can be performed with these resources, among others, part of speech annotation, pattern recognition, semantic unit analysis, concordances, information extraction, paraphrasing and translation.[9] Barreiro (2008) and Barreiro (2010) describe in detail the initial dictionary and its enhancement with new linguistic knowledge, namely inflectional, derivational and morpho-syntactic properties, and semantic relations that permitted the generation of paraphrases.

Initially, Port4NooJ contemplated paraphrases involving support verb constructions or their stylistic or aspectual variants and corresponding single verbs (*fazer/realizar/efetuar uma apresentação* (*make a presentation (of)*) = *apresentar* (*present*)), compound and single adverbs (*de uma forma interativa* (*in an interactive way*) = *interativamente* (*interactively*); *com entusiasmo* (*with enthusiasm*) = *entusiasticamente* (*enthusiastically*)), relatives and participial adjectives (*que foram escritos* (*that were written*) = *escritos* (*written*)), relatives and possessive constructions *o papel que a Europa tem/desempenha* (*the role that Europe plays*) = *o papel da Europa* (*the role of Europe*)), and active/passive constructions (*A solta B* (*A releases B*) = *B é solto por A* (*B is released by A*)), among others. In Sect. 5, we will describe the new Port4NooJ paraphrases resulting from the transformation of human intransitive adjective constructions described in the lexicon-grammar tables.

## 5   Lexicon-Grammar of Human Intransitive Adjectives

The lexicon-grammar tables explored in this study describe 4,248 human intransitive adjectives, i.e. adjectives that select a human noun as subject and do not require any complement. These adjectives were grouped into 15 subclasses, which present different lexico-syntactic properties, as illustrated in Table 1.

These properties relate specifically to: (i) the syntactic nature of the human subject: depending on the adjective class, this position can be headed by a human noun (Nhum), (ii) a complex noun phrase involving an appropriate noun (Nap de Nhum), and/or (iii) by a finite clause (QueF), (iv) the nature of the copular verb (Cop) selected by the adjective: there are adjectives that co-occur

---

[8] The morphological component includes a morphological grammar to process contracted forms.

[9] Due to its bilingual characteristics, Port4NooJ dictionary can be used in translation (with embedded paraphrasing).

**Table 1.** Classes and distributional properties of human intransitive adjectives

| | N0 | | QueF | Cop | | CCI | |
|---|---|---|---|---|---|---|---|
| | Nhum | Nap de Nhum | | ser | estar | um Adj | |
| SAHP1 | + | + | + | + | - | - | Ele é honesto – He is honest<br>A atitude dele é honesta – His attitude is honest<br>Que ele faça isso é honesto da sua parte<br>– Doing that is honest on his part |
| SAHP2 | + | + | - | + | - | - | Ele é atlético – He is athletic<br>O seu porte é atlético – His look is athletic |
| SAHP3 | + | - | - | + | - | - | Ele é comunicativo – He is communicative |
| SAHC1 | + | + | + | + | - | + | Ele é arrogante  – He is arrogant<br>Ele é um arrogante – He is an arrogant<br>O comportamento dele é arrogante – His behavior is arrogant<br>Que ele se comporte assim é arrogante da sua parte<br>– Behaving like that is arrogant on his part |
| SAHC2 | + | + | - | + | - | + | Ele é introvertido – He is introvert<br>Ele é um introvertido – He is an introvert<br>A sua maneira de ser é introvertida – His way of being is introvert |
| SAHC3 | + | - | - | + | - | + | Ele é aselha – He is clumsy<br>Ele é um aselha – He is a clumsy |
| EAHP2 | + | + | - | - | + | - | Ele está bronzeado – He is tanned<br>O rosto dele está bronzeado – His face is tanned |
| EAHP3 | + | - | - | - | + | - | Ele está zangado – He is angry |
| SEAHP2 | + | + | - | + | + | - | Ele (é + está) cadavérico – He is cadaverous<br>O seu rosto (é + está) cadavérico – His face is cadaverous |
| SEAHP3 | + | - | - | + | + | - | Ele (é + está) calvo – He is bald |
| SEAHC2 | + | + | - | + | + | + | Ele (é + está) esquelético – He is cadaverous<br>As pernas dele (são + estão) esqueléticas – His legs are cadaverous<br>Ele é um esquelético – He is a cadaverous |
| SEAHC3 | + | - | - | + | + | + | Ele (é + está) alucinado – He is alucinated<br>Ele é alucinado – He is alucinated<br>Ele é um alucinado – He is an alucinated |
| SAN | + | + | - | + | - | - | Ele é português – He is Portuguese<br>A sua nacionalidade é portuguesa – His nationality is Portuguese |
| SAF | + | - | - | + | - | +/- | Ele é socialista – He is socialist<br>Ele é um fascista – He is a fascist |
| SEAD | + | +/- | - | +/- | +/- | +/- | Ele é autista – He is autistic<br>Ele está constipado – He has a cold<br>Ele (é + está) anorético – He is anoretic<br>As suas pernas estão inchadas – His legs are swollen<br>Ele é cleptomaníaco – He is Kleptomaniac<br>Ele é um cleptomaníaco – He is a kleptomaniac |

only with the verb *ser* or *estar*, while others co-occur with both copula verbs, (v) the possibility of the predicative adjective appearing in another indefinite construction (CCI).

In addition to these generic properties, the tables describe specific distributional and transformational attributes of the adjective, allowing recognition and generation of a variety of syntactic constructions where each adjective occurs. Some of these properties depend on the adjective subclass, while others depend on the adjective itself. Figure 2 presents some examples extracted from the lexicon-grammar table that describes the adjectives of nationality. As illustrated in the table, the adjectives of this semantico-syntactic class require a human noun (Nhum) to fill the subject position (N0). In predicative context, they are linked to its subject by the verb *ser*, and they cannot be quantified or intensified (Quant). Moreover, in adnominal context, the adjectives of nationality can only occur in post-adnominal position. However, there are some properties in the table that may vary depending on the adjective. For example, even though all the adjectives of this class are able to modify the generic predicative classifier *origem* (designation of origin), the presence of a more specific noun classifier, such as *nacionalidade* (nationality/country of origin), *naturalidade* (place of origin), *etnia* (ethnicity) or *raça* (race), depends on the semantics of the adjective (i.e., on the nature of the locative noun with which each adjective is associated). The information described in Table 2 allows generating the following constructions, among others:

```
Ele é açoriano – He is Azorean
Ele é um açoriano [simpático] – He is a [friendly] Azorean
Ele é um rapaz açoriano – He is an Azorean boy
Ele é de origem açoriana – He is of Azorean origin
Ele é de naturalidade açoriana – He is of Azorean origin
Ele é dos Açores – He is from Azores
```

| Nhum | Nap de Nhum | Npabst | Nap | Npc | QueF | ser | estar | andar | ficar | permanecer | viver | encontrar-se | sentir-se | mostrar-se | revelar-se | tornar-se | UM Nclas | Um +Modif | Adv Quant | Superlativo | ser de | origem | nacionalidade | naturalidade | etnia | raça | Adjectivo | N Adj | Adj N | Nhum Vop A Prep Nap | (de + em) (E+Def) Nap | em Dposs Nap | em Vinf | para com Nhum | CCI | (Seu + Meu) Adj! | Nome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | + | + | + | - | - | + | + | + | + | - | - | abissínio | + | - | - | - | - | - | - | - | - | Abissínia |
| + | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | + | + | + | - | - | + | + | + | + | - | - | acadiano | + | - | - | - | - | - | - | - | - | Acádia |
| + | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | + | + | + | - | - | + | + | + | + | - | - | açoriano | + | - | - | - | - | - | - | - | - | Açores |
| + | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | + | + | + | - | - | + | + | + | + | - | - | afegão | + | - | - | - | - | - | - | - | - | Afeganistão |
| + | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | + | + | + | - | - | + | + | - | - | + | + | africano | + | - | - | - | - | - | - | - | - | África |

**Fig. 2.** Lexicon-grammar table SAN describing adjectives of nationality

Figure 3 presents an excerpt of the table describing the predicative adjectives classified as SAHC1. The adjectives of this class select the verb *ser*, and some of their aspectual or stylistic variants, in particular *mostrar-se* (show self as),

| | N0 | | | | | Vcop | | | | | | | | | | | Class | | Quant | | | P_Adn | | Reestr N0 | | | (W+E) | | | | |
| Nhum | Nap de Nhum | Npalsd | Npc | QueF | N-hum | ser | estar | andar | ficar | permanecer | viver | encontrar-se | sentir-se | mostrar-se | revelar-se | tornar-se | UM Nclas | Um +Modif | Adv Quant | Superlativo | Adjectivo | N Adj | Adj N | Nhum Vop A Prep Nap | (de + em) (E+Def) Nap | em Dposs Nap | em Vinf | para com Nhum | CCI | (Seu + Meu) Adj! | Nome |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | adulador | + | - | + | - | + | + | + | + | + | adulação |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | agourento | + | - | + | - | + | + | - | + | + | mau agouro |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | alarmista | + | - | + | - | + | + | - | + | + | alarmismo |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | alentador | + | - | + | - | + | + | - | + | - | alento |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | altruísta | + | - | + | - | + | + | - | + | - | altruísmo |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | + | ambicioso | + | - | + | - | + | + | - | + | + | ambição |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | + | anormal | + | - | + | - | + | + | + | + | + | anormalidade |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | antidesportista | + | - | + | - | + | + | - | + | + | antidesportivismo |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | antipático | + | + | + | - | + | + | - | + | + | antipatia |
| + | + | + | - | + | - | + | - | - | - | - | - | - | - | + | + | + | + | + | + | - | antipatriota | + | - | + | - | + | - | - | + | + | antipatriotismo |

**Fig. 3.** Lexicon-grammar table SAHC1 describing a subclass of predicative adjectives

*revelar-se* (reveal self as), and/or *tornar-se* (become). Their subject position can be filled by a human noun (Nhum), an appropriate noun (Nap), or a finite clause. Most adjectives described in this table can be quantified by an adverb (Adv), and some can also receive a degree morpheme (Sup). In adnominal context, there are adjectives, like *antipático* (unfriendly), that can occur both in post- and pre-adnominal position. Among other properties, this table provides information on particular syntactic constructions derived from a set of transformations involving the subject (Reest N0). Below, we present some examples represented in this lexicon-grammar.

```
O rapaz antipático - The unfriendly boy
Um antipático rapaz - An unfriendly boy
Ele (é + mostrou-se + revelou-se) antipático
    - He (is + showed himself as + revealed himself as) unfriendly
Ele é um antipático - He is such an unfriendly person
Ele é uma pessoa antipática - He is an unfriendly person
A sua maneira de ser é antipática - His personality is unfriendly
Que ele faça isso é antipático da sua parte - What he is doing is unfriendly
Ele revela antipatia - He reveals unfriendliness
Ele é antipático na sua maneira de ser - He is unfriendly in his ways
Ele é antipático para com toda a gente - He is unfriendly with everybody
Seu antipático! - Unfriendly you!
```

# 6    Integration of the Lexicon-Grammar Tables in Port4NooJ

## 6.1    From Lexicon-Grammar Tables to NooJ Dictionaries

The conversion of the 15 lexicon-grammar tables into a NooJ dictionary of human intransitive adjectives was mostly done automatically with different scripts. The process is hence easily adaptable to integrate other lexicon-grammar tables. For each entry in a lexicon-grammar table, we applied the following main steps:

1. If adjective is already in Port4NooJ, merge the lexicon-grammar properties with every homograph adjectival entry in Port4NooJ dictionary, by adding the new properties to that entry, otherwise create a new entry;
2. Create inflectional (FLX) and derivational (DRV) codes and corresponding rules as needed;
3. Check for missing FLX and DRV codes, and create new ones as needed.

### 6.1.1   Representation of Lexicon-Grammar Table Properties

The properties `+IH` and `+Table=<table_name>` were added to all human intransitive adjectives. The first property indicates that the adjective is a human intransitive adjective, and the second one refers to the lexicon-grammar table where the adjective properties are formalized.

For each different column in a lexicon-grammar table, a property `+<name_of_ prop>` was created. If the adjective row is marked with the value `+`, then that property was added to the adjective entry. Properties that have a value other than `+/-` were added as `+<name_of_prop>=value_of_prop`. For properties `Nome` and `Verbo`, instead of creating `+Nome=<value>` and `+Verbo=<value>`, a script translates the pair(s) adjective/noun and adjective/verb, if they exist, into a derivation paradigm and creates attributes `+DRV=A2N<drv_code1>:<flx_code1>` and `+DRV=A2V<drv_code2>:<flx_code2>`, respectively.

The `drv` code is determined and formalized automatically by finding the radical between the adjective and the noun or verb. For example, the adjective *alegre* (*happy*) is associated with the corresponding noun *alegria* (happiness) and the corresponding verb *alegrar* (*become happy*) through derivation rules (cf. `A2B143` and `A2V6` below), which replace the adjectival ending -*e* with the noun and verb endings -*ia* and -*ar*, respectively.

```
alegr(e|ia) => A2N143 = <B1>ia/N
alegr(e|ar) => A2V6 = <B1>ar/V
```

The inflection of the derived word (`flx_code1` or `flx_code2`) is determined by consulting Port4NooJ (cf. `FLX=CASA` for the noun and `FLX=FALAR` for the verb).

```
alegria,N+FLX=CASA+AB+state+EN=joy+SYNN=contentamento
alegrar,V+FLX=FALAR+Aux=1+PRECVagree-type+Subset=...
```

In cases where the derived forms did not exist, their codes were assigned automatically. The FLX code of the base form, the adjective, was determined in the same way: the inflection code is looked up in Port4NooJ or assigned automatically in cases where the word does not exist.

Additional properties were created to account for specific knowledge required in paraphrasing. For example, the property `+TopDET={o|a|os|as|undef}` indicates the determiner that co-occurs with a toponym:

```
o homem abissínio <-> o homem da Abissínia
the Abyssinian man <-> the man from Abyssinia

o homem açoriano <-> o homem dos Açores
the Azorean man <-> the man from the Azores

o homem português <-> o homem de Portugal
the Portuguese man <-> the man from Portugal
```

The value of +TopDET was determined automatically by consulting the AC/DC corpora (28 corpora covering different variants of Portuguese in a total of 1,279 million words) and counting the distribution of the determiner that occurs (or its absence) in the context of the prepositions *de* (*of*) and *em* (*in*). For each toponym in the lexicon-grammar tables, a CQP query, the language used to query the AC/DC corpora, with the following structure was used to consult each corpora:

```
de+em+d(o+a+os+as)+n(o+a+os+as) <toponym>
```

In cases where the toponym did not occur in the corpora, +TopDET=undef was used to distinguish those cases from toponyms that do not accept determiners (i.e., the property +TopDET was not added to the adjective entry).

### 6.1.2   Integration with eSPERTo Dictionary Entries

After the properties of each adjective in the lexicon-grammar tables were created, the script merged those properties with the information corresponding to that adjective in the Port4NooJ entries. When the adjective already existed in Port4NooJ, the lexicon-grammar properties were added to all the adjective homograph entries in Port4NooJ. This means that, for example, the following Port4NooJ entries:

```
velho,A+FLX=ALTO+AB+class+EN=vintage
velho,A+FLX=ALTO+AN+Hum+EN=elder
velho,A+FLX=ALTO+NAV+Apred+EN=old
```

became the following new entries:

```
velho,A+FLX=ALTO+AB+class+EN=vintage+IH+Table=SEAHP3+Nhum
+Vcopser+Vcopestar+Vcopencontrarse+Vcopsentirse+Vcoptornarse
+UMNclas+UmModif+AdvQuant+Superlativo+NAdj+DRV=A2N164:CASA
+DRV=A2V67:AGRADECER

velho,A+FLX=ALTO+AN+Hum+EN=elder+IH+Table=SEAHP3
+Nhum+Vcopser+Vcopestar+Vcopencontrarse+Vcopsentirse+Vcoptornarse
+UMNclas+UmModif+AdvQuant+Superlativo+NAdj+DRV=A2N164:CASA
+DRV=A2V67:AGRADECER

velho,A+FLX=ALTO+NAV+Apred+EN=old+IH+Table=SEAHP3+Nhum
+Vcopser+Vcopestar+Vcopencontrarse+Vcopsentirse+Vcoptornarse
+UMNclas+UmModif+AdvQuant+Superlativo
+NAdj+DRV=A2N164:CASA+DRV=A2V67:AGRADECER
```

Initially, this process was done blindly, i.e., the Port4NooJ entries were not checked for properties that excluded them from being human intransitive adjectives, and that, consequently, should not receive lexicon-grammar attributes. In a second round, entries with at least the attribute `+AB`, i.e., adjectives that are classified as "abstract", such as the first entry above, should be discarded to obtain a more accurate version of the dictionary of human intransitive adjectives.

When the adjective did not exist in Port4NooJ, new entries were created. This happened for different reasons: the adjective was missing from Port4NooJ (e.g., *abissínio*), it derived from other base form (e.g., *arranhado* is the past participle of *arranhar*) or had another part of speech tag in Port4NooJ (e.g., *solteiro* in Port4Nooj is a noun only). In any of those cases, the inflection code was assigned automatically given that the suffixes of the human intransitive adjectives were very regular. A few exceptional adjectives with less productive suffixes were missing FLX codes. Those entries were reviewed by linguists and their codes were assigned manually. New FLX codes and corresponding inflectional paradigms were created as needed. All other properties of new adjectival entries came from the lexicon-grammar tables:

```
abissínio,A+FLX=ALTO+IH+Table=SAN+Nhum+Vcopser+Vcoptornarse+UMNclas+UmModif+
NclassPserde+NclassPorigem+NclassPnacionalidade+NclassPnaturalidade+NAdj+
Top=Abissínia+TopDET=a
(no entry in Port4Nooj)
```

```
arranhado,A+FLX=ALTO+IH+Table=EAHP2+Nhum+NapdeNhum+Npc+Vcopestar+AdvQuant+
Superlativo+NAdj+NhumVopAPrepNap+deemEDefNap+DRV=A2N4:BALÃO+DRV=A2V2:FALAR+
Reflexivo
(In Port4Nooj: arranhar,V+FLX=FALAR...)
```

```
solteiro,A+FLX=ALTO+IH+Table=SEAHP3+Nhum+Vcopser+Vcopestar+Vcopficar+
Vcoppermanecer+Vcopencontrarse+UMNclas+UmModif+Superlativo+NAdj
(In Port4NooJ: solteiro,N+FLX=ANO+AN+des+EN=bachelor)
```

## 6.2 From Lexicon-Grammar Tables to NooJ Grammars

Syntactic grammars in NooJ can be described and used in two different ways: for syntactic parsing, and for transformational analysis. We explored both, as described in Sects. 6.2.1 and 6.2.2. However, for the time being, eSPERTo is generating paraphrases through syntactic parsing.

### 6.2.1 Option 1: Syntactic Parsing

NooJ syntactic grammars that are used to parse a text, need to describe for each input the corresponding paraphrases that will be generated in the output. For example, the possibility of having the indefinite article, a construction common to several tables, can be described by recognizing the sequence without the article, and generating the construction including the determiner, or the reverse. However, we are duplicating information by swapping the input with the output

(cf. top path with bottom path in Fig. 4). In the case of just two equivalent constructions, this is not a big problem. In the case of a set with more than two paraphrases (e.g., *o homem americano | o homem dos EUA | o homem de origem americana | o homem de nacionalidade americana |* etc.), the recognition of all constructions in the set and the generation of the alternative constructions, would require at least $n \times (n-1)$ paths, where $n$ is the number of paraphrases in the set. For example, the grammar in Fig. 5 recognizes only *o homem americano* and generates the corresponding paraphrases. Similar grammars would have to be constructed for each paraphrase to be also recognized in a text.

### 6.2.2   Option 2: Transformational Module

A better description than option 1 is to represent that the construction without the indefinite article (top path in Fig. 4) and the construction with the indefinite article (bottom path in Fig. 4) are equivalents, as described in Fig. 6. This equivalency is expressed through the use of the global variable @A. That grammar can then be used in the transformational module to generate all the equivalent constructions.
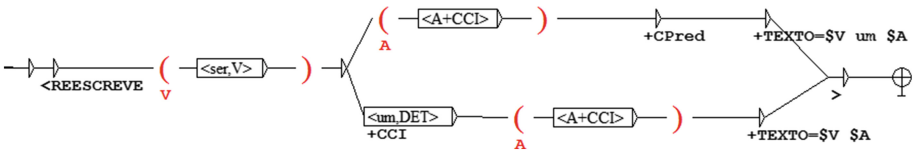


**Fig. 4.** Characterizing indefinite constructions: paraphrasing through parsing
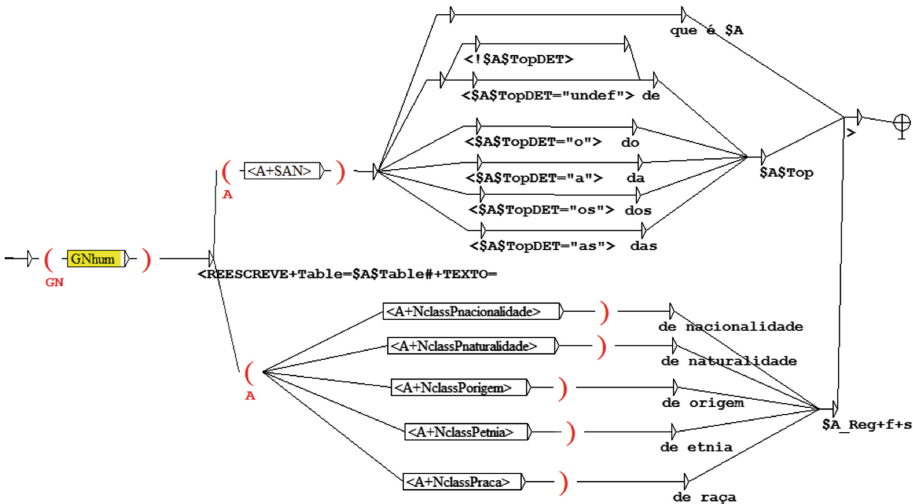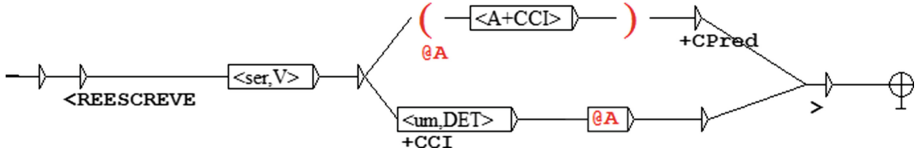


**Fig. 5.** Paraphrasing constructions involving patronymic adjectives

**Fig. 6.** Characterizing indefinite constructions: paraphrasing through transformational analysis

## 7    Preliminary Results

Port4NooJ dictionary formalizes 40,336 lemmas that recognize 1,006,424 word forms. There are 13,051 entries formalizing adjectives that correspond to 6,115 different adjectives.

The new standalone dictionary of human intransitive adjectives integrated in Port4NooJ includes 5,177 entries, that correspond to 4,138 different adjectives. Table 2 shows, for each lexicon-grammar table, how many adjectives existed already in Port4NooJ, and how many were added. Only 26 % of the adjectives for-

**Table 2.** Statistics on the merge between human intransitive adjectives and Port4NooJ adjectives

| Table | In Port4NooJ | New | % In |
|---|---|---|---|
| EAHP2 | 18 | 86 | 17 % |
| EAHP3 | 54 | 212 | 20 % |
| SAF | 70 | 233 | 23 % |
| SAHC1 | 115 | 228 | 34 % |
| SAHC2 | 41 | 176 | 19 % |
| SAHC3 | 54 | 464 | 10 % |
| SAHP1 | 303 | 246 | 55 % |
| SAHP2 | 142 | 225 | 39 % |
| SAHP3 | 97 | 262 | 27 % |
| SAN | 108 | 543 | 17 % |
| SEAD | 39 | 148 | 21 % |
| SEAHC2 | 14 | 40 | 26 % |
| SEAHC3 | 15 | 52 | 22 % |
| SEAHP2 | 53 | 86 | 38 % |
| SEAHP3 | 32 | 92 | 27 % |
| Total | 1155 | 3093 | 26 % |

malized in the lexicon-grammar tables were in Port4NooJ already[10]. This means that the number of different adjectives in Port4Nooj increased about 50 %.

A few grammars were constructed that explore the information in the new dictionary to extend eSPERTo paraphrase knowledge. We started by developing grammars to recognize and paraphrase (i) constructions involving patronymic adjectives, (ii) characterizing indefinite constructions, (iii) the possibility of alternating Vcop *ser* and *estar* with other aspectual variants, and (iv) cross constructions.



**Fig. 7.** Using properties of the human intransitive adjectives in noun phrase grammars

The information in the new dictionary of human intransitive adjectives was also used to improve the recognition of human noun phrases (see Fig. 7).

## 8   Conclusions and Future Work

We successfully integrated 15 lexicon-grammar tables describing the distributional properties of human intransitive adjectives into Port4NooJ by creating a standalone dictionary of human intransitive adjectives and by creating grammars that use information provided by the new dictionary to describe equivalent constructions involving those adjectives. In this way, we extended eSPERTo paraphrasing capabilities.

In the near future, we intend to: (i) create additional grammars to recognize the remaining constructions formalized in lexicon-grammar tables of human intransitive adjectives; (ii) revise and evaluate the new resources; (iii) integrate

---

[10] The difference between the number of different adjectives (4,138) and the number of adjectives in the lexicon-grammar tables (4,248) is due to a few adjectives belonging to more than one table as they have different meanings and distributional properties. One such adjective is *tonto*, which means either *dizzy* or *foolish/silly*. In the first case, *tonto* only occurs with Vcop *estar* as formalized in SEAD table of disease adjectives: *ele está tonto* (*he is dizzy*), whereas, in the second case, *tonto* only occurs with Vcop *ser* as formalized in table SAHC1: *ele é tonto* (*he is foolish/silly*).

and adapt additional lexicon-grammar tables, such as the ones formalizing constructions with Vsup *ser de* (Baptista 2000) and Vsup *fazer* (Chacoto 2005).

We will also use the Port4NooJ paraphrase knowledge to annotate a corpus with paraphrases. This corpus will be used to develop, train and test the eSPERTo's hybrid paraphrase acquisition engine. In turn, the new paraphrases will be merged with the existing paraphrases in Port4NooJ.

# References

Baptista, J.: Sintaxe dos Predicados Nominais construídos com o verbo suporte SER DE. Ph.D. thesis. Universidade do Algarve, Faro, Portugal (2000)

Baptista, J., Mamede, N., Markov, I.: Integrating a lexicon grammar of verbal idioms in a Portuguese NLP system. PARSEME General Meeting, Athens, 10–11 March 2014, Poster Session (2014)

Barreiro, A.: Port4NooJ: Portuguese linguistic module and bilingual resources for machine translation. In: Blanco, X., Silberztein, M. (eds.) Proceedings of the 2007 International NooJ Conference, pp. 19–47. Cambridge Scholars Publishing, Newcastle upon Tyne (2008)

Barreiro, A.: Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation. Ph.D. thesis. Porto, Portugal: Universidade do Porto (2009)

Barreiro, A.: Linguistic resources and applications for Portuguese processing and machine translation. In: Kuti, J., Silberztein, M., Váradi, T. (eds.) Applications of Finite-State Language Processing: Selected Papers from the NooJ 2008 International Conference, pp. 41–51. Cambridge Scholars Publishing, Newcastle upon Tyne (2010)

Barreiro, A.: SPIDER: a system for paraphrasing in document editing and revision — applicability in machine translation pre-editing. In: Gelbukh, A. (ed.) CICLing 2011, Part II. LNCS, vol. 6609, pp. 365–376. Springer, Heidelberg (2011). ISBN: 978-3-642-19436-8

Barreiro, A., et al.: OpenLogos rule-based machine translation: philosophy, model, resources and customization. Mach. Transl. **25**(2), 107–126 (2011)

Carvalho, P.: Análise e Representação de Construções Adjectivais para Processamento Automático de Texto. Adjectivos Intransitivos Humanos. Ph.D. thesis. Universidade de Lisboa (2007)

Chacoto, L.: O Verbo Fazer em Construções Nominais Predicativas. Ph.D. thesis. Universidade do Algarve (2005)

Machonis, P.: English phrasal verbs: from lexicon-grammar to natural language processing. South. J. Linguist. **34**(1), 21–48 (2010)

Mota, C., et al.: The Logos legacy in the eSPERTo paraphrasing system. In: Turing's Footsteps. Logos Contribution to Machine Translation and Natural Language Processing (2015, forthcoming)

Scott, B.(Bud): The Logos model: an historical perspective. Mach. Transl. **18**(1), 1–72 (2003). ISSN: 0922-6567

Silberztein, M.: La formalisation des langues: l'approche de NooJ. ISTE, Londres, p. 426 (2015)

Vietri, S.: The formalization of Italian lexicon-grammar tables in a NooJ pair dictionary/grammar. In: Kuti, J., Silberztein, M., Váradi, T. (eds.) Applications of Finite-State Language Processing: Selected Papers from the NooJ 2008 International Conference, pp. 138–147. Cambridge Scholars Publishing, Newcastle-upon-Tyne (2010)