Port4NooJ v3.0: Integrated Linguistic Resources for Portuguese NLP

Cristina Mota[1], **Paula Carvalho**[1,2], **Anabela Barreiro**[1]
[1]L2F / INESC-ID Lisboa, [2]Universidade Europeia | Laureate International Universites
cmota@ist.utl.pt, {pcc,abarreiro}@inesc-id.pt

Abstract

This paper introduces Port4NooJ v3.0, the latest version of the Portuguese module for NooJ, highlights its main features, and details its three main new components: (i) a lexicon-grammar based dictionary of 5,177 human intransitive adjectives, and a set of local grammars that use the distributional properties of those adjectives for paraphrasing (ii) a polarity dictionary with 9,031 entries for sentiment analysis, and (iii) a set of priority dictionaries and local grammars for named entity recognition. These new components were derived and/or adapted from publicly available resources. The Port4NooJ v3.0 resource is innovative in terms of the specificity of the linguistic knowledge it incorporates. The dictionary is bilingual Portuguese-English, and the semantico-syntactic information assigned to each entry validates the linguistic relation between the terms in both languages. These characteristics, which cannot be found in any other public resource for Portuguese, make it a valuable resource for translation and paraphrasing. The paper presents the current statistics and describes the different complementary and synergic components and integration efforts.

Keywords: paraphrasing, sentiment analysis, named entity recognition

1. Introduction

NooJ (Silberztein, 2015; Silberztein, 2016) is a multilingual linguistic environment to develop linguistic resources, and process written texts. Currently, NooJ includes over 20 language modules. Port4NooJ is the Portuguese language module of NooJ, and it includes linguistic resources, such as: (i) a large coverage dictionary with English transfers[1]; (ii) rules to formalize and document Portuguese inflectional and derivational descriptions, and (iii) local grammars, namely morphological, disambiguation, semantico-syntactic, morphological grammars to inflect and generate multiword expressions, and translation grammars. Port4NooJ different components interact among them and are used to process texts. Several processing functions can be performed with these resources, among others, part of speech annotation, semantic analysis, named entity recognition, translation and paraphrasing. The module can be downloaded from the NooJ website[2].

In this paper we present the Port4NooJ v3.0, focusing in the description of its three main new components: (i) a lexicon-grammar based dictionary of human intransitive adjectives, and a set of local grammars that use the distributional properties of those adjectives for paraphrasing, (ii) a polarity dictionary for sentiment analysis, and (iii) linguistic resources for named entity recognition. There is no other public resource for Portuguese that incorporates such a variety of linguistic knowledge.

This research work was developed in the scope of the eSPERTo[3] project. The main objective of this project is twofold: (i) develop 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, and (ii) develop 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 online 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. 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.

2. Port4NooJ Module

The initial Port4NooJ linguistic resources derive from OpenLogos. OpenLogos is an open source derivative of the commercial Logos system downloadable from the DFKI website[4], and available at INESC-ID.[5] In order to create Port4NooJ, the OpenLogos English-Portuguese dictionary was converted into NooJ format and enhanced with new properties, including derivational and morpho-syntactic and semantic relations that allowed generation of paraphrases for Portuguese (Barreiro, 2009). These paraphrases are used to feed the linguistic engine of the eSPERTo's paraphrasing system[6], which is based on the NooJ technology (Silberztein, 2015).

[1] Due to its bilingual characteristics, this dictionary can be used in 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.

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

[3] 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.

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

[5] http://www.l2f.inesc-id.pt/openlogos/demo.html. The Logos system was built on the Logos Model (Scott, 2003; Scott, forthcoming; Barreiro et al., 2011).

[6] http://esperto.l2f.inesc-id.pt/

PoS	Lemmas	Inflected Forms
N	14046	120,231
A	13059	196,235
V	9548	677,727
DET	262	471
PRO	156	264
PREP	288	288
CONJ	161	164
ADV	2643	15,363
Other	231	316
Total	40394	1,011,059

Table 1: Port4NooJ lexicon: PT-Dict 2.0

In Port4NooJ dictionaries, each dictionary entry is represented by its lemma and contains information on its part-of-speech (PoS), inflectional paradigm (FLX), semantico-syntactic properties (SAL), and corresponding English transfer (EN). The "transfer" is the translation of a word, which is disambiguated with the SAL knowledge inherited from the Logos Model (Table 1 shows the distribution of entries by part of speech tags in Port4NooJ's main dictionary – PT-Dict 2.0). In Logos terminology, SAL stands for Semantico-Syntactic Abstraction Language (SAL), which plays a very important role in the translation of each word in context.

The dictionary entry for the Portuguese noun *mesa* (*table*) has the following structure:

```
mesa,N+FLX=CASA+SAL=COsurf+EN=table
```

The entry is represented by (i) its lemma: *mesa*, (ii) its PoS, noun (N), (iii) its inflectional paradigm: FLX=CASA (*mesa* inflects according to paradigm class CASA, i.e. inflects like the noun *casa* (*house/home*), the example word to represent the paradigm), (iv) its semantico-syntactic properties: SAL=COsurf, which stands for a concrete noun surface, and (v) its English transfer: *EN=table*.

2.1. The SAL Property

One important characteristic of the Port4NooJ dictionary is that it already provides some level of disambiguation by the use of SAL. For example, there is an entry for *corredor* (*runner*) and another entry for *corredor* (*hallway*), the first one classified as an animate noun denoting a profession or other human designation (SAL =ANdes) and the second one defined as a place that has the general structure of a path (SAL=PLpath).

Words integrate different hierarchical ontology classes and subclasses, according to their linguistic attributes (in the Logos Model, *supersets*, *sets* and *subsets*). Accordingly in the Port4NooJ dictionary, syntactic-semantic (SAL) properties provided for each entry represent these ontological classes and subclasses. Table 2 illustrates SAL properties for the entries *cão*, *vestido*, *cidade*, *sair*, *português*, *feliz*, and *coerentemente*.

If the word is highly polysemous, such as the verbs *raise*, the disambiguation needs to take place by the application of disambiguation grammars, in Logos terminology, the Semtab rules, which assign different transfers for that particular word.

Word	SAL
lobo (wolf)	common noun, warm-blooded vertebrate animal, mammal
país (country)	common noun, agentive proper name denoting a geographic place, geographical entity, and geographical location
correr (run)	motional intransitive verb

Table 2: SAL properties for words of several PoS

2.2. Multiword Units

Port4NooJ contains different types of multiword units or compound words. In particular, it includes invariable compounds of general language, i.e., closed word classes (mostly grammatical words) such as adverbs, prepositions, pronouns, conjunctions and numeric expressions:

```
a curto prazo,ADV+TEMP+EN=in the short run
a favor de,PREP+CAUS+EN=in favor of
cada um,PRO+INDEF+SG+EN=each one
de quem,INT+ThatType+EN=whose
quem quer que seja,REL+WhateverType+EN=whoever
além disso,CONJ+COOR+EN=besides
um quarto,NUM+frac+EN=one fourth
```

A dictionary of multiwords is currently under development and comprises nominal expressions such as *cabo de vassoura* (*broomstick*) or *luz solar* (*sunlight*); verbal expressions, such as *marcar pontos* (*score*) or *piscar o olho* (*wink*); adjectival expressions such as *fraco de espírito* (*feeble-minded*), *cor-de-rosa* (*pink*) or *norte-americano* (*North-American*); adverbial expressions, such as *com entusiasmo* (*enthusiastically*) or *de parte* (*aside*).

2.3. Support Verb Constructions

Meyers et al. (2004) present evidence to consider suport verb constructions (SVCs) as multiwords, based on predicate-argument structure. Extended syntactic and semantic information was added to the lexicon, in order to formalize and process them. In particular, (i) 33 support verbs were identified as such in the lemma dictionary with the attribute +SUP, (ii) nominalizations that are predicative nouns were assigned the attributes +Npred+Nom, and were formalized by adding derivation codes to 4,968 verbal entries, which allow the generation of the corresponding predicative noun, and (iii) the attribute +VSUP is assigned the support verb that occurs with the predicative noun in the SVC. The following example illustrates the formalization of a support verb (*fazer*), a verbal entry that has a corresponding nominalization, which is a predicative noun (*ameaçar*) and the derived nominalization *ameaça*.

```
fazer,V+AUX+SUP+FLX=FAZER+Aux=1+PREVDIbid-type
       +Subset=581+EN=make+SVB
ameaçar,V+FLX=COMEÇAR+Aux=1+PREVfailType+Subset=492
     +EN=threaten+VSUP=fazer+DRV=NDRV16:CASA
ameaça,ameaçar,N+Npred+Nom+FLX=COMEÇAR+Aux=1
     +RECTR52+Subset=142+EN=menace+VSUP=fazer
     +DRV=NDRV16:CASA+f+s
```

Autonomous predicate nouns, i.e., predicate nouns which do not have a morpho-syntactic or semantic relation with a verb or adjective (non-nominalizations) (Gross, 1982) are formalized directly in the lemma dictionary and are assigned only the attribute +Npred, such as in:

```
barulho,N+Npred+FLX=ANO+ME+abs+EN=noise+VSUP=fazer
```

827 lemma entries also include aspectual or stylistic variants of the main support verb, which is marked with the attribute +VSTYLE as in the following entry:

```
visitar,V+FLX=FALAR+Aux=1+OBHUM68+Subset=515
        +EN=visit+VSUP=fazer+VSTYLE=efectuar
        +VSTYLE=realizar+DRV=NDRV16:CASA
```

2.4. Inflectional and Derivational Descriptions

Inflectional paradigms are independent standard pattern models (prototypes) based on morphological suffixation rules. These rules cover variation in gender and number (adjectives and nouns), person (verbs and pronouns), tense (verbs), diminutives, augmentatives and superlatives (nouns, adjectives and some adverbs), and nominalizations.
In general, NooJ dictionaries are connected with inflectional and derivational descriptions, which can be represented in the form of graphs or plain rules, for simple words and compounds.
Port4NooJ contains 310 inflectional paradigms, and 738 derivational paradigms.

2.5. Grammars

NooJ grammars are represented by finite state transducers (FSTs); i.e., graphs that represent many different linguistic phenomena and can be used to add annotations to the text, or to filter out annotations. Local grammars describe frozen or semi-frozen phenomena or morphological phenomena. Syntactic grammars are used to disambiguate a word, perform an active to passive transformation or a syntactic agreement check and describe phrase and sentence structure. Grammars for multiword units can describe phrases such as SVCs and other multiwords. Grammars can also be used for semantic analysis and representation of named entities and paraphrases, for example, and we have also been using them to perform translation.

2.5.1. Morphological Grammars

Port4NooJ includes a morphological grammar to process contracted forms such as *das* (EN: *of the*) resulting from the contraction of the preposition *de* (EN: *of*) and the determiner *as* (EN: *the*), or *neste* (EN: *in this*) resulting from the contraction between preposition em (EN: *in*) and the demonstrative pronoun *este* (EN: *this*).
When applying the contracted forms grammar to text, during the normalization phase, it is possible to decompose the word in its different basic constituents.

2.5.2. Syntactic Grammars

Whereas morphological grammars are applied at the word level, by contrast, syntactic grammars are applied at the phrase and sentence level. They can be used for identifying and annotating both syntactic patterns, and semantic units. More specifically they can perform local, structural and transformational analysis.
Port4NooJ includes grammars to: (i) identify and annotate dates and temporal expressions, (ii) disambiguate words or sequences of words, i.e., to filter out lexical or syntactic annotations in the text, (iii) to paraphrase several types of constructions, and (iv) to translate simple sentences.
In the following sections, we describe the new three main components of Port4NooJ: (i) a lexicon grammar based dictionary of human intransitive adjectives for paraphrasing, (ii) a polarity lexicon for sentiment analysis, and (iii) a set of linguistic resources for named-entity recognition.

3. Lexicon-Grammar of Human Intransitive Adjectives for Paraphrasing

Carvalho (2007) studied, formalized, and classified the distributional properties of human intransitive adjectives in lexicon-grammar tables, corresponding to 15 semantico-syntactic classes: disease, membership, geographical adjectives, such as nationality, and 12 generic human adjective subclasses. The 12 subclasses are based on the auxiliary verbs with which they co-occur: either *ser*, *estar* or both, and also based on the possibility of being preceded by an indefinite article, and finally, by 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.
The properties formalized in the lexicon-grammar tables enable eSPERTo to paraphrase:

- adjective, noun and verb morphologically related constructions: *está zangado* (*is angry*) = *zangou-se* (*got* (*self*) *angry*) = *esteve envolvido numa zanga* (*was involved in anger*);
- adjective constructions supported by different copulative verbs: *estar perdido* (*be lost*) = *andar perdido* (*walk around lost*);
- constructions involving nationality and other membership relations: *de origem portuguesa* (*of Portuguese origin/roots*) = *portugueses* (*Portuguese*) = *de Portugal* (*from Portugal*), *benfiquista* (*Benfica fan*) = *do Sport Lisboa e Benfica* (*a fan of Sport Lisboa e Benfica*);
- 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* (*was moderated in his comments*) = *os seus comentários foram moderados* (*his comments were moderated*) = *foi moderado* (*was moderated*);
- generic noun phrases: *é um indivíduo estúpido* (*he is a fool*) = *é um estúpido* (*he is a fool*) = *é estúpido* (*he is a fool*).

As referred in Mota et al. (forthcoming), which describes the detailed process of integrating the lexicon grammar tables into Port4NooJ, the new standalone dictionary of human intransitive adjectives of Port4NooJ includes 5,151 entries, corresponding to 4,138 different adjectives. Given

Table	Lemmas
SAHP1	818
SAN	676
SAHC3	561
SAHP2	498
SAHC1	441
SAHP3	435
SAF	326
EAHP3	317
SAHC2	238
SEAHP2	209
SEAD	203
SEAHP3	158
EAHP2	137
SEAHC3	72
SEAHC2	62
Total	5151

Table 3: Distribution of adjectives by table attribute after integration into Port4NooJ

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 3 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 derivations that already made part of Port4NooJ. In cases where the derivation did not exist, new derivational descriptions (1,202) 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 toponymms. 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.

```
Norte da Europa,N+Top+TopDET=o+Adj=nórdico
Tunísia,N+Top+TopDET=a+Adj=tunisino
```

This dictionary does not distinguish the types of toponyms, but that information can be inferred from the corresponding adjective properties. For instance, *tunisino* has the attribute +NclassPnacionalidade, but *nórdico* does not, which means the first *Tunísia* is a country, but *Norte da Europa* is not:

```
tunisino,A+FLX=ALTO+AN+des+EN=Tunisian+Table=SAN
  +Nhum+Vcopser+Vcoptornarse+UMNclas+UmModif
  +NclassPserde+NclassPorigem+NclassPnacionalidade
  +NclassPnaturalidade+NAdj+DRV=A2NTop491:HOLANDA
  +TopDET=a
nórdico,A+FLX=ALTO+Table=SAN+Nhum+Vcopser+UMNclas
  +UmModif+NclassPserde+NclassPorigem+NclassPnatu
  ralidade+NAdj+DRV=A2NTop361:CANADA+TopDET=o
```

	PoS	Lemmas	Inflected Forms
Before	N	1,081	1,280
	A	4,779	16,863
	V	489	29,504
	Idioms	666	34,700
	Total	7,014	82,347
After	N	1,308	17,331
	A	5,840	83,840
	V	1,263	80,459
	ADV	0	4,012
	Idioms	620	44,034
	Total	9,031	229,676

Table 4: SentiLex-PT: POS tag and idiomatic expression distribution before and after integration into Port4NooJ

	Target	Lemmas	Inflected Forms
Before	N0	6,550	55,184
	N0 & N1	456	26,796
	N1	7	366
After	N0	7,924	150,817
	N0 & N1	1,102	78,513
	N1	5	346

Table 5: SentiLex-PT: target distribution before and after integration into Port4NooJ

About 27% of SAN adjectives are indeed derived from country names.

A first set of grammars was also constructed to extend eSPERTo's paraphrastic knowledge. These grammars 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.

4. SentiLex-PT: a Polarity Lexicon for Sentiment Analysis

SentiLex-PT is a sentiment lexicon to mine social judgments from Portuguese texts, i.e., to detect opinions about human entities (Silva et al., 2012). SentiLex-PT's dictionary contains 7,014 lemmas that generate 82,347 inflected forms (see Table 4 for a distribution of PoS tag and idiomatic expressions).

SentiLex-PT's lexicon entries correspond to human predicates, i.e., predicates modifying human nouns, compiled from different publicly available resources (corpora and dictionaries). The sentiment attributes for each entry are (i) the target of sentiment, (ii) the predicate polarity, and (iii) the polarity assignment. Most entries were manually labeled regarding polarity, but some adjectives had their attributes assigned automatically. The inflected forms associated with the verbs and idiomatic expressions, and their corresponding morphological attributes, were extracted semi-automatically from LABEL-LEX, a publicly available lexicon for Portuguese, developed by Ranchhod et al. (2004).

The integration of SentiLex-PT into Port4NooJ consisted in converting each entry of the lemma dictionary to the NooJ format according to the following procedure:

	Target	Polarity	Lemmas	Inflected Forms
Before	N0	-	4,598	53,658
	N0	0	860	7,704
	N0	+	1,550	20,667
	N1	-	245	14,562
	N1	0	178	10,249
	N1	+	39	2,299
After	N0	-	5,474	132,539
	N0	0	1,155	28,598
	N0	+	2,397	68,193
	N1	-	597	42,021
	N1	0	423	30,563
	N1	+	87	6,275

Table 6: SentiLex-PT: polarity assignment distribution before and after integration into Port4NooJ

1. Convert the polarity properties of the SentiLex lemma into NooJ format;
2. If the lemma exists in Port4NooJ then add the converted properties to the Port4NooJ entry;
3. Otherwise, create a new entry by automatic assigning an inflectional paradigm to the lemma and adding the converted properties to the new entry; if it is not possible to assign an inflectional paradigm, then those entries require human revision;
4. Add the entry to a new lemma dictionary, unless inflectional code assignment was impossible to achieve.

After following the previous procedure, a new lemma dictionary was created with 9,196 entries, of which 165 needed manually assigned inflectional paradigms. Table 4 shows the distribution of part of speech tags before and after the conversion.[7] The number of entries in SentiLex is higher after conversion to Port4NooJ, because Port4NooJ includes English transfers for each Portuguese dictionary entry. In many cases, however, the transfer does not correspond to the word formalized in SentiLex and should be removed after manual revision.

About 28% of the 7,014 lemmas of SentiLex (i.e., 1,994 lemmas) already existed in Port4NooJ, and were hence merged with the homographic lemmas of Port4NooJ. We created 53 new inflectional paradigms for idioms from inflectional paradigms that already existed in Port4NooJ.

5. Resources for Named Entity Recognition

Mota (2009) described a named entity tagger for Portuguese that recognizes and annotates person, location and organization names. The tagger includes a module implemented in NooJ that identifies candidate named entities and surrounding contexts that will be classified based on a co-training algorithm. This module is comprised of:

(i) a Portuguese dictionary with 1.3 million inflected forms[8];

Profissao	1581	cantor	singer
Cargo	67	ministro	minister
Titulo	30	barão	baron
Parentesco	29	tio	uncle
Tratamento	9	senhor	mister
Filiacao	7	membro	member
Patrio	530	francês	French
Org+Head	83	universidade	university
Emp+Head	25	livraria	bookstore
Org+Dep	8	departamento	department
Geo	29	rio	river
GeoP	9	país	country
Construcao	4	jardim	park
Total	2411		

Table 7: Priority dictionary of specific words for NER

(ii) several priority dictionaries that help prevent future ambiguities that would produce incorrect chunking and, hence, imprecise named entity extraction. These priority dictionaries (Table 7 include:

- 81 lemmas that may be in the beginning of a sequence of proper names where the complete sequence is not a person, organization or location (e.g., *protocolo* (EN: *protocol*));
- words that should be delimited as independent names when occurring in the beginning of a sequence of proper names, such as *parlamento europeu* (EN: *European Parlament*), *secretaria de estado* (EN: *State Department*);
- 1,724 nouns that in general precede person names, such as names of job titles (e.g., *presidente* – EN: *president*) and occupations (e.g., *jornalista* – EN: journalist);
- 530 patronymic adjectives (e.g., *nortenho* (EN: *northern*) and *americano* (EN: *American*));
- 162 grammatical words that are ambiguous with nouns, adjectives or verbs. These words were assigned only the grammatical POS tag. For example, the conjunction *e* (EN: *and*) that is ambiguous with the noun that refers to the letter *e*, was added to this dictionary only as conjunction.

(iii) a set of chunking grammars that: identify and tag sequences of words that are named entity candidates, multiword cardinals, ordinals and determiners, particular noun phrases, and sequences of auxiliary verbs;

(iv) a set of local grammars that identify pairs of named entity candidates and their surrounding contexts, among the following types of contexts: proper name in the context of a noun phrase, proper name in the left context of a verb, proper name in the right context of a verb, coordination of proper names, apposition involving proper names, and proper name within an age context.

Integrating in Port4NooJ the previous resources that were specifically developed for named entity recognition is fairly

[7] It excludes the entries for which an inflectional paradigm will be manually assigned.

[8] This dictionary is the NooJ version of the Lusolex dictionary created by L2F at INESC-ID (Wittmann et al., 2000).

straightforward, because they were already in NooJ format. Besides minor updates to the grammars due to NooJ evolving since they were first created, the main modification that we introduced was to remove from the priority dictionaries adjectival entries marked with the attribute +Patrio, assigned to nationality adjectives. When those entries existed in the dictionary of human intransitive adjectives with the attribute +Table=SAN, assigned to geographical adjectives, including nationality adjectives, they were eliminated. About 65% of the entries marked +Patrio (i.e., 342 entries) were hence removed, leaving

However, in order to make the resources fully compatible with Port4NooJ, i.e., so they can run based on Port4NooJ main dictionary, PT-Dict 2.0, instead of Lusolex 3.0, it is necessary to either add the equivalent information to the grammars, in the case of the attributes assigned to the same word in Port4Nooj and Lusolex are different, or create a new set of grammars that only uses attributes from Port4NooJ.

Ideally, we create a new dictionary that merges information from the two Portuguese dictionaries. Even assuming that Lusolex can be made publicly available, Port4NooJ dictionary has at least one feature that makes the merging task more challenging: many predicative nouns are derived from their equivalent verbal predicates. This means that a noun in Lusolex may not exist in Port4NooJ as a lemma – in such case, we should not add it to the merged dictionary.

6. Conclusions and Future Work

In comparison to other available public resources, Port4NooJ offers two strong competitive advantages, its bilingual nature and the integration of semantico-syntactic knowledge associated to each entry, characteristics that make possible its use in complex natural language processing tasks, such as translation and paraphrasing.

In this paper, we presented three 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. Furthermore, in the near future we envisage to improve eSPERTo's paraphrasing capabilities by integrating additional lexicon-grammar tables, such as of the nominal predicative constructions with Vsup *ser de* (Baptista, 2000) and Vsup *fazer* (Chacoto, 2005).

Acknowledgements

This research work was supported by Fundação para a Ciência e Tecnologia (FCT), under project eSPERTo – EXPL/MHC-LIN/2260/2013, UID/CEC/50021/2013, and post-doctoral grant SFRH/BPD/91446/2012.

7. Bibliographical References

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

Barreiro, A., Scott, B., Kasper, W., and Kiefer, B. (2011). Openlogos rule-based machine translation: Philosophy, model, resources and customization. *Machine Translation*, 25(2):107–126.

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

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

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

Gross, M. (1982). Une classification des phrases «figées» du français. *Revue québécoise de linguistique*, 11(2):151–185.

Meyers, A., Reeves, R., and Macleod, C. (2004). Np-external arguments a study of argument sharing in english. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 96–103. Association for Computational Linguistics.

Mota, C., Carvalho, P., Raposo, F., and Barreiro, A., (forthcoming). *Generating Paraphrases of Human Intransitive Adjective Constructions with Port4NooJ*. Communications in Computer and Information Science. Springer.

Mota, C. (2009). *How to keep up with language dynamics: A case study on named entity recognition*. Ph.D. thesis, Instituto Superior Técnico, Maio.

Ranchhod, E., Carvalho, P., Mota, C., and Barreiro, A. (2004). Portuguese large-scale language resources for nlp applications. In *LREC*. European Language Resources Association.

Scott, B. (2003). The Logos Model: An Historical Perspective. *Machine Translation*, 18(1):1–72.

Scott, B. (forthcoming). *Language, Brains and the Computer*.

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

Silberztein, M. (2016). *Formalizing Natural Languages: the NooJ Approach*. Wiley Eds.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In *Computational Processing of the Portuguese Language*, pages 218–228. Springer.

Wittmann, L., Ribeiro, R. D., Pêgo, T., and Batista, F. (2000). Some language resources and tools for computational processing of portuguese at inesc. In *LREC*. European Language Resources Association.