ISSN 2007-9737

A Rule-Based Meronymy Extraction Module for Portuguese

Ilia Markov[1], Nuno Mamede[2,3], Jorge Baptista[3,4]

[1] Instituto Politécnico Nacional, Center for Computing Research,
Mexico City, Mexico

[2] Universidade de Lisboa, Instituto Superior Técnico,
Lisbon, Portugal

[3] INESC-ID Lisboa/L2F - Spoken Language Lab,
Lisbon, Portugal

[4] Universidade do Algarve/FCHS and CECL,
Faro, Portugal

markovilya@yahoo.com, {Nuno.Mamede, Jorge Baptista}@l2f.inesc-id.pt

Abstract. In this article, we improve the extraction of semantic relations between textual elements as it is currently performed by STRING, a hybrid statistical and rule-based Natural Language Processing (NLP) chain for Portuguese, by targeting *whole-part* relation (*meronymy*), that is, a semantic relation between two entities of which one is perceived as a constituent part of the other, or between a set and its member. In this case, we focus on the type of meronymy involving human entities and *body-part nouns* (*Nbp*) (e.g., *O Pedro partiu uma perna* 'Pedro broke a leg': `WHOLE-PART(Pedro,perna)` '`WHOLE-PART(Pedro,leg)`'). In order to extract this type of whole-part relations, a rule-based meronymy extraction module has been built and integrated in the grammar of the STRING system. The module was evaluated with promising results.

Keywords. Whole-part relation, meronymy, body-part noun, disease noun, Portuguese.

1 Introduction

Automatic identification of semantic relations is an important step in extracting meaning out of texts, which may help several other Natural Language Processing (NLP) tasks such as question answering, text summarization, machine translation, information extraction, information retrieval, among others [22]. For example, for questions *What are the components of X? What is Y made of?* and the like, the discovery of whole-part relations is necessary to assemble the right answer. The whole-part relations acquired from a collection of documents are used in answering questions that normally cannot be handled based solely on keyword matching and proximity [23].

For automatic text summarization, where the most important information from a document or set of documents is extracted, semantic relations are useful for identifying related concepts and statements, so a document can be compressed [30]. For example, imagining that one wants to summarize medical reports, where a lot of body-part nouns (henceforward, *Nbp*) and human entities are mentioned, whole-part relation extraction would be relevant to correctly associate the patients' names and their organs' nouns.

Zhang et al. [61] showed that whole-part relations could be used in the NLP task of opinion mining. When one is talking about an object (product), one often refers to its parts and not to the whole, like in the sentence *Neste hotel, o quarto era limpo, as camas eram feitas de lavado todos os dias, e os pequenos almoços eram opíparos* 'In this hotel, the room was clean, the sheets were changed regularly, and the breakfast was sumptuous'. In these cases, if there is a whole-part relation established

doi: 10.13053/CyS-19-4-2255

between the parts and the general product (the whole), one would then see whether the opinion about the general product is positive or not.

Identification of meronymic relations can also be helpful in several anaphora resolution problems. For instance, comparing the sentences *O Pedro partiu o braço* 'Pedro broke the arm' and *O Pedro partiu-lhe o braço* (lit: Pedro broke him the arm), while the *Nbp braço* 'arm' refers to the subject in the first sentence, it refers to the antecedent of the dative pronoun *lhe* 'him' in the second sentence. Thus, correctly establishing whole-part relation can improve anaphora resolution module by providing more information into machine learning systems.

Furthermore, the identification of whole-part relations could benefit semantic role labeling. For example, in the previous sentences, the subject *Pedro* is the EXPERiENCER in the first case, while it becomes the AGENT in the second one, and the EXPERiENCER is now the dative pronoun *lhe* 'him', to which the *Nbp braço* 'arm' is meronymically related. Thus, finding the correct whole-part relation holding between the nouns in these sentences would allow establishing their semantic roles more accurately.

The goal of this work is to improve the extraction of semantic relations between textual elements in STRING[1], a hybrid statistical and rule-based NLP chain for Portuguese [34]. This work will target whole-part relation (*meronymy*), that is, a semantic relation between two entities of which one is perceived as a constituent part of the other, or between a set and its member. In this case, we focus on the type of meronymy involving human entities and *Nbp* in Portuguese. Though STRING already extracts some types of semantic relations [4, 5, 10], meronymic relations are not yet being detected, in spite of the large set of *Nbp* that have already been semantically tagged in its lexicon. In other words, we expect to enhance the system's semantic relation extraction module by capturing meronymic relations.

This paper is structured as follows: Section 2 describes related work on whole-part dependency extraction; Section 3 explains with some detail how this task was implemented in STRING; Section 4 presents the evaluation procedure and the results of the task; Section 5 describes the error analysis, focusing on false positive and false negative cases, and provides a second evaluation of the system's performance, once some of those problems were corrected; Section 6 draws the conclusions from this work and points to the future work by providing possible directions for expanding and improving the module here developed.

2 State of the Art

2.1 Whole-Part Relations

Whole-part relations (also known as *meronymy*) are a type of semantic relation that holds between two elements in a sentence, one that denotes a *whole* and another that denotes a *part*. Meronymy is a complex relation that should not be treated as a single relation, but as a collection of relations [27].

A well-known classification of whole-part relations was developed by Winston et al. [60]. Six types of whole-part relations were distinguished based on the different ways the *parts* contribute to the structure of the *whole*. These include component-integral object (*wheel - car*), member-collection (*soldier - army*), portion-mass (*meter - kilometer*), stuff-object (*alcohol - wine*), feature-activity (*paying - shopping*), and place-area (*oasis - desert*).

Ittoo and Bouma [28] reported that in WordNet [14] whole-part relations are divided into three basic types: Member-of (e.g., *UK* IS-MEMBER-OF *NATO*), Stuff-of (e.g., *carbon* IS-STUFF-OF *coal*), all other whole-part relations under the general name of Part-of (e.g., *leg* IS-PART-OF *table*).

Other classifications, proposed by Odell [41] and Gerstl and Pribbenow [21], are based on the work of Winston et al. [60].

In the taxonomy developed by Keet and Artale [29] there is a distinction between *transitive mereological* whole-part relations and *intransitive meronymic* ones. The distinction consists in that meronymic relations are not necessarily transitive (the fact that A is meronymically related to B and B to C does not mean that A is also meronymically related to C). Intransitivity of 'part of' relations can

[1] https://string.l2f.inesc-id.pt/ [last access: 12/11/2015]. All other URLs in this document were also verified on this date.

be demonstrated by the example *hand-musician-orchestra*, where the inalienable part (*hand*) of an entity whole (*musician*) is not a part of a collective entity whole (*orchestra*). Keet and Artale [29] classify mereological relations into the four following categories: involved-in (*chewing - eating*), located-in (*city - region*), contained-in (*tool - trunk*), structural part-of (*engine - car*); while meronymic relations these authors identify are member-of (*player - team*), constituted-of (*clay - statue*), sub-quantity-of (*meter - kilometer*), participates-in (*enzyme - reaction*).

In our work, we focus on a specific type of whole-part relations involving human body parts. Ittoo and Bouma [28] propose that, in information extraction tasks, focusing on a particular whole-part relation type is likely to give more stable results than using general sets of whole-part relations as seeds for machine learning algorithms:

> “We believe that the traditional practice of initializing IE algorithms with general sets that mix seeds denoting different part-whole relation types leads to inherently unstable results [. . .] Furthermore, general seeds are unable to capture the specific and distinct patterns that lexically realize the individual types of part-whole relations [. . .] This instability strongly suggests that seeds instantiating different types of relations should not be mixed, particularly when learning part-whole relations, which are characterized by many subtypes. Seeds should be defined such that they represent an ontologically well-defined class, for which one may hope to find a coherent set of extraction patterns” [28, p. 1334].

In this work, we are neutral to the classifications suggested above, even though the whole-part relations here studied can fall into *component-integral object* of Winston et al. [60], or into the general part-of case, as in the classification provided by WordNet [14].

According to our review of related work and to a recent review of the literature on semantic relation extraction in Portuguese [1], no work on whole-part relations' extraction, of the specific kind and in the specific context of application here aimed at, have been identified for this language[2]. The current work also aims at extracting a specific type of whole-part relations from real, running texts, involving *Nbp* and human entities, and *not* lexical semantic relations between common nouns (as the relation between *soldier - army*, or *engine - valve*). To our knowledge, no existing Portuguese NLP system (e.g., a parser) detects this type of semantic relations.

Furthermore, and unlike the previous references, we adopt a rule-based approach, using the tools and resources available in STRING. This is done under the scope of a specific and developing NLP chain STRING, built for European Portuguese. It will be seen that most of the relations here targeted also apply to other Portuguese varieties, including the Brazilian Portuguese.

2.2 Whole-Part Relations Extraction

In NLP, various information extraction techniques have been developed in order to capture whole-part relations in texts.

Hearst [25] tried to find lexical correlates to the *hyponymic* relations (type-of relations) by searching in unrestricted, domain-independent text for cases where known hyponyms appear in proximity. For example, in the construction *NP, NP and other NP*, as in 'temples, treasuries, and other civic buildings', the first two terms would be considered as hyponyms of the last term. In other patterns, like *such NP as NP, or/and NP*, as in 'works by such authors as Herrick, Goldsmith, and Shakespeare', the last three terms are considered as hyponyms of the term “author”. Hearst proposed six lexico-syntactic patterns; he then tested the patterns for validity and used them to extract relations from a corpus. To validate his acquisition method, he compared the results of the algorithm with information found in WordNet. Hearst reports that when the set of

[2] At the later stages of this research, we came to know the work of Cláudia Freitas [18]; however, since all the lexicon, grammar rules, and evaluation procedures had already been accomplished by then, we decided not to take it into consideration at this time but to use it in future work.

152 relations that fit the restrictions of the experiment (both the hyponyms and the hypernyms are unmodified) was looked up in WordNet,

> "180 out of the 226 unique words involved in the relations actually existed in the hierarchy, and 61 out of the 106 feasible relations (i.e., relations in which both terms were already registered in WordNet) were found." [25, p. 544].

Hearst claims that he tried applying the same technique to meronymy, but without great success.

Berland and Charniak [7] addressed the acquisition of meronyms using manually crafted patterns, similar to Hearst [25], in order to capture textual elements that denote whole objects (e.g., *building*) and then to harvest possible part objects (e.g., *room*). More precisely,

> "given a single word denoting some entity that has recognizable parts, the system finds and rank-orders other words that may denote parts of the entity in question." [7, p. 57].

The authors used the North American News Corpus (NANC), a compilation of the wire output of a certain number of newspapers; the corpus includes about 1 million words. Their system output was an ordered list of possible parts according to some statistical metrics. They report that their method finds parts with 55% accuracy for the top 50 words ranked by the system and a maximum accuracy of 70% over their top 20 results. The authors report that they came across various problems, such as tagger mistakes, idiomatic phrases, and sparse data - the source of most of the noise.

A lexical knowledge base MindNet [57, 51] was created from dictionary definitions by automatic tools. It has been maintained by the Microsoft NLP research group since 2005 [58], and it is supposedly accessible for on-line browsing.[3] In its creation, a broad-coverage parser generates syntactic trees, to which rules are applied that generate corresponding structures of semantic relations. Thus, a rule-based approach is used in MindNet in order to extract semantic structures from dictionary definitions. The authors also applied their methods for processing free texts, more precisely, the entire text of the Microsoft Encarta Encyclopedia. The only results that the authors present are the number of extracted relations but no evaluation was provided. The structure of MindNet is based on dictionary entries. For each word entry, MindNet contains a record for each word sense, and provides information such as their POS and textual definition. Each word sense is explicitly related to other words. MindNet contains a broad set of semantic (and syntactic) relations, including Hypernym, Location, Manner, Material, Means, Modifier, and Part. Relation paths between words in MindNet are useful for determining word similarity. For example, there are several paths between the words *car* and *wheel*, including not only simple relations like (*car*,`Modifier`,*wheel*) but also paths of length two, like (*car*,`Hypernym`,*vehicle*,`Part`,*wheel*), and longer.

Girju et al. [22, 23] present a supervised, domain independent approach for the automatic detection of whole-part relations in text. The algorithm identifies lexico-syntactic patterns that encode whole-part relations. Classification rules have been generated for different patterns such as genitives, noun compounds, and noun phrases containing prepositional phrases to extract whole-part relations from them. The classification rules were learned automatically through an *iterative semantic specialization* (*ISS*) procedure applied on the noun constituents' semantic classes provided by WordNet. The rules produce semantic conditions that the noun constituents matched by the patterns must satisfy in order to exhibit a whole-part relation. Thus, the method discovers semi-automatically the whole-part lexico-syntactic patterns and learns automatically the semantic classification rules needed for the disambiguation of these patterns. For training purposes the authors used WordNet, the LA Times (TREC9) text collection that contains 3 GB of news articles from different journals and newspapers, and the SemCor collection [40]. From these documents the authors formed a large corpus of 27,963 negative examples and 29,134 positive examples of well distributed subtypes of whole-part relations which provided a

[3] http://stratus.research.microsoft.com/mnex/Main.aspx, currently unavailable.

set of classification rules. The rules were tested on two different text collections: LA Times and Wall Street Journal. The authors report an overall average precision of 80.95% and recall of 75.91%. The authors also state that they came across a large number of difficulties due to the highly ambiguous nature of syntactic constructions.

Van Hage et al. [24] developed a method for learning whole-part relations from vocabularies and text sources. The authors' method learns whole-part relations by

> "first learning phrase patterns that connect parts to wholes from a training set of known part-whole pairs using a search engine, and then applying the patterns to find new part-whole relations, again using a search engine." [24, p. 30].

The authors reported that they were able to acquire 503 whole-part pairs from the AGROVOC Thesaurus[4] to learn 91 reliable whole-part patterns. They changed the patterns' part arguments with known entities to introduce web-search queries. Corresponding whole entities were then extracted from documents in the query results, with a precision of 74%.

The Espresso algorithm [45] was developed in order to harvest semantic relations in a text. Espresso is based on the framework adopted in [25]:

> "It is a minimally supervised bootstrapping algorithm that takes as input a few seed instances of a particular relation and iteratively learns surface patterns to extract more instances." [45, para. 3].

Thus, the algorithm extracts surface patterns by connecting the seeds (tuples) in a given corpus. The algorithm obtains a precision of 80% in learning whole-part relations from the Acquaint (TREC-9) newswire text collection with almost 6 million words.

Thereby, for the English language, it appears that the acquisition of whole-part relation pairs by way of machine learning techniques achieves fairly good results.

Next, in this work, we focus on state-of-the-art relations' extraction in Portuguese, in the scope of ontology building.

2.3 Existing Ontologies for Portuguese and Related Work on Whole-Part Relations

Some work has already been done on building *knowledge bases* for Portuguese, most of which include the concept of whole-part relations. These knowledge bases are often referred to as *lexical ontologies*, because they have properties of a lexicon as well as properties of an ontology [26, 49]. Well-known, existing lexical ontologies for Portuguese are Portuguese WordNet.PT [36, 37], later extended to WordNet.PT Global - Rede Léxico-Conceptual das Variedades do Português (Lexical-Conceptual Network for Portuguese Varieties) [38]; MWN.PT - MultiWordNet of Portuguese[5] [48]; PAPEL - Palavras Associadas (Associated Words) Porto Editora Linguateca[6] [44]; and Onto.PT[7] [43]. Some of these ontologies are not freely available for the general public, while others only provide the definitions associated to each lexical entry without the information on whole-part relations. Furthermore, the type of whole-part relation targeted in this work, involving a human entity and meronymically related *Nbp*, cannot be adequately captured using those resources (or, at least, only those resources).

Some attention was also paid to two well-known parsers of Portuguese in order to discern how they handled the whole-part relations extraction: the PALAVRAS parser [8], consulted using the Visual Interactive Syntax Learning (VISL) environment[8], and LX Semantic Role Labeller[9] [9]. Judging from the available on-line versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly.

In conclusion, the available resources for extracting whole-part relations in Portuguese are inadequate or insufficient for the task of automatic extraction of *human–Nbp* whole-part relations from

[4] http://www.fao.org/agrovoc

[5] http://mwnpt.di.fc.ul.pt/

[6] http://www.linguateca.pt/PAPEL/

[7] http://ontopt.dei.uc.pt/

[8] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/dependency.php

[9] http://lxcenter.di.fc.ul.pt/services/en/LXSemanticRoleLabeller.html

running texts, as those we have targeted here. Besides, we adopt a rule-based approach in order to extract this kind of relations, which differs from other approaches in the literature, mostly focused on part-whole lexical relations (between common nouns) and based on machine learning techniques.

3 Meronymy Extraction Module in STRING

3.1 STRING Overview

STRING [34] [10] performs all the basic steps of natural language processing (tokenization, sentence splitting, POS-tagging, POS-disambiguation, and parsing) for Portuguese texts. The architecture of STRING is given in Fig. 1.

STRING has a modular, pipeline structure, where (i) the preprocessing stage (tokenization, sentence splitting, text normalization) and lexical analysis are performed by LexMan [59] (ii) followed by RuDriCo [13], which applies manually built disambiguation rules, handles contractions, and several special types of compound words, (iii) the MARv module [50]) then performs POS-disambiguation, using HMM and the Viterbi algorithm, and, finally, (iv) the XIP rule-based parser (Xerox Incremental Parser) [2] segments sentences into chunks (or elementary sentence constituents: NP, PP, etc.) and extracts dependency relations among the chunks' heads (SUBJect, MODifier, etc.). XIP also performs named entity recognition (NER) [53, 33, 54, 42]. A set of post-parser modules has also been developed to handle certain NLP tasks such as anaphora resolution [35], temporal expressions' normalization [39], and slot-filling [11].

As part of the parsing process, XIP executes *dependency rules*. Dependency rules extract different types of dependencies between the nodes of the sentence's chunking tree, namely, the chunk heads. Dependencies can thus be viewed as equivalent to (or representing) the syntactic relations holding between different elements in a sentence. Some of the dependencies extracted by XIP represent rather complex relations, such as the notion of *subject* (SUBJ) or *direct object* (CDIR), which imply a higher level of analysis of a given sentence. Other dependencies are much simpler and sometimes quite straightforward, like the determinative dependency DETD, holding between an article and the noun it determines, e.g., *o livro* 'the book' - DETD(livro,o) 'DETD(book,the)'. Some dependencies can also be seen as auxiliary dependencies, they are required to build the more complex ones. The next rule extracts a syntactic dependency PREPD between the preposition introducing a prepositional phrase (more precisely, a prepositional chunk PP) and its head, as in the relation between *em* 'in' and *João*, in sentence 1:

1. *O Pedro confia em_o João*[11]
   (lit: Pedro trusts in_the João)
   'Pedro trusts João'

```
|PP#1{prep#2,?*,#3}|
if (HEAD(#3,#1))
    PREPD(#3,#2)
```

A dependency rule is composed of three parts: *structural conditions*, *dependency conditions*, and *actions*, which are performed in that order. The rule above, thus, reads as follows:

— Firstly, the structural conditions state the context of application of the rule; this is defined between two pipe signs '|'; the first to delimit the left context and the second to define the right context of the matching string; in this context, the nodes/chunks already built, their part-of-speech, and any other relevant features can be expressed using regular expressions; in this case, a prepositional phrase PP is defined as variable #1, which must be constituted by an introducing preposition numbered as variable #2, a non-defined string of elements (eventually none) (?*), and a final variable #3;

[10] https://string.l2f.inesc-id.pt/

[11] In Portuguese, the preposition is often contracted with the article, so the correct form would be *O Pedro confia no João*. The contraction was solved in this example, for clarity purposes.

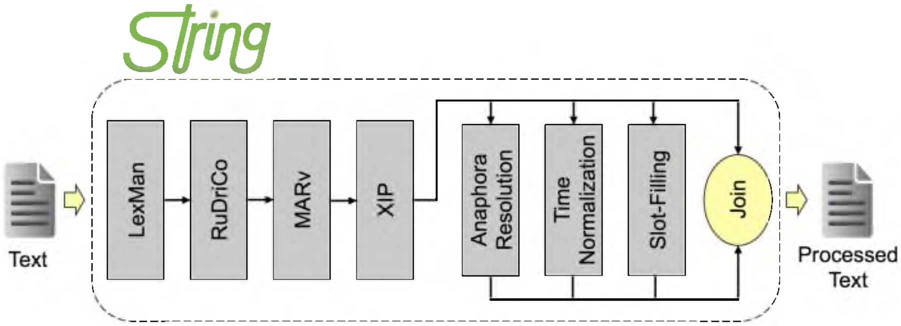

Fig. 1. STRING Architecture (from [34]).

— Secondly, the dependency conditions express the set of dependencies that must have been already extracted (or, on the contrary, should not have been extracted); if these conditions are verified, the rule is fired; in this case, a condition is defined that a HEAD dependency must exist between the PP chunk and the variable #3; notice that the HEAD dependency had already been built in a previous stage of parsing, when the chunking module determines this elementary constituent: the formal definition of a PP chunk is, in fact, a phrase introduced by a preposition and ending in a noun; the HEAD dependency is then extracted between the PP chunk and that noun;

— Thirdly, the actions are defined, that is, which dependencies are to be extracted and/or modified; in this case, the PREPD dependency is extracted, linking the preposition and the head of the PP.

3.2 Meronymy Extraction Module in STRING

Next, we describe the way some of whole-part dependencies involving *Nbp* are extracted in the Portuguese grammar for the XIP. To this end, a new module of the rule-based grammar was built, which is the first step towards a meronymy extraction module for Portuguese, and which contains most of the rules required for this work.

Occasionally, other parts of the grammar and some files in the lexicons had to be adapted, as new features had to be defined or new lexical entries were required, or new features had to be added to some already existing entries. For example, it was necessary to provide extensive lists of *Nbp*, including compounds, which had not yet been introduced in the lexicon. More rarely, other dependency rule files were slightly adapted to accommodate the new meronymy module.

In order to better present the different syntactic-semantic situations that the meronymy extraction module targets, some of the more simple cases are illustrated first and then some of the more complex situations follow. Example 2 is a simple case where there is a determinative PP complement *de N* 'of N' depending on the *Nbp*, so that the meronymy is overtly expressed in the sentence:

2. *O Pedro partiu o braço do João*
   'Pedro broke the arm of João'

The rule that captures the meronymy relation between *João* and *braço* 'arm' is

```
IF (MOD[POST](#2[UMB-Anatomical-human],
    #1[human]) &
    PREPD(#1,?[lemma:de]) &
    CDIR[POST](#3,#2) &
    ~WHOLE-PART(#1,#2)
   )
    WHOLE-PART(#1,#2)
```

The rule reads as follows: first, the parser determines the existence of a [MOD]ifier dependency, already calculated, between an *Nbp* (variable #2) and a human noun (variable #1); notice that, according to XIP conventions, the governor of the dependency is its first argument, hence *João* is said to be a modifier of *braço* 'arm'; this modifier must also be introduced by the preposition *de* 'of' which is expressed by the dependency PREPD; then, a constraint is defined that the *Nbp* must be a direct object (CDIR) of a given verb (variable #3); and, finally, that there is still no previously calculated WHOLE-PART dependency between the *Nbp* and the human noun (variable #1); this last constraint is meant to ensure that there is only one meronymy relation between each *Nbp* and a given noun; if all these conditions are met, then, the parser builds the WHOLE-PART relation between the human determinative complement and the *Nbp*.

The next example 3 demonstrates the case of sentences with an *Nbp* as a direct object and a dative complement *a Nhum* 'to Nhum', which is the "owner" of that *Nbp*.[12]

3. *O Pedro partiu o braço ao João*
'Pedro broke the arm to João'

The meronymy relation between *João* and *braço* 'arm' is captured by the following rule:

```
IF (^MOD[POST](#3,#1[human]) &
    PREPD(#1,?[lemma:a]) &
    CDIR[POST](#3,
    #2[UMB-Anatomical-human]) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
   )
    CINDIR(#3,#1),
    WHOLE-PART(#1,#2)
```

This is how the rule should read: first, the default MOD[ifier] dependency between the verb and the prepositional phrase *a Nhum* 'to Nhum' has to be changed (and it is, thus, preceded by the symbol '^') into an indirect complement (CINDIR) dependency; to do this, the system verifies if there is a syntactic relation between the preposition and the head noun of this PP, which is expressed by the dependency PREPD; then, the system checks if the *Nbp* is the direct object (CDIR) of a given verb (variable #3); and, lastly, if there is still no previously calculated CINDIR and WHOLE-PART dependencies; in this case, the parser builds a CINDIR dependency between the verb and the *Nhum* and a WHOLE-PART dependency between the *Nhum* and the *Nbp*. The dative complement can also be reduced to a dative *clitic* pronoun (v.g., *lhe* 'to him or her') and this can sometimes be fronted to a preverbal position. Specific rules were built in order to capture these situations.

Next, we describe the cases that involve determinative possessive pronouns. Even though these pronouns have their source in a *de N* 'of N' determinative complement, they are captured not as independent chunks but as determinants (dependency POSS) of the NP head noun. Furthermore, in Portuguese, possessives agree in gender and number with the noun they determine and not with their antecedent (as in English), e.g.:
o teu braço 'your_2nd-sg.masc.sg. arm_masc.sg.'
a tua mão 'your_2nd-sg.fem.sg. hand_fem.sg.'
os teus braços 'your_2nd-sg.masc.pl. arm_masc.pl.'
as tuas mãos 'your_2nd-sg.fem.pl. hand_fem.pl.'
and in the case of third-person possessive pronouns (v.g., *seu* 'his', *sua* 'her', *seus* 'their', *suas* 'their'), the pronoun can refer both to a singular or plural antecedent:

4. *O Pedro partiu o seu braço*
'Pedro broke his arm'

In this case, there is a [POSS]essive dependency between an *Nbp* and the possessive pronoun, so, if no WHOLE-PART relation had yet been extracted for that *Nbp*, then the parser establishes this dependency with the possessive. Notice that the meronymy module does not solve the anaphoric relation the possessive pronoun entails. This task is performed by a subsequent module of STRING.

Next, in example 5, we present the (apparently) more simple case of a sentence with just a human subject and an *Nbp* direct object:

5. *O Pedro partiu um braço*
'Pedro broke an arm'

In Portuguese, in the absence of a determinative complement, a possessive determiner, or a dative complement (eventually reduced to a clitic dative pronoun), sentences like 5 are preferably interpreted as holding a whole-part relation between

[12] Syntactically, this dative complement can be analyzed as the result from the dative restructuring ([32, 3]) of the *Nbp de Nhum* base phrase.

the human subject and the object *Nbp*. Thus, if there is a subject and a direct complement dependencies holding between a verb and a human, on one side, and the verb and an *Nbp*, respectively; and if no WHOLE-PART dependency has yet been extracted for that *Nbp*, either for that human subject or another element in the same sentence, then the WHOLE-PART dependency is extracted. The corresponding rules were built, but due to lack of space they are not presented in this paper.

There may be a relation within the same sentence between different *Nbp*, like in example 6. In this case, the WHOLE-PART relation should be established not only between the subject of the sentence and the *Nbp*, but also between the various *Nbp*s in the sentence.

6. *A Ana pinta as unhas dos pés*
(lit: Ana paints the nails of the feet)
'Ana paints the toenails'

In example 6, there is a meronymic relation between *Ana* and *unhas* 'nails', but also between *pés* 'feet' and *unhas* 'nails', so that two WHOLE-PART relations should be extracted.

There may be a relation within the same sentence between an *Nbp* and a noun that designates a part of that same *Nbp*, and which we will call *npart* (e.g., *ponta da língua* 'tip of the tongue', *costas das mãos* 'back of the hands', *palma da mão* 'palm', *canto do olho* 'canthus', *asa do nariz* 'nostrils', *lóbulo da orelha* 'ear lobe', etc.). This case differs from the previous one because, on the one hand, the whole-part relation should be established between the human noun and the *Nbp* and **not** the *npart* that precedes it; and, on the other hand, a second whole-part relation should also be established between the determinative *npart* and the *Nbp*, although this *npart* is not, by itself, an *Nbp*. Example 7 and the set of dependencies below illustrate this situation:

7. *O Pedro tocou com a ponta da língua no gelado da Ana*
'Pedro touched with the tip of the tongue the ice cream of Ana'

WHOLE-PART(Pedro,língua)
'WHOLE-PART(Pedro,tongue)' - correct;
WHOLE-PART(língua,ponta)
'WHOLE-PART(tongue,típ)' - correct;
WHOLE-PART(Pedro,ponta)
'WHOLE-PART(Pedro,típ)' - incorrect.

The set of *npart* varies according to the *Nbp*, and each set has to be established *a priori*. For example, for the *Nbp pé* 'foot' we can include the nouns *peito* 'instep', *alto* 'top', *cova or arco* 'arch', *dorso* 'instep', *planta* 'sole', and *ponta* 'tiptoe'. This is done by way of rules that add the feature *npart* to the nouns in the set associated to each *Nbp*, in the context of a determinative complement *de N* 'of N' of that *Nbp*. So far, 54 rules were built to associate the *Nbp* with their parts. All in all, 27 general rules have been built and implemented in STRING in order to extract whole-part relations involving *Nbp*.

We now turn to another type of meronymic relation. In some cases, a whole-part relation is implicit, and though *Nbp* are involved, they are not mentioned directly (e.g., *gastrite-estômago* 'gastritis-stomach'). In these cases, we decided that a whole-part relation between the human entity and the 'hidden' *Nbp* should be established. This was done to account for the interpretation of sentences such as *O Pedro tem uma gastrite e por isso não bebe café, que dizem que faz mal ao estômago.* 'Pedro has a gastritis and therefore [he] does not drink coffee, as they say it is bad for the stomach', where the presence of the *Nbp estômago* 'stomach' only makes sense because it is implied, somehow, by the disease noun *gastrite* 'gastritis'. These type of entailment relations contribute to the cohesion of texts, and the (hidden) meronymy here involved certainly plays an important role.

At this time, we focus on predicative nouns designating *diseases* (*Nsick*). High lexical constraints apply in this relation: for each disease predicative noun, the specific *Nbp* that is involved must be explicitly indicated in the lexicon. In order to adequately parse these constructions, we also distinguish three different sentence types. The first type is the case where a disease noun is built with the support verb *ter* 'have', example 8:

8. *O Pedro tem uma gastrite*
'Pedro has gastritis'

The rule that captures the meronymy relation between *Pedro* and *estômago* 'stomach' is given below:

```
IF (CDIR[POST](#1[lemma:ter],
    #2[lemma:gastrite]) &
    SUBJ(#1,#3) &
    ~WHOLE-PART(#3,?)
   )
    WHOLE-PART[hidden=+]
    (#3,##noun#[surface:estômago,
    lemma:estômago])
```

The rule reads as follows: first, the system checks if the disease noun (in this case, *gastrite* 'gastritis') is the direct object (CDIR) of the verb *ter* 'have' (variable #1); secondly, the system verifies if there is an explicit subject (variable #3) for the verb; and if there is still no WHOLE-PART relation between that subject and the other node; in this case, the system builds the WHOLE-PART dependency between the subject of the verb and the "hidden" *Nbp*, for which it creates a new (dummy) noun node. To express that a "hidden" noun is involved in this relation, a special tag "hidden" is also introduced in the dependency.

The next type of sentences (example 9) involves the support verb *estar com* 'be with' (more punctual aspect than *ter* 'have'):

9. *O Pedro está com uma gastrite*
(lit: Pedro is with a gastritis)
'Pedro has gastritis'

While the overall linguistic situation is similar to the case above, here different dependencies are extracted upon which the WHOLE-PART relation is to be built: the disease noun is normally parsed as a [MOD]ifier of *estar* 'to be' and there is a preposition *com* 'with' introducing it. Finally, many support verb and predicative noun constructions can be reduced to complex NPs, where the predicative noun is the head of the NP and its subject becomes a determinative *de N* 'of N' complement (eventually followed by any other complement of the predicative noun), as in sentence 10.

10. *A gastrite do Pedro é grave*
'Pedro's gastritis is severe'

So far, 29 different pairs (*disease nouns*, *Nbp*) have been encoded in the lexicon, with 3 rules for each pair. In the future, the previous identification of support verb constructions (currently, a work in progress) will allow for a significant simplification of these relations.

To conclude this section, we have also addressed the issue of ambiguity raised by idioms involving *Nbp*. As it is well known, there are many frozen sentences (or idioms) that include *Nbp* as one of their elements. However, for the overall meaning of these expressions, the whole-part relation is often irrelevant, as in the following example:

11. *O Pedro perdeu a cabeça*
(lit: Pedro lost the [=his] head)
'Pedro got mad'

The overall meaning of this expression has nothing to do with the *Nbp*, so that, even though we may consider a whole-part relation between *Pedro* and *cabeça* 'head', this has no bearing on the semantic representation of the sentence, equivalent in 11 to "get mad". The STRING strategy to deal with this situation is, first, to capture frozen or fixed sentences, and then, after building all whole-part dependencies, exclude/remove only those containing elements that were also involved in fixed sentences' dependencies. In this way, two general modules, for fixed sentences and whole-part relations, can be independently built, while a simple "cleaning" rule removes the cases where meronymy relation is irrelevant.

Frozen sentences are initially parsed as any ordinary sentence, and then the idiomatic expression is captured by a special dependency (FIXED), which takes as its arguments the main lexical items of the idiom. The number of arguments varies according to the type of idiom. In example 11 above, this corresponds to the dependency: FIXED(perdeu,cabeça) 'FIXED(lost,head)', which is captured by the following rule:

```
IF (VDOMAIN(?,#2[lemma:perder]) &
    CDIR[post](#2,#3[surface:cabeça])
   )
    FIXED(#2,#3)
```

This rule captures any VDOMAIN, that is, a verbal chain of auxiliaries and the main verb whose lemma is *perder* 'lose', and a post-positioned direct complement whose head is the surface form *cabeça* 'head'.

Next, the rules that exclude WHOLE-PART relation come into play. In case there are both a FIXED dependency and WHOLE-PART relation, a rule like the one shown below removes the latter, that is, it considers the sentence to be idiomatic and the meronymy to be irrelevant for the sentence's overall meaning.

```
IF (FIXED(#1,?,?,?,?,?,#2) &
    ^WHOLE-PART(#3,#4) &
    (#3::#1 || #3::#2 || #4::#1 ||
     #4::#2 ||
       ((#3 > #1) & (#3 < #2)) ||
        ((#4 > #1) &
        (#4 < #2))  ))~
```

In order to better understand the formalism here adopted, consider an apparently more complex example 12 of idiom:

12. *O Pedro anda com a cabeça à razão de juros*
'Pedro has a lot on his mind/getting mad with so many problems'

The rule that captures provisorily this idiom construes the FIXED dependency with 7 arguments: FIXED(anda,com,cabeça,a,razão,de,juros), while another rule also captures the WHOLE-PART dependency between the subject and the *Nbp* *cabeça* 'head': WHOLE-PART(Pedro,cabeça) 'WHOLE-PART(Pedro,head)'. This is when the "cleaning" rule above takes place. It, first, verifies if both FIXED and WHOLE-PART dependencies are present and signals the latter ('^') to be removed (1st line); then it checks if they have common arguments (2nd line), comparing the corresponding nodes, in this case, the nodes #3 and #4 against #2 (and also against #1, though so far no idiom has been considered where the first argument is not a verb). This part of the rule captures all cases where an argument of the whole-part relation is also involved in the fixed dependency. Finally (3rd line), the rule verifies whether any of the nodes of the WHOLE-PART relation are between the first and the last node of the FIXED expression. The conditions of the 2nd and the 3rd lines are in disjunction: if at least one of the conditions matches, the rule fires and removes the WHOLE-PART dependency. Thus, considering example 12 and the corresponding (provisory) dependencies above, the 1st line conditions are matched, but none of the 2nd line; nevertheless, as the condition ((#4 > #1) & (#4 < #2)) is matched, that is, the noun *cabeça* 'head' is between the first and the last argument of the FIXED dependency, then the meronymy is removed.

In the case of idioms that involve *Nbp*, like example 13, it has been noticed that these frozen sentences never allow determinative complements of the frozen head nouns, or the meaning of the sentence becomes literal, example 14 (which is signaled by '°'), as in example 14 below:

13. *O Pedro partiu a cara ao João*
(lit: Pedro broke the face to João)
'Pedro hit João' (not necessarily in the face)

14. °*O Pedro partiu o lado direito da cara ao João*
'Pedro broke the right side of the face to João'

In order to deal with this condition, a specific "cleaning" rule was introduced at the end of the fixed sentences module:

```
IF (^FIXED(#1,#2) &
    MOD(#2,#3[npart])) ~
```

This rule acts before the meronymy module and removes the FIXED dependency whenever a *npart* is involved. Thus, after this rule, instead of getting the incorrect output:

```
FIXED(partiu,cara)
'FIXED(broke,face)'
```

that would preclude the meronymy rules to be triggered, only the correct dependencies are extracted:

```
WHOLE-PART(João,cara)
'WHOLE-PART(João,face)'
```

and

```
WHOLE-PART(cara,lado)
'WHOLE-PART(face,side)'.
```

Similar rules were necessary for `FIXED` dependencies with 3 or more arguments.

In order to capture the idioms involving *Nbp*, we built about 400 of such rules, from 10 formal classes of idioms [6].

4 Evaluation

4.1 Evaluation Corpus

The 1st fragment of the CETEMPúblico corpus [52] was used in order to extract sentences that involve *Nbp*. This fragment of the corpus contains 14,715,055 tokens (147,567 types), 6,256,032 (147,511 different) simple words and 260,943 sentences. The existing STRING lexicons of *Nbp* and *Nsick* were adapted to the DELA format to be used within the UNITEX corpus processor [46, 47] along with the remaining available resources for European Portuguese, distributed with the system.

Using the *Nbp* (151 lemmas) and the *Nsick* (29 lemmas) dictionaries, 16,746 *Nbp* and 79 *Nsick* instances were extracted from the corpus (excluding the ambiguous noun *pelo* 'hair' or 'by-the', which did not appeared as an *Nbp* in this fragment). Some of these sentences were then excluded, for they consist of incomplete utterances or include more than one *Nbp* per sentence. A certain number of particularly ambiguous *Nbp*; e.g., *arcada* 'arcade', *articulação* 'articulation', *lobo* 'lobe', *médio* 'middle', *membro* 'part', *membro superior* 'upper limb', *miúdos* 'kids', *órbita* 'orbit', *órgão* 'organ', *rádio* 'radius', *raiz* 'root', *tecido* 'tissue', and *temporal* 'temporal' that showed little or no occurrence at all in the *Nbp* sense were discarded from the extracted sentences. Finally, the sentences that lacked a full stop were corrected, in order to prevent errors from STRING's sentence splitting module. In the end, a set of 12,659 sentences with *Nbp* was retained for evaluation.

Based on the distribution of the remaining 103 *Nbp*, a random stratified sample of 1,000 sentences was selected, keeping the proportion of their total frequency in the corpus. This sample also includes a small number of disease nouns (6 lemmas, 17 sentences).

4.2 Inter-annotator Agreement

The output sentences were then divided into 4 subsets of 225 sentences each. Each subset was then given to a different, native-speaker annotator, and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement. From the 100 sentences that were annotated by all the participants in this process, we calculated the Average Pairwise Percent Agreement, the Fleiss' Kappa [16], and the Cohen's Kappa coefficient of inter-annotator agreement [12] using ReCal3: Reliability Calculator [17], for 3 or more annotators.[13] The four annotators achieved the following results. First, the Average Pairwise Percent Agreement, that is, the percentage of cases each pair of annotators agreed with each other is shown in Table 1.

The Average Pairwise Percent Agreement is 85.031%, which is relatively high. The best agreement is shown by the pair of annotators 1 and 3 (90.741%).

Next, the Fleiss' Kappa inter-annotator agreement coefficient is shown in Table 2. Fleiss' Kappa[14]

> "works for any number of raters giving categorical ratings [. . .], to a fixed number of items. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly."

[13] http://dfreelon.org/utils/recalfront/recal3/

[14] http://en.wikipedia.org/wiki/Fleiss'_kappa

Table 1. Average Pairwise Percent Agreement

Average pairwise pct. agr.	Pairwise pct. agr. annot. 1 & 4	Pairwise pct. agr. annot. 1 & 3	Pairwise pct. agr. annot. 1 & 2	Pairwise pct. agr. annot. 2 & 4	Pairwise pct. agr. annot. 2 & 3	Pairwise pct. agr. annot. 3 & 4
85.031%	86.111%	**90.741%**	82.407%	81.481%	80.556%	88.889%

Table 2. Fleiss' Kappa

Fleiss' Kappa	Observed Agreement	Expected Agreement
0.625	0.85	0.601

In our case, Fleiss' Kappa equals 0.625 and indicates that observed agreement of 0.85 is higher than expected agreement of 0.601.

Finally, the Average Pairwise Cohen's Kappa (CK) is shown in Table 3.

The Average Pairwise Cohen's Kappa is 0.629. Again, the pair of annotators 1 and 3 achieved the best Cohen's Kappa coefficient of 0.757. According to Landis and Koch [31] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [15], these results correspond to an inter-annotator agreement halfway between "fair" and "good".

Then, for these overlapping 100 annotated sentences, another linguist manually checked the few instances of disagreement among annotators and decided a unique, definitive tag.

In view of these results, we can reasonably assume that, for the remaining independently annotated and non-overlapping 900 sentences of the corpus, the annotation is sufficiently consistent and could be used for the evaluation of the system output.

4.3 Evaluation of the System's Overall Performance

The system performance was evaluated adopting the usual evaluation metrics of precision, recall, F-measure, and accuracy. The results are shown in Table 4, where TP=*true positives*; TN=*true negatives*; FP=*false positives*; FN=*false negatives*; the first line corresponds to the 100 sentences that were subject to multiple annotators' classification, while the 900 sentences are the remainder instances of the sample taken form the corpus.

The number of instances (TP, TN, FP, and FN) is higher than the number of sentences, as one sentence may involve several instances of whole-part relations. The relative percentages of the TP, TN, FP, and FN instances are similar between the sets with 100 and 900 sentences. This explains the similarity of the evaluation results and seems to confirm our decision to use with enough confidence the set of the remaining 900 sentences as a golden standard for the evaluation of the system output. The recall is relatively small, which can be explained by the fact that in many sentences the *whole* and the *part* are too far away from each other, and too many elements are intervening between the human noun and the target *Nbp*. The precision is somewhat better. The accuracy is relatively high for the same reason that there is a great number of *true negatives*, which, as it was mentioned before, occur because in many cases there is no whole-part relation to be extracted, even though there is an *Nbp* in the sentence.

The results for *Nsick*, though the number of instances is small, show precision of 0.5, recall of 0.11, F-measure of 0.17, and accuracy of 0.76.

5 Error Analysis

The results of the evaluation of the task showed that there were 62 false positive cases and 132 false negatives. We begin this section by the analysis of some false positives cases and then move on to the false negatives.

Table 3. Average Pairwise Cohen's Kappa (CK)

Average pairwise CK	Pairwise CK annot. 1 & 4	Pairwise CK annot. 1 & 3	Pairwise CK annot. 1 & 2	Pairwise CK annot. 2 & 4	Pairwise CK annot. 2 & 3	Pairwise CK annot. 3 & 4
0.629	0.65	**0.757**	0.59	0.558	0.518	0.699

Table 4. System's performance for *Nbp*

Number of sentences	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
100	8	73	7	14	0.53	0.36	0.43	0.79
900	73.5	673	55	118	0.57	0.38	0.46	0.81
Total:	81.5	746	62	132	0.57	0.38	0.46	0.81

5.1 False Positives

5.1.1 Disambiguation of *Nbp* in Context

To begin with, we tackled a number of cases with the ambiguous noun *língua* 'tongue/language'. In order to preclude the building of whole-part relation in cases such as *língua portuguesa* 'Portuguese language', *a língua de Camões* 'the language of Camões', *professor de língua* (lit: teacher of language) 'language teacher', etc., where the noun *língua* 'language' is not used in the sense of an anatomical part, we adopted one of the following strategies: we removed the *Nbp* (`sem-anmov`) feature from the noun's lexical set of features. This is carried out by the following rules, which are applied before the chunking stage:

— In the case of gentilic adjectives, one rule had to be done for each one of this type of adjectives:

```
2> noun[lemma:língua,sem-anmov=~],
adj[gentcontinent=+].
2> noun[lemma:língua,sem-anmov=~],
adj[gentregion=+].
2> noun[lemma:língua,sem-anmov=~],
adj[gentcountry=+].
2> noun[lemma:língua,sem-anmov=~],
adj[gentcity=+].
```

— In the case of combinations of *língua* 'tongue/language' with renowned authors of a given language, a PP structure has to be spelled out; so far, we built rules for over a dozen authors, epitomes of their national languages, which occurred with some frequency in the CETEMPúblico corpus:

```
2> noun[lemma:língua,sem-anmov=~],
prep[lemma:de],
noun[lemma:Camões].
```

e.g., *língua de Camões*.

```
2> noun[lemma:língua,sem-anmov=~],
prep[lemma:de],
noun[lemma:Shakespeare].
```

e.g., *língua de Shakespeare*.

— A similar rule is necessary for PP complements with country names (*a língua de Portugal* 'Portugal's language'):

```
2> noun[lemma:língua,sem-anmov=~],
prep[lemma:de],
noun[country=+].
```

Besides, there could also be a determiner for such examples as *a língua do Brasil é o Português* (lit: the language of the Brazil is the Portuguese) Thus, a second rule is necessary:

```
2> noun[lemma:língua,sem-anmov=~],
prep[lemma:de],
art[lemma:o], noun[country=+].
```

This rule is required because some country names are obligatorily preceded by a definite article (*o Brasil* 'the Brazil', *os Estados Unidos* 'the United States', etc.)

5.1.2 Difficult Cases

A certain number of cases were found where the use of the *Nbp* is clearly figurative, but it is neither an idiom nor a compound word, so we were unable to devise any strategy to avoid capturing the whole-part relation:

15. *À farta ementa associou-se um acontecimento a que certamente não foi alheio o* ***dedo*** *organizativo de José Perdigão, que no filho encontrou precioso instrumento...*
'To the abundant menu, an event was associated, which was certainly not unconnected with the organizational finger of José Perdigão, who found in [his] son a [precious=] most valuable tool...'

```
WHOLE-PART(José Perdigão,dedo)
'WHOLE-PART(José Perdigão,finger)'
```

In this case, the whole-part relation is correctly extracted, but the *Nbp dedo* 'finger' is not to be interpreted literally but figuratively, and can be connoted with other idioms such as *meter o dedo/a mão em* 'sb. put [one's] finger/hand in sth.' 'to have a role in / to interfere with'.

5.2 False Negatives

5.2.1 Noun or NP Modifiers (not involving verbs)

The rules that have been developed only involve verb arguments (subject or complements) and did not consider the situations where an *Nbp* is a modifier of a noun or an adjective. Therefore, in several situations, the whole-part relations have not been captured. For example:

16. `133>TOP{NP{Um mágico} PP{de carapuço} PP{em a cabeça} .}`
'A magician with a hood over the head'

In this case, there is only a complex NP, with all the PP depending on the head noun *mágico* 'magician'. It is also possible to consider that in these cases an adjective or a verb past participle has been zeroed (e.g., *Um mágico de carapuço enfiado/posto/colocado na cabeça* 'A magician with a hood stucked/placed over the head'). The meronymy module did not contemplate these complex NPs, including those with a zeroed adjective or past-participle, as most of the rules always involved a verb argument. This will have to be taken into consideration in future work.

5.2.2 Missing Features

One of the main reasons why the whole-part relation has not been captured derived from the fact that many human nouns are still unmarked with the human feature (or any of its subsumed features). For example, in the sentence:

17. *Numa espécie de altar, um transexual padece com uma coroa de agulhas espetadas na cabeça, apoiado a umas muletas, provavelmente a sua cruz, nesta paródia à crucificação*
'In a kind of altar, a transsexual suffers with a crown of needles stuck in his head, supported by crutches, probably his cross, in this parody of the crucifixion'

In this case, the whole-part relation between the subject of *padecer* 'suffer' and the body-part *cabeça* 'head' was not captured just because the noun *transexual* (*id.*) had not been attributed the feature `human`.

In some cases, the rules have not been triggered because the human entity is expressed by a personal pronoun and this category is not marked with the human feature. In the following sentence, the system also failed to establish the whole-part relation because it cannot ascribe the human feature to the relative pronoun *que* 'who' that is the subject of the relative clause:

18. *Segundo o responsável do hospital, o doente –* ***que*** *também sofreu graves ferimentos na cabeça – poderia ser ainda sujeito a uma segunda intervenção cirúrgica*
'According to the head of the hospital, the patient - **who** also suffered serious head injuries - could still be subjected to a second surgical intervention'

However, the antecedent of the pronoun has been correctly extracted:

```
ANTECEDENT_RELAT(doente,que)
'ANTECEDENT_RELAT(patient,who)'
```

According to [35], relative pronouns are among the most successful cases of anaphora resolution in STRING. Therefore, it is possible that after this module comes into play, the features of the antecedent are inherited by the pronoun and the whole-part module be allowed to process the sentence again.

An opposite situation occurs when some features associated to the *Nbp* preclude the correct extraction of the whole-part dependency. *Corpo* 'body' is one of that cases and a very complex one. It is an element of several compound nouns, which are identified during lexical analysis and do not interfere in the dependency extraction step. Furthermore, it can be an *Nbp* and also a collective noun functioning as a type of determiner as in

19. *O corpo (=conjunto) dos docentes da faculdade*
'The staff of the (= set) of the teachers of the faculty'

Because of this a `QUANTD` (quantifying) dependency is extracted between *corpo* 'body' and the immediately following `PP`, which prevents the extraction of whole-part relation; therefore, rules were build to partially disambiguate this particular noun by removing the features associated to its collective noun interpretation:

```
3> noun[lemma:corpo,sem-anmov=+,
sem-sign=~,sem-cc=~,sem-ac=~,sem-hh=~,
sem-group-of-things=~],prep[lemma:de],
(art[lemma:o]),noun[lastname=+].
3> noun[lemma:corpo,sem-anmov=+,
sem-sign=~,sem-cc=~,sem-ac=~,sem-hh=~,
sem-group-of-things=~],prep[lemma:de],
(art[lemma:o]),noun[firstname=+].
```

These rules read as follows: if the noun *corpo* 'body' is followed by the preposition *de* 'of' and a first or a last proper name, then we remove all the other features of *corpo* 'body' except the one that marks it as an *Nbp* and a concrete-countable (`sem-cc`). The rules do not solve all the cases, naturally, since the distinction between the determiner and the *Nbp* cannot yet be detected as it would require a previous word sense disambiguation module.

5.2.3 Ambiguous `FIXED` Expressions, Incorrectly Captured

In some cases, the `FIXED` expressions have been incorrectly captured as whole-part relations because they are ambiguous and have been used in the literal sense. For example:

20. *Ele arrancava-me os cabelos todos* 'He pulled out all my hair'

```
FIXED(arrancava,cabelos)
FIXED(pulled out,hair)
```

In the idiom *arrancar os cabelos* (lit: to pluck the hair) 'to despair', there is obligatory co-reference between the subject and the *Nbp*, so there is no way the sentence could be interpreted figuratively. The problem, thus, relies in the incorrect representation of the constraints of the idiom, not of the grammar. In this case, the correct relation should be `WHOLE-PART(me,cabelos)` `'WHOLE-PART(my,hair)'`.

5.2.4 No Syntactic Relation Between Whole and Part

In some cases the *whole* and the *part* are not syntactically related (and can be far away from each other in a sentence):

21. *O facto do corpo ter sido encontrado na cozinha, leva os bombeiros a suspeitar que a vítima, com graves problemas de saúde, tenha desmaiado e caído ă lareira, o que poderă ter estado na origem do incêndio*
'The fact that the body was found in the kitchen, makes the firefighters to suspect that the victim, with serious health problems, had fainted and fallen into the hearth, which may have been the origin of the fire'

In this example, the *part corpo* 'body' is the subject of the *ter sido encontrado* 'have been found', while the *whole vítima* 'victim' is the subject of *tenha desmaiado* 'had fainted'; each noun is in a different subclause, and there is no syntactic dependency between the two nouns. However, the annotator was able to identify this meronymic relation `WHOLE-PART(vítima,corpo)` `'WHOLE-PART(victim,body)'`, which is beyond the scope of our current parser.

5.2.5 Difficult Cases

In spite of our best efforts, some *Nbp* were still missing from the lexicon, as in the case of *defesas imunitárias* 'immune defenses':

22. *O que se pensa que acontece na artrite reumatóide é que a cartilagem é atacada pelas defesas imunitárias do doente, como se ela fosse um autêntico "corpo estranho"*
'What we think happens in rheumatoid arthritis is that the cartilage is attacked by the immune defenses of the patient as if it was an authentic "foreign body"'

In such cases, we have completed the dictionary.

In the next example, there is also a problem with the compound noun *cabelo(s) branco(s)* 'white hair(s)':

23. *Um deles, de óculos e cabelo branco, olha para o relógio e depois perscruta com alguma inquietação as bancadas a meia nau*
'One of them, wearing glasses and with white hair, looks at his watch and then peers restlessly to the seats at midship'

For the moment, even though *cabelo(s) branco(s)* 'white hair(s)' is already tokenized as a compound noun, it has not been given the *Nbp* feature; therefore, the system did not capture any meronymic relation for this element. Even if the compound had been recognized, the problem would have remained in the missed apposition relation of the two PPs with the subject complex NP, whose head is a pronoun (namely, *um deles* 'one of them'). Since no dependency exists between the subject (*um* 'one') and the apposition, and also because the subject is a pronoun (see above), no feature is there to trigger the meronymy rules.

5.3 Evaluation after Error Analysis

Once all the corrections were taken into consideration, we ran the system again in order to carry out the second evaluation of its performance. The results are shown in Table 5.

The precision was improved by 0.13 (from 0.57 to 0.70), the recall by 0.11 (from 0.38 to 0.49), the F-measure by 0.12 (from 0.46 to 0.58), and the accuracy by 0.04 (from 0.81 to 0.85). The results for *Nsick* remained the same (so we do not repeat them here). Since only some of the detected errors were corrected at this stage, and some can still be improved by extending the current work to so far unaddressed situations (dependencies on nouns, anaphora resolution, to name a few), it is expectable that higher levels of performance will be achieved in future work.

6 Conclusions and Future Work

6.1 Conclusions

This work addressed the problem of extraction of whole-part relations (*meronymy*), that is, a semantic relation between two entities of which one is perceived as a constituent part of the other, or

Table 5. Post-error analysis system's performance for *Nbp*

Number of sentences	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
100	10	75	4	12	0.71	0.45	0.56	0.84
900	90	688	39	91	0.70	0.50	0.58	0.86
Total:	100	763	43	103	0.70	0.49	0.58	0.85

between a set and its member. As a type of semantic relations, whole-part relations contribute to cohesion and coherence of a text and can be useful in several Natural Language Processing (*NLP*) tasks such as question answering, text summarization, machine translation, information extraction, information retrieval, anaphora resolution, semantic role labeling, among others. This work targeted a special type of whole-part relations that involve human entities and *body-part nouns* (*Nbp*) in Portuguese. To extract whole-part relations, a new module of the rule-based grammar was built and integrated in STRING, a hybrid statistical and rule-based NLP chain for Portuguese [34].

An overview of related work has been done, paying a particular attention to whole-part relations extraction in Portuguese. Two well-known parsers of Portuguese were reviewed in order to discern how they handled the whole-part relations extraction: the PALAVRAS parser [8], consulted using the Visual Interactive Syntax Learning (VISL) environment, and LX Semantic Role Labeller [9]. Judging from the available on-line versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly. Furthermore, according to our review of the related work and to a recent review of the literature on semantic relations extraction [1], no other references to whole-part relation extraction for Portuguese have been identified. This paper can be seen at a first approach to this complex issue.

In order to extract whole-part relations, a rule-based meronymy extraction module has been built and integrated in the grammar of the STRING system. It contains 29 general rules addressing the most relevant syntactic constructions triggering this type of meronymic relations, and a set of 87 rules for the 29 *disease nouns* (*Nsick*), in order to capture the underlying *Nbp* (e.g., *gastrite-estômago* 'gastritis-stomach').

A set of around 400 rules has also been devised to prevent the whole-part relations being extracted in the case the *Nbp* are elements of idiomatic expressions (e.g., *O Pedro perdeu a cabeça* (lit: Pedro lost the [=his] head) 'Pedro got mad'). This work also addresses the cases where a whole-part relation holds between two *Nbp* in the same sentence (e.g., *A Ana pinta as unhas dos pés* (lit: Ana paints the nails of the feet) 'Ana paints her toes' nails') and the case of determinative nouns that designate parts of an *Nbp*, though they are not *Nbp* themselves (e.g., *O Pedro encostou a ponta da língua ao gelado da Ana* 'Pedro touched with the tip of the tongue the ice cream of Ana'). Each one of these cases triggers different sets of dependencies. 54 rules were built to associate the *Nbp* with their parts, to handle the cases where there is an *Nbp* and a noun that designates a part of that same *Nbp*.

For the evaluation of the work, the first fragment of the CETEMPúblico corpus [52] (14,7 million tokens and 6,25 million words) was used in order to extract sentences that involve *Nbp* and *Nsick*. Using the *Nbp* (151 lemmas) and the *Nsick* (29 lemmas) dictionaries, specifically built for the STRING lexicon, 16,746 *Nbp* and 79 *Nsick* instances were extracted from the corpus.

In order to produce a golden standard for the evaluation, a random stratified sample of 1,000 sentences was selected, keeping the proportion of the total frequency of *Nbp* in the source corpus. This sample also includes a small number of *Nsick* (6 lemmas, 17 sentences). The 1,000 output sentences were divided into 4 subsets of 225 sentences each. Each subset was then given to a different annotator (native Portuguese speaker), and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement. The annotators were asked to append the whole-part dependency, as it was previously defined in a set of guidelines, using the XIP format.

To assess inter-annotator agreement we used ReCal3: Reliability Calculator [17] for 3 or more annotators. The results showed that the Average Pairwise Percent Agreement equals 0.85, the Fleiss' Kappa inter-annotator agreement is 0.62, and the Average Pairwise Cohen's Kappa is 0.63. According to Landis and Koch [31] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [15], these results correspond to an inter-annotator agreement halfway between "fair" and "good". In view of these results, we assumed that the remaining independent and non-overlapping annotation of the corpus by the four annotators is sufficiently consistent and can be used as a golden standard for the evaluation of the system's output.

After confronting the produced golden standard against the system's output, the results for *Nbp* show precision of 0.57, recall of 0.38, F-measure of 0.46, and accuracy of 0.81. The recall is relatively small (0.38), which can be explained by the fact that in many sentences, the *whole* and the *part* are not syntactically related and are quite far away from each other; naturally, human annotators were able to overcome these difficulties. In some cases, the rules were not triggered because some human nouns and personal pronouns are unmarked with the human feature. Besides, as we focused on verb complements alone, the situations where an *Nbp* is a modifier of a noun or an adjective (and not a verb) have not been contemplated in this project, which produced a significant number of *false negatives*. Other, quantitatively less relevant, cases were also submitted to the detailed error analysis made after the system's first evaluation. The problem derived from pronouns (especially relative pronouns) not having the human feature raises the issue of the adequate placing of the meronymy module in the STRING pipeline architecture: if some part of this task could also be performed after anaphora resolution, it is likely that better results would be produced.

The precision of the task is somewhat better (0.57). The accuracy is relatively high (0.81) since there is a large number of *true negative* cases. The results for *Nsick*, though the number of instances is small, show a precision of 0.5, a recall of 0.11, an F-measure of 0.17, and an accuracy of 0.76.

A detailed error analysis was performed to determine the most relevant cases for these results, which led to implementation of some solutions. A second evaluation of the system's performance was carried out, with the same golden standard, and the results showed that the precision improved by 0.13 (from 0.57 to 0.70), the recall by 0.11 (from 0.38 to 0.49), the F-measure by 0.12 (from 0.46 to 0.58), and the accuracy by 0.04 (from 0.81 to 0.85). The results for *Nsick* remained the same. Because of the limited size of the sample (only 1,000 sentences) and the significant increase in the results on the evaluation of the meronymy module performance (around 0.12), it is likely that, by continuing this work, a higher performance threshold may be attained.

To conclude, this work can be considered as a first attempt to extract whole-part relations in Portuguese, in this case, involving human entities and *Nbp*. A rule-based module was built, integrated in the STRING system, and evaluated with promising results.

6.2 Future Work

In future work, we will address the extraction of other types of whole-part relations such as component-integral object (*pedal - bicycle*), member-collection (*player - team*), place-area (*grove - forest*), among others [60]. The intention is also to use the list of *Nbp* provided by Cláudia Freitas [18] in order to complete the existing *Nbp* lexicon in STRING (common vocabulary for vehicles, human collective nouns, place-botanic, place-human building, place-geographic, tools, plants, animals, etc.). However, it is not obvious if for some of these classes of objects the strategy used here can be adequate; eventually, other strategies must be adopted such as a machine learning approach that will capture words associated to the lexical classes in patterns that are prone to be interpreted in this way. We also intend to target the extraction of other types of whole-part relations using syntactic dependency-based n-grams – the concept is introduced in detail in [55, 56] – and other syntactic information, such as subcategorization frames [19, 20] within a machine learning approach.

Another line of future work will be the improvement of recall by focusing on the *false negative* cases already found, which have shown that several syntactic patterns have not been paid enough attention yet. Thus, the focus will shift to the situations where an *Nbp* is a modifier of a noun or an adjective (and not a verb): e.g., *Um mágico de carapuço (enfiado) na cabeça* 'A magician with a hood (stuck) over the head'. Furthermore, significant work will be required to complete the coverage of human nouns or, more precisely, to enrich the existing lexicon with the appropriate human feature, probably resorting to machine learning techniques. A more general (and more complex) issue is the tagging of personal pronouns with the features corresponding to their human antecedent, which will certainly improve the recall of the task. However, this raises the issue of the order of application of the anaphora resolution module and the meronymy module built here. Attention should also be paid to the idioms that correspond to support verb constructions (*dar uma/a mão a* 'give a hand to', *estar em as mãos de* 'to be in one's hands', and others) and the integration of this type of expressions in STRING in order to prevent the system of extracting whole-part relations in these cases.

Acknowledments

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013; Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institutions; and the Mexican Government (scholarship CONACYT).

References

1. **Abreu, S., Bonamigo, T., & Vieira, R.** (**2013**). A review on relation extraction with an eye on portuguese. *J. Braz. Comp. Soc.*, Vol. 19, No. 4, pp. 553–571.
2. **Ait-Mokhtar, S., Chanod, J., & Roux, C.** (**2002**). Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, Vol. 8, No. 2/3, pp. 121–144.
3. **Baptista, J.** (**1997**). Conversão, nomes parte-do-corpo e restruturação dativa. **Castro, I.**, editor, *Actas do XII Encontro da APL*, volume I – Linguística, pp. 51–59.
4. **Baptista, J., Cabarrão, V., & Mamede, N.** (**2012**). Classification directives for events and relations extraction between named entities in portuguese texts. Technical report, Instituto Superior Técnico, Universidade do Algarve.
5. **Baptista, J., Mamede, N., Hagège, C., & Maurício, A.** (**2012**). Time expressions in portuguese. guidelines for identification, classification and normalization. Technical report, Universidade do Algarve, Instituto Superior Técnico, Xerox Research Centre Europe.
6. **Baptista, J., Mamede, N., & Markov, I.** (**2014**). Integrating verbal idioms into an nlp system. **Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T., & Nunes, M. G.**, editors, *Computational Processing of Portuguese Language, PROPOR 2014*, volume 8775 of *LNAI/LNCS*, Springer, pp. 250–255.
7. **Berland, M. & Charniak, E.** (**1999**). Finding parts in very large corpora. *Proceedings of the 37th annual meeting of the ACL on Computational Linguistics*, Morristown, NJ, USA. ACL, pp. 57–64.
8. **Bick, E.** (**2000**). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus Univ. Aarhus, Denmark: Aarhus Univ. Press.
9. **Branco, A. & Costa, F.** (**2010**). A deep linguistic processing grammar for portuguese. **Pardo, T., Branco, A., Klautau, A., Vieira, R., & Lima, V.**, editors, *Computational Processing of Portuguese, PROPOR 2010*, LNAI/LNCS 6001, Springer, pp. 86–89.
10. **Cabrita, V., Baptista, J., & Mamede, N.** (**2013**). Diretivas de classificação e anotação de corpora para a extração de relações entre eventos. Technical report, Instituto Superior Técnico.
11. **Carapinha, F.** (**2013**). *Extração Automática de Conteúdos Documentais*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.
12. **Cohen, J.** (**1960**). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46.

13. **Diniz, C. F. P.** (**2010**). *Um Conversor baseado em regras de transformação declarativas*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

14. **Fellbaum, C.** (**1998**). *WordNet: An Electronic Lexical Database*. MIT, Cambridge.

15. **Fleiss, J.** (**1981**). *Statistical methods for rates and proportions*. New York: John Wiley, Heidelberg, second edition.

16. **Fleiss, J. L.** (**1971**). Measuring nominal scale agreement among many raters. *Psych. Bull.*, Vol. 76, No. 5, pp. 378–382.

17. **Freelon, D.** (**2010**). Recal: Intercoder Reliability Calculation as a Web Service. *Intl. J. of Internet Science*, Vol. 5, No. 1, pp. 20–33.

18. **Freitas, C.** (**20.05.2014**). Esqueleto - anotaÇão das palavras do corpo humano. Technical Report Versão 5: 20.05.2014.

19. **Gelbukh, A.** (**1999**). Syntactic disambiguation with weighted extended subcategorization frames. *Proceedings of PACLING-99, Pacific Association for Computational Linguistics*, University of Waterloo, Canada, pp. 244–249.

20. **Gelbukh, A.** (**2014**). Unsupervised learning for syntactic disambiguation. *Computación y Sistemas*, Vol. 18, No. 2, pp. 329–344.

21. **Gerstl, P. & Pribbenow, S.** (**1995**). Midwinters, end games, and body parts: a classification of part-whole relations. *Intl. J. of Human Computer Studies*, Vol. 43, pp. 865–890.

22. **Girju, R., Badulescu, A., & Moldovan, D.** (**2003**). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of HLT-NAACL*, volume 3, pp. 80–87.

23. **Girju, R., Badulescu, A., & Moldovan, D.** (**2006**). Automatic discovery of part-whole relations. *Computational Linguistics*, Vol. 21(1), pp. 83–135.

24. **Hage, W. V., Kolb, H., & Schreiber, G.** (**2006**). A method for learning part-whole relations. *The Semantic Web - ISWC 2006, LNAI/LNCS*, Vol. 4273, pp. 723–725.

25. **Hearst, M.** (**1992**). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conf. on Computational Linguistics*, volume 2 of *COLING 92*, ACL Morristown, NJ, USA, pp. 539–545.

26. **Hirst, G.** (**2004**). Ontology and the lexicon. **Staab, S. & Studer, R.**, editors, *Handbook on Ontologies*, Springer, pp. 209–230.

27. **Iris, M., Litowitz, B., & Evens, M.** (**1988**). Problems of the part-whole relation. **Evens, M.**, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, Cambridge Univ. Press, pp. 261–288.

28. **Ittoo, A. & Bouma, G.** (**2010**). On learning subtypes of the part-whole relation: Do not mix your seeds. *Proceedings of the 48th Annual Meeting of the ACL*, Univ. of Groningen, pp. 1328–1336.

29. **Keet, M. & Artale, A.** (**2008**). Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, Vol. 3(1), pp. 91–110.

30. **Khoo, C. & Na, J.-C.** (**2006**). Semantic relations in information science. *Annual Review of Information Science and Technology*, Vol. 40, pp. 157–229.

31. **Landis, J. & Koch, G.** (**1977**). The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174.

32. **Leclère, C.** (**1995**). Sur une restructuration dative. *Language Research*, Vol. 31-1, pp. 179–198.

33. **Loureiro, J.** (**2007**). *Reconhecimento de Entidades Mencionadas (Obra, Valor, Relações de Parentesco e Tempo) e Normalização de Expressões Temporais*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

34. **Mamede, N., Baptista, J., Diniz, C., & Cabarrão, V.** (**2012**). STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese. *Computational Processing of Portuguese, PROPOR 2012*.

35. **Marques, J.** (**2013**). *Anaphora Resolution*. Master's thesis, Univ. of Lisbon/IST and INESC-ID Lisboa/L2F.

36. **Marrafa, P.** (**2001**). *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto Camões.

37. **Marrafa, P.** (**2002**). Portuguese wordnet: general architecture and internal semantic relations. *DELTA*, Vol. 18, pp. 131–146.

38. **Marrafa, P., Amaro, R., & Mendes, S.** (**2011**). Wordnet.pt global – extending wordnet.pt to portuguese varieties. *Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, Edinburgh, Scotland. ACL Press, pp. 70–74.

39. **Maurício, A.** (**2011**). *Identificação, Classificação e Normalização de Expressões Temporais*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

40. **Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T.** (**1993**). A Semantic Concordance. *Proceedings of the ARPA Workshop on Human Language Technology*.

41. **Odell, J.** (**1994**). Six different kinds of composition. *J. of Object-Oriented Programming*, Vol. 5(8), pp. 10–15.

42. **Oliveira, D.** (**2010**). *Extraction and Classification of Named Entities*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

43. **Oliveira, H.** (**2012**). *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. Ph.D. thesis, Univ. of Coimbra/Faculty of Science and Technology.

44. **Oliveira, H., Gomes, P., Santos, D., & Seco, N.** (**2008**). Papel: A dictionary-based lexical ontology for portuguese. **Teixeira, A., Lima, V., Oliveira, L., & Quaresma, P.**, editors, *Computational Processing of the Portuguese Language, PROPOR 2008*, LNAI/LNCS 5190, Aveiro, Portugal. Springer, pp. 31–40.

45. **Pantel, P. & Pennacchiotti, M.** (**2006**). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. *Proceedings of Conf. on Computational Linguistics/ACL (COLING/ACL-06)*, Sydney, Australia, pp. 113–120.

46. **Paumier, S.** (**2000**). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée.

47. **Paumier, S.** (**2014**). *Unitex 3.1beta, User Manual*. Univ. Paris-Est Marne-la-Vallée.

48. **Pianta, E., Bentivogli, L., & Girardi, C.** (**2002**). Multiwordnet: developing an aligned multilingual database. *1st Intl. Conf. on Global WordNet*, Mysore, India, pp. 293–302.

49. **Prévot, L., Huang, C., Calzolari, N., Gangemi, A., Lenci, A., & Oltramari, A.** (**2010**). Ontology and the lexicon: a multi-disciplinary perspective (introduction). In **Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., & Prévot, L.**, editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 1. Cambridge Univ. Press, pp. 3–24.

50. **Ribeiro, R.** (**2003**). *Anotação Morfossintáctica Desambiguada do Português*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

51. **Richardson, S., Dolan, W., & Vanderwende, L.** (**1998**). Mindnet: Acquiring and structuring semantic information from text. *Proceedings of 17th International Conference on Computational Linguistics*, COLING'98, pp. 1098–1102.

52. **Rocha, P. & Santos, D.** (**2000**). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. **Nunes, M.**, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, São Paulo: ICMC/USP, pp. 131–140.

53. **Romão, L.** (**2007**). *Reconhecimento de Entidades Mencionadas em LÃngua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

54. **Santos, D.** (**2010**). *Extracção de relações entre entidades mencionadas*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

55. **Sidorov, G.** (**2013**). Non-continuous syntactic N-grams. *Polibits*, Vol. 48, pp. 67–75.

56. **Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L.** (**2013**). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, Vol. 41, No. 3, pp. 853–860.

57. **Vanderwende, L.** (**1995**). Ambiguity in the acquisition of lexical information. *Proceedings of the AAAI 1995 Spring Symposium, Working notes of the symposium on representation and acquisition of lexical knowledge*, pp. 174–179.

58. **Vanderwende, L., Kacmarcik, G., Suzuki, H., & Menezes, A.** (**2005**). Mindnet: An automatically-created lexical resource. *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, Association for Computational Linguistics, pp. 8–9.

59. **Vicente, A.** (**2013**). *LexMan: um Segmentador e Analisador Morfológico com transdutores*. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

60. **Winston, M., Chaffin, R., & Herrmann, D.** (**1987**). A taxonomy of part-whole relations. *Cognitive Science*, Vol. 11, pp. 417–444.

61. **Zhang, L., Liu, B., Lim, S. H., & O'Brien-Strain, E.** (**2010**). Extracting and ranking product features in opinion documents. *Proceedings of the 23rd International Conference on Computational Linguistics COLING '10: Posters*, Stroudsburg, PA, USA, pp. 1462–1470.

Ilia Markov received his Bachelor degree in Computer Engineering in 2001 from the Kaliningrad

State Technical University, Russia. He obtained his Master degree in Language Sciences in 2012 from the University of Algarve, Portugal. He is currently a PhD student at Instituto Politécnico Nacional, Center for Computing Research, Mexico. His main research interests include natural language processing, computational linguistics, and information retrieval.

Nuno Mamede graduated in Electrotechnical and Computers Engineering by the Instituto Superior Técnico (IST), Lisbon, in 1981, and received his MSc and PhD degrees in Electrotechnic and Computers Engineering, from the same University in 1985 and 1992, respectively. In 1982 he started as lecturer and since 2006 he holds a position of Associate Professor in Instituto Superior Técnico, where he has taught Digital Systems, Object Oriented Programming, Programming Languages, knowledge representation, Natural Language Processing. He has been a researcher at INESC-ID Lisboa, in Lisbon, since its creation in 1980. He participated in the foundation of L2F where hold a position in the Executive Board. His activities have been in the areas of Written Natural Language Processing, namely on Syntactic Processing, Named Entity Recognition, and Natural Language Interfaces to Data Bases. He has authored a significant number of scientific papers. He is a member of AAAI, ACM and ACL.

Jorge Baptista received his Bachelor and Master degrees in Linguistics from the Faculty of Letters of the University of Lisbon, in 1990 and 1995, respectively. He has a PhD in Linguistics (syntax) from University of Algarve (2001). He is an Associate Professor at University of Algarve and an invited researcher at L2F, INESC-ID Lisbon. His main research interests are in computational and theoretical linguistics (syntax, grammar, large coverage lexica, corpus linguistics, machine translation).

Article received on 02/06/2015; accepted on 06/11/2015.
Corresponding author is Ilia Markov.