Automatic Identification of Whole-Part Relations in Portuguese

Ilia Markov

Dissertation for obtaining the Master Degree in

Language Sciences

Advisor: Prof. Doutor Jorge Manuel Evangelista Baptista (Univ. Algarve / FCHS)
Co-advisor: Prof. Doutor Nuno João Neves Mamede (Univ. Lisboa / IST)

Faro, 2014

Automatic Identification of Whole-Part Relations in Portuguese
Declaração de autoria do trabalho

Declaro ser o(a) autor(a) deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Acknowledgements

First, and foremost, I would like to thank my supervisor, Prof. Jorge Baptista, for giving me his expert opinion on each detail of this work and for encouraging me to go further. He was always ready to help both scientifically and personally, and I felt his support in any situation. His enthusiasm, energy, knowledge, experience, and a permanent desire to work hard made the accomplishment of this work possible. I thank my destiny for meeting this person and having a chance to learn so much from him.

I am very grateful to my co-supervisor, Prof. Nuno Mamede, who I had the pleasure to meet and work with, for being always there for me to hear all my questions, discuss and give his insight on the topics addressed in this dissertation. I very appreciate his cooperation and will to help; it was a privilege working with him.

I would also like to thank the members of the L^2F group at INESC-ID Lisboa: David Martins de Matos, Ricardo Ribeiro, Fernando Batista, and Hugo Meinedo, for being always available and eager to help, despite all the busyness.

Pursuing my higher education was made possible by the Erasmus Mundus Action 2 2011-2574 Triple I - Integration, Interaction and Institutions scholarship. I am also thankful to Unitex/GramLab for the student participant scholarship for attending the 2nd Unitex/GramLab Workshop at The University Paris-Est Marne-la-Vallée.

A final word of gratitude is dedicated to my parents and friends for all their support.

Faro, June 5^{th} 2014

Ilia Markov

Resumo

Neste trabalho, procurou-se melhorar a extração de relações semânticas entre elementos textuais tal como é atualmente realizada pela STRING, um sistema híbrido de Processamento de Linguagem Natural (*PLN*), baseado em métodos estatísticos e regras híbrido, e desenvolvido para o Português. Visaram-se as relações *todo-parte* (*meronímia*), que pode ser definida como uma relação semântica entre uma entidade que é percebido como parte integrante de outra entidade, ou a relação entre um membro e um conjunto de elementos. Neste caso, vamos-nos concentrar num tipo de meronímia envolvendo entidades humanas e nomes *parte-do-corpo* (*Npc*); *e.g.*, *O Pedro partiu uma perna*: `WHOLE-PART(Pedro,perna)`. Para extrair este tipo de relações parte-todo, foi construído um módulo de extração de relações meronímicas baseado em regras e que foi integrado na gramática do sistema de STRING.

Cerca de 17.000 instâncias de *Npc* foram extraídas do primeiro fragmento do corpus CETEMPúblico para a avaliação deste trabalho. Foram também recolhidos 79 casos de *nomes de doença* (*Nd*), derivados a partir de um *Npc* subjacente (*e.g.*, *gastrite-estômago*). A fim de produzir um corpus de referência (golden standard) para a avaliação, foi selecionada uma amostra aleatória estratificada de 1.000 frases, mantendo a proporção da frequência total de *Npc* no corpus. Esta amostra também inclui um pequeno número de *Nd* (6 lemas, 17 frases). Essas instâncias foram repartidas e anotadas por quatro falantes nativos de português. 100 frases foram dadas a todos os anotadores a fim de calcular o acordo inter-anotadores, que foi considerado entre "razoável" (fair) e "bom" (good).

Comparando a saída do sistema com o corpus de referência, os resultados mostram, para as relações parte-todo envolvendo *Npc*, 0,57 de precisão, 0,38 de cobertura (recall), 0,46 de medida-F e 0,81 de acurácia. A cobertura foi relativamente pequena (0,38), o que pode ser explicada por vários fatores, tais como o facto de, em muitas frases, o *todo* e a *parte* não estarem relacionadas sintaticamente e até se encontrarem por vezes bastante distantes. A precisão é um pouco melhor (0,57). A acurácia é relativamente elevada (0,81), uma vez que existe um grande número de casos *verdadeiro-negativos*. Os resultados para os nomes de doença, embora o número de casos seja pequeno, mostram uma 0,50 de precisão, 0,11 de cobertura, 0,17 de medida-F e 0,76 de acurácia. A cuidadosa análise de erros realizada permitiu detetar as principais causas para este desempenho, tendo sido possível, em alguns casos, encontrar soluções para diversos problemas. Foi então realizada uma segunda avaliação do desempenho do sistema, verificando-se uma melhoria geral dos resultados: a precisão melhorou +0,13 (de 0,57 para 0,70), a cobertura +0,11 (de 0,38 para 0,49), a medida-F +0,12 (de 0,46 para 0,58) e a acurácia +0,04 (de 0,81 para 0,85). Os resultados

para os *Nd* permaneceram idênticos.

Em suma, este trabalho pode ser considerado como uma primeira tentativa de extrair relações parte-todo, envolvendo entidades humanas e *Npc* em Português. Um módulo baseado em regras foi construído e integrado no sistema STRING, tendo sido avaliado com resultados promissores.

Palavras-chave: relação todo-parte, meronímia, nome parte-do-corpo, nome de doença, Português.

Abstract

In this work, 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* relations (*meronymy*), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. 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)`. 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.

Around 17,000 *Nbp* instances were extracted from the first fragment of the CETEMPúblico corpus for the evaluation of this work. We also retrieved 79 instances of *disease nouns* (*Nsick*), which are derived from an underlying *Nbp* (*e.g.*, *gastrite-estômago* 'gastritis-stomach'). 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). These instances were annotated by four native Portuguese speakers, and for 100 of them the inter-annotator agreement was calculated and was deemed from "fair" to "good".

After confronting the produced golden standard against the system's output, the results for *Nbp* show 0.57 precision, 0.38 recall, 0.46 F-measure, and 0.81 accuracy. The recall is relatively small (0.38), which can be explained by many factors such as the fact that in many sentences, the *whole* and the *part* are not syntactically related. The precision 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 0.50 precision, 0.11 recall, 0.17 F-measure, and 0.76 accuracy. A detailed error analysis was performed, some improvements have been made, and a second evaluation of the system's performance was carried out. It 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.

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

Keywords: whole-part relation, meronymy, body-part noun, disease noun, Portuguese.

Resumo Alargado

Neste trabalho, procuramos melhorar a extração de relações semânticas entre elementos textuais tal como é atualmente realizada pelo sistema STRING, um sistema híbrido, com base em regras e métodos estatísticos, de Processamento de Linguagem Natural (*PLN*) desenvolvido para o Português. Neste sentindo, visamos as relações *parte-todo* (*meronímia*), ou seja, um tipo de relação semântica entre uma entidade que é percebido como parte integrante de outra entidade, ou um membro de um conjunto. Neste caso, concentram-nos no tipo de meronímia envolvendo entidades humanas e *nomes parte-do-corpo* (*Npc*). Enquanto um tipo de relações semânticas, as relações parte-todo contribuem para a coesão e coerência de um texto e a sua identificação pode ser útil em várias tarefas de PLN, como sistemas de pergunta-resposta, sumarização de texto, tradução automática, extração de informação, recuperação de informação, resolução de anáfora, anotação de papéis semânticos, entre outras.

Foi feita uma revisão dos principais trabalhos relacionados, prestando uma atenção especial à extração relações parte-todo em Português. Dois analisadores sintáticos de Português bem conhecidos foram considerados, a fim de discernir como lidam com a extração de relações parte-todo: o analisador PALAVRAS, consultado através do sistema VISL, e o LX-Anotador de Papéis Semânticos. A julgar pelas versões em linha ou demos desses sistemas disponíveis, aparentemente, nenhum destes analisadores extrai relações parte-todo, pelo menos de forma explícita. Além disso, de acordo com a nossa análise dos trabalhos relacionados e outros comentários recentes da literatura sobre a extração de relações semânticas, não foram identificados outras menções de sistemas de extração de relações parte-todo para o Português.

Para extrair relações parte-todo, foi construído um módulo de extração de meronímia, baseado em regras e integrado na gramática do sistema de STRING. Este módulo contém 29 regras gerais, que tratam das construções sintáticas mais relevantes que desencadeiam este tipo de relações meronímica; e um conjunto de 87 regras para 29 *nomes de doença* (*Nd*), a fim de capturar os *Npc* subjacentes. Um conjunto de cerca de 400 regras também foi desenvolvido para evitar que as relações parte-todo fossem extraídas no caso de os *Npc* constituírem elementos de expressões fixas idiomáticas. Este trabalho também abordou as situaões em que há uma relação dentro da mesma frase entre diferentes *Npc*; por exemplo: *A Ana pinta as unhas dos pés*. Também foram tratados os casos que envolvem um nome determinativo e um *Npc* e em que esse determinante designa uma parte do *Npc*; *e.g.*, *O Pedro encostou a ponta da língua ao gelado*. Cada um destes casos desencadeia conjuntos de dependências diferentes. 54 regras foram construídas para associar certos *Npc* com os nomes determinativos que designam as suas partes.

Para a avaliação do trabalho utilizou-se o primeiro fragmento do corpus CETEMPúblico (14,7 milhões de tokens e 6,25 milhões de palavras) para extrair as frases que envolvem *Npc* e *Nd*. Usando os dicionários de *Npc* (151 lemas) e de *Nd* (29 lemas), construído especificamente para léxico STRING, foram extraídos do corpus 16.746 frases com *Npc* e 79 casos de *Nd*. A fim de produzir um texto anotado de referência para a avaliação, foi selecionada uma amostra aleatória estratificada de 1.000 frases, mantendo a proporção da frequência total de *Npc* no corpus. Esta amostra também inclui um pequeno número de *Nd* (6 lemas, 17 frases). As 1.000 frases de saída foram divididas em quatro conjuntos de 225 frases cada. Cada conjunto foi então dado a um anotador diferente (falante nativo de Português), e um conjunto comum de 100 frases foram adicionados a cada grupo, a fim de avaliar a concordância entre anotadores. Foi pedido aos anotadores que acrescentassem a cada frase a dependência parte-todo, tal como fora previamente definida num conjunto de diretrizes de anotação, utilizando o formato do parser da cadeia. Para avaliar a concordância entre anotadores usamos a ferramenta ReCal3, para 3 ou mais anotadores. Os resultados mostraram que o acordo médio entre pares de anotadores é de 0,85, a medida de acordo entre anotadores Fleiss-Kappa é de 0,62, e o acordo médio Cohen-Kappa é de 0,63. Segundo Landis e Koch, este números correspondem ao limite inferior de acordo "substancial"; no entanto, de acordo com Fleiss, estes resultados correspondem a um acordo entre anotadores a meio caminho entre "razoável" ("fair") e "bom". Em vista destes resultados, assumiu-se que para o restante da amostra, anotada de forma independente e sem sobreposição pelos quatro anotadores, o processo de anotação era suficientemente consistente e podia ser utilizado como um padrão de referência para a avaliação da saída do sistema.

Depois de definir este padrão de referência, este foi comparado com a saída do sistema. Para os *Npc*, os resultados mostram 0,57 de precisão, 0,38 de cobertura (ou abrangência; "recall"), 0,46 de medida-F, e 0,81 de acurácia ("accuracy"). A cobertura é relativamente reduzida (0,38), o que pode ser explicado pelo facto de, em muitas frases, os elementos que designam o *todo* e a *parte* não estarem sintaticamente relacionado e se encontrarem muito longe uns dos outros; no entanto, os anotadores foram capazes de ultrapassar estas dificuldades, assinalando a relação meronímica. Outros casos relevantes foram aqueles em que as regras não foram acionados por causa de alguns substantivos humanos e os pronomes pessoais, em geral, são se encontrarem marcados na cadeia com o traço de humano; as situações em que um *Npc* é um modificador/complemento de um substantivo ou um adjetivo (e não de um verbo), situação que não tinha sido contemplada neste estudo. Estes casos, levantam o problema da localização deste módulo da meronímia na arquitetura da cadeia de processamento: uma parte desta tarefa deve ser também realizada após a resolução de anáforas.

A precisão da tarefa é um pouco melhor (0,57). A acurácia é relativamente elevada (0,81) uma vez que existe um grande número de casos verdadeiros-negativos. Os resultados para os *Nd*, embora o número de casos seja pequeno, mostram uma precisão de 0,50, 0,11 de cobertura, 0,17 de medida-F e 0,76 de acurácia. Realizou-se uma análise de erro detalhada para determinar os casos que mais contribuíram para estes resultados, o que levou a que, para algumas situações identificadas, se pudesse propor e implementar diferentes soluções. Foi então realizada uma segunda avaliação do desempenho do sistema e esta mostrou que a precisão melhorava cerca de 0,13 (de 0,57 para 0,70), a cobertura 0,11

(de 0,38 para 0,49), a medida-F 0,12 (de 0,46 para 0,58) e a acurácia 0,04 (de 0,81 para 0,85). Os resultados para os *Nd* permaneceram idênticos.

Para concluir, este trabalho pode ser considerado como uma primeira tentativa de extrair relações parte-todo em Português, neste caso, envolvendo entidades humanas e *Npc*. Foi construído um módulo baseado em regras, que foi integrado no sistema STRING e avaliado com resultados promissores.

Table of Contents

Acknowledgements iii

Resumo v

Abstract vii

ResumoAlargado ix

List of Figures xviii

List of Tables xix

1 Introduction 1

1.1 Context . . . 1

1.2 Goals . . . 2

1.3 Structure . . . 2

2 Related Work 3

2.1 Whole-Part Relations . . . 3

2.2 Whole-Part Relations Extraction . . . 5

2.3 Existing Ontologies for Portuguese . . . 8

2.3.1 WordNet . . . 8

2.3.2 PAPEL . . . 10

2.3.3 Onto.PT . . . 11

2.4 Related Work on Whole-Part Relations Extraction in Portuguese . . . 12

2.4.1 PALAVRAS Parser . . . 12

2.4.2 LX Semantic Role Labeller . . . 15

3 Whole-Part Dependencies Extraction Module in STRING 19

3.1 Overview of STRING . . . 19

3.2 Dependency Rules in XIP . . . 21

3.3 The Basic Whole-Part Dependencies Involving Body-Part Nouns . . . 22

3.3.1 Determinative Complements . . . 23

3.3.2 Dative Complements . . . 25
3.3.3 Subject *Nbp* and Determinative Complements . . . 26
3.3.4 Dative Pronouns . . . 29
3.3.5 Possessive Pronouns . . . 30
3.3.6 Complex Dative Restructuring with Subject *Nbp* . . . 31
3.3.7 Subject *Nhum* and Direct Object *Nbp* . . . 37
3.3.8 Subject *Nhum* and Prepositional Phrase with *Nbp* . . . 37
3.4 Determinative Nouns of *Nbp* . . . 43
3.4.1 Relations between *Nbp* . . . 43
3.4.2 Relation between *Nbp* and Parts of *Nbp* . . . 44
3.5 Complex Relations Involving Derived Nouns . . . 46
3.6 Frozen Sentences (idioms) and Exclusion of Whole-Part Relations . . . 48

4 Evaluation **53**
4.1 Evaluation Corpus . . . 53
4.2 Annotation Campaign . . . 55
4.3 Inter-annotator Agreement . . . 56
4.4 Evaluation of the Whole-Part Dependencies Involving *Nbp* and *Nsick* . . . 60
4.4.1 Definition of Evaluation Measures . . . 60
4.4.2 Problematic Cases . . . 61
4.4.3 Evaluation of the System's Overall Performance . . . 63
4.4.4 Evaluation of the System Performance for *Nsick* . . . 63
4.5 Error Analysis . . . 64
4.5.1 False-positives . . . 64
4.5.2 False-negatives . . . 71
4.6 Post-Evaluation . . . 75

5 Conclusions and Future Work **77**
5.1 Conclusions . . . 77
5.2 Future Work . . . 79

Bibliography **81**

A *Nbp* Whole-Part Extraction Rules **91**
A.1 General Rules . . . 91
A.2 Disease Nouns . . . 97

B *Nbp* Lexicon **99**
B.1 Parts of *Nbp* . . . 99
B.2 *Nbp* Disambiguation . . . 100

C Distribution of *Nbp* **101**

D Annotation Guidelines **105**

E Golden Standard **107**

List of Figures

2.1 Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara do João* (lit: Pedro washed the face of João) 'Pedro washed João's face'. . . . 13

2.2 Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara ao João* (lit: Pedro washed the face to João) 'Pedro washed João's face'. . . . 14

2.3 Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara* 'Pedro washed the face'. 15

2.4 Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara do João* (lit: Pedro washed the face of João) 'Pedro washed João's face'. . . . 16

2.5 Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara ao João* (lit: Pedro washed the face to João) 'Pedro washed João's face'. . . . 16

2.6 Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara* 'Pedro washed the face'. . . . 17

3.1 STRING Architecture (from [Mamede-et-al-2012]). . . . 19

3.2 WHOLE-PART relations for the sentence *O Pedro partiu o braço do João* 'Pedro broke the arm of João'. . . . 24

3.3 WHOLE-PART relations for the sentence *O Pedro partiu o braço dele* (lit: Pedro broke the arm of him) 'Pedro broke his arm'. . . . 25

3.4 WHOLE-PART relations for the sentence *A rapariga de olhos azuis* 'The girl with blue eyes'. . . . 26

3.5 WHOLE-PART relations for the sentence *O Pedro partiu o braço ao João* 'Pedro broke the arm to João'. . . . 27

3.6 WHOLE-PART relations for the sentence *O braço do Pedro está partido* (lit: The arm of Pedro is broken) 'Pedro's arm is broken'. . . . 28

3.7 WHOLE-PART relations for the sentence *O braço dele está partido* (lit: The arm of him is broken) 'His arm is broken'. . . . 28

3.8 WHOLE-PART relations for the sentence *O Pedro partiu-lhe o braço* 'Pedro broke him the arm'. . . . 30

3.9 WHOLE-PART relations for the sentence *O Pedro não lhe partiu o braço* (lit: Pedro did_not to-him break the arm) 'Pedro did not break his arm'. . . . 31

3.10 WHOLE-PART relations for the sentence *O Pedro partiu o seu braço* 'Pedro broke his arm'. . . . 32

3.11 WHOLE-PART relations for the sentence *Os braços doem-lhe* (lit: The arms hurt him) 'His arms are hurting'. . . . 32

3.12 WHOLE-PART relations for the sentence *Os braços não lhe doem* (lit: The arms do_not to-him hurt) 'His arms are not hurting'. . . . 33

3.13 Initial, incorrect parse for the sentence: *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'. . . . 34

3.14 First step of the parsing for the sentence *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'. . . . 35

3.15 Correct parsing for the sentence *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'. . . . 36

3.16 WHOLE-PART relations for the sentence *Não lhe doem os braços* (lit: Not to-him are_hurting the arms) 'His arms are not hurting'. . . . 36

3.17 WHOLE-PART relations for the sentence *O Pedro partiu um braço* 'Pedro broke an arm'. . . . 37

3.18 WHOLE-PART relations for the sentence *O Pedro coçou na cabeça* (lit: Pedro scratched on the head) 'Pedro scratched the head'. . . 38
3.19 WHOLE-PART relations for the sentence *O Pedro espalhou óleo nas pernas à Joana* 'Pedro spread oil on the legs of Joana'. 39
3.20 WHOLE-PART relations for the sentence *O Pedro feriu-se no braço* (lit: Pedro wounded himself in the arm) 'Pedro wounded his arm'. 40
3.21 WHOLE-PART relations for the sentence *O Pedro bateu-me nas pernas* (lit: Pedro hit me in the legs) 'Pedro hit my legs'. 41
3.22 WHOLE-PART relations for the sentence *O Pedro andava de braços cruzados* 'Pedro walked with arms crossed'. 42
3.23 WHOLE-PART relations for the sentence *O Pedro levava o Zé pela mão* 'Pedro led Ze by the hand'. 43
3.24 WHOLE-PART relations for the sentence *A Ana pinta as unhas dos pés* (lit: Ana paints the nails of the feet) 'Ana paints the toenails'. 44
3.25 WHOLE-PART relations for the sentence *O Pedro tem uma gastrite* 'Pedro has gastritis'. 47
3.26 WHOLE-PART relations for the sentence *O Pedro está com uma gastrite* (lit: Pedro is with a gastritis) 'Pedro has gastritis'. 47
3.27 WHOLE-PART relations for the sentence *A gastrite do Pedro é grave* 'Pedro's gastritis is severe'. 48
3.28 Frozen sentences (idioms) and exclusion of whole-part relations. 50

List of Tables

4.1 10 most frequent *Nbp*. 54
4.2 Number of *Nsick*. 54
4.3 Distribution of the annotations in the corpus. 57
4.4 Average Pairwise Percent Agreement. 59
4.5 Fleiss' Kappa. 59
4.6 Average Pairwise Cohen's Kappa (CK). 59
4.7 System's performance for *Nbp*. 63
4.8 System's performance for *Nsick*. 64
4.9 Post-error analysis system's performance for *Nbp*. 76

C.1 Distribution of *Nbp*. 101

Chapter 1

Introduction

1.1 Context

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 and others [Girju-et-al-2003]. For example, for questions like *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 keywords matching and proximity [Girju-et-al-2006]. 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 [Khoo-2006]. 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 relations extraction would be relevant to correctly associate the patients' names and their organs' nouns.

[Zhang-et-al-2010] showed that whole-part relations can be used in the NLP task of opinion mining. Once one is talking about an object (product), one can often refer 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 between the parts and the general product (the whole), one can see if 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 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. 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 to establish their semantic roles more accurately.

Modules for anaphora resolution [Marques-2013] and semantic roles labeling [Talhadas-2014] have been already developed in STRING[1], a hybrid statistical and rule-based NLP chain for Portuguese [Mamede-et-al-2012]. These modules take place at the last steps of the parsing processing. Therefore, our specific meronymy extraction module will also be implemented in the final stages of the processing chain, but before these modules come into action, in order for them to take advantage of the whole-part relations.

1.2 Goals

The goal of this work is to improve the extraction of semantic relations between textual elements in STRING. At this time, only the first steps have been taken in the direction of semantic parsing. This work will target whole-part relations (*meronymy*), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. 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 [Baptista-et-al-2012a], [Baptista-et-al-2012b], [Cabrita-et-al-2013], 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 relations extraction module by capturing meronymic relations.

1.3 Structure

This dissertation is structured as follows: Chapter 2 describes related work on whole-part dependencies extraction; Chapter 3 explains with some detail how this task was implemented in STRING; Chapter 4 presents the evaluation procedure, the results of the task, and the error analysis; Chapter 5 draws the conclusions from this work and points to the future work by providing possible directions for expanding and improving the module here developed.

[1] https://string.l2f.inesc-id.pt/ [last access: 05/06/2014]. All other URLs in this document were also verified on this date.

Chapter 2

Related Work

THIS chapter presents related work, and it is organized in the following way: Section 2.1 provides a brief definition of whole-part relations and succinctly describes different proposals of classification of whole-part relations; in Section 2.4, we present an overview of whole-part relations extraction techniques for the English language; Section 2.3 presents the outline of the existing lexical ontologies for Portuguese: WordNet, PAPEL, and Onto.PT; in Section 2.4, we describe in some detail how two well-known Portuguese parsers (PALAVRAS and LX-SRL) address the extraction of whole-part relations.

2.1 Whole-Part Relations

Whole-part relations (also known as *meronymy*)[1] 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 be treated as a collection of relations, not as a single relation" [Iris-et-al-1988].

A well-known classification of whole-part relations was developed by Winston *et al.* [Winston-et-al-1987]. Six types of whole-part relations were distinguished based on the way parts contribute to the structure of the whole, these consist on:

1. Component-Integral object (*wheel - car*);
2. Member-Collection (*soldier - army*);
3. Portion-Mass (*meter - kilometer*);
4. Stuff-Object (*alcohol - wine*);
5. Feature-Activity (*paying - shopping*);
6. Place-Area (*oasis - desert*).

[1] In the bibliography the term *part-whole* is also often used, but we decided to adopt *whole-part* since in our NLP chain the convention has been adopted to put the governor of the dependency first and the subordinate term second.

As Ittoo and Bouma [Ittoo-and-Bouma-2010] reported, in WordNet [Miller-1995], [Fellbaum-1998], [Fellbaum-2010] whole-part relations are divided into three basic types:

1. Member-of (*e.g.*, *UK* IS-MEMBER-OF *NATO*);
2. Stuff-of (*e.g.*, *carbon* IS-STUFF-OF *coal*);
3. all other whole-part relations under the general name of Part-of (*e.g.*, *leg* IS-PART-OF *table*).

Other classifications, proposed by Odell [Odell-1994] and Gerstl and Pribbenow [Gerstl-and-Pribbenow-1995], are based on the work of Winston *et al.* [Winston-et-al-1987]. Gerstl and Pribbenow [Gerstl-and-Pribbenow-1995] identify different kinds of whole according to their inherent compositional structure: complexes, collections, and masses.
In the taxonomy developed by Keet and Artale [Keet-and-Artale-2008] there is a distinction between *transitive mereological*[2] whole-part relations and *intransitive meronymic* ones. The distinction consists in that 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 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 [Keet-and-Artale-2008] classify mereological relations into the four following categories:

1. involved-in (*chewing - eating*);
2. located-in (*city - region*);
3. contained-in (*tool - trunk*);
4. structural part-of (*engine - car*).

while meronymic relations these authors identify are:

1. member-of (*player - team*);
2. constituted-of (*clay - statue*);
3. sub-quantity-of (*meter - kilometer*);
4. participates-in (*enzyme - reaction*).

In our work, we focus on a specific type of whole-part relations involving *Nbp*. Ittoo and Bouma [Ittoo-and-Bouma-2010] have shown that in information extraction tasks focusing on particular whole-part relation type gives more stable results than using general sets of whole-part relations as seeds for machine-learning algorithms:

[2]Mereology is a sub-discipline in philosophy that concerns the investigation of the whole-part relations.

> "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" [Ittoo-and-Bouma-2010, p. 1334].

In this work, we are neutral to the suggested classifications, even though the whole-part relations here studied can fall into *component-integral object* [Winston-et-al-1987] or into the general *part-of* case, in the classification provided by WordNet.

According to our review of related work and to a recent review of the literature on semantic relations extraction [Abreu-et-al-2013], no works on whole-part relations extraction for Portuguese have been identified[3]. In the Linguateca[4] Joint Evaluation campaigns, a proposal was made for a track on identifying relations between named entities[5]. Some of these relations included (indirect) anaphora and a special type of relation (*v.g.*, `TIPOREL="inclui"` and `TIPOREL="incluido"`), which can in some cases be approximated to the meronymy relation here studied. A detailed presentation of a system for extracting these semantic relations is presented in [Bruckschen-et-al-2008].

The current work also aims at extracting a specific type of whole-part relations, involving *Nbp*, but we adopt a rule-based approach, using the tools and resources available in STRING. This is done under the scope of developing NLP chain STRING for European Portuguese.

2.2 Whole-Part Relations Extraction

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

Hearst [Hearst-1992] 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". The author 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, the author compared the results of the algorithm with information found in WordNet. The author reports that when the set of 152 relations that fit the restrictions of the

[3] At the later stages of this project (May, 2014), we came to know the work of Cláudia Freitas [Freitas-2014]; however, since all the work has been already accomplished, we decided not to take it into consideration at the moment but to use it in the future work.

[4] www.linguateca.pt

[5] www.linguateca.pt/aval_conjunta/HAREM/ReRelEM.html

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." [Hearst-1992, p. 544].

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

Berland and Charniak [Berland-and-Charniak-1999] addressed the acquisition of meronyms using manually-crafted patterns, similar to [Hearst-1992], 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." [Berland-and-Charniak-1999, 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 is about 1 million words. Their systems 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 [Vanderwende-1995, Richardson-et-al-1998] was created from dictionary definitions by automatic tools. It has been maintained by the Microsoft NLP research group up until 2005 [Vanderwende-et-al-2005], and it is supposedly accessible for on-line browsing.[6] In its creation, a broad-coverage parser generates syntactical 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.* [Girju-et-al-2003], [Girju-et-al-2006] 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

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

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 [Miller-et-al-1993]. 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 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.* [Van-Hage-et-al-2006] 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." [Van-Hage-et-al-2006, p. 30].

The authors reported that they were able to acquire 503 whole-part pairs from the AGROVOC Thesaurus[7] 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 [Pantel-and-Pennacchiotti-2006] was developed in order to harvest semantic relations in a text. Espresso is based on the framework adopted in [Hearst-1992]:

> "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." [Pantel-and-Pennacchiotti-2006, § 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.

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

2.3 Existing Ontologies for Portuguese

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 [Hirst-2004], [Prevot-et-al-2010].

The following sections will briefly describe the existing lexical ontologies for Portuguese: WordNet, PAPEL, and Onto.PT.

2.3.1 WordNet

Princeton WordNet[8] [Miller-1995], [Fellbaum-1998], [Fellbaum-2010] is an online lexical database developed at Princeton University[9]. WordNet is a database of words and collocations that groups the words into *synsets*. A synset is a grouping of synonymous words and pointers that describe the relations between this synset and other synsets. The relations represented in WordNet are synonymy, antonymy, hyperonymy/hyponymy, meronymy, troponymy, and entailment.

WordNet is created manually by experts which makes it a highly reliable linguistic resource, but has the disadvantage of its production, development and maintenance being highly costly and very time-consuming. Besides, its lexical coverage and growth are constrained by these production factors.

WordNet made it possible for many NLP applications to be enhanced with new capabilities; furthermore, it was used in various NLP tasks such as question-answering [Pasca-and-Harabagiu-2001, Clark-et-al-2008], text categorisation [Elberrichi-et-al-2006, Rosso-et-al-2004], text summarisation [Bellare-et-al-2004, Plaza-et-al-2010], information retrieval [Voorhees-1998], sentiment analysis [Esuli-and-Sebastiani-2007, Williams-and-Anand-2009], query expansion [Navigli-and-Velardi-2003], determining similarities [Seco-et-al-2004, Agirre-et-al-2009a], intelligent search [Hemayati-et-al-2007], and word sense disambiguation [Resnik-1995, Banerjee-and-Pedersen-2002, Gomes-et-al-2003, Agirre-et-al-2009b].

Whole-part relations are captured by the concept of meronymy, which is applied in WordNet to detachable objects, like *leg*, which is a part of the *body*, or in relation to collective nouns, such as the link between the concepts of *ship* and *fleet*.[10] WordNet was initially developed for the English language, but later on the same framework was adopted for other languages as well.

Portuguese WordNet.PT [Marrafa-2001], [Marrafa-2002], later extended to WordNet.PT Global - *Rede Léxico-Conceptual das variedades do Português* [Marrafa-et-al-2011], is a resource developed by the University of Lisbon[11] in partnership with Instituto Camões[12]. This project aimed at developing a broad-coverage wordnet for the European Portuguese variant. WordNet.PT contains a large set of semantic relations, covering: general-specific; whole-part; equivalence; opposition; categorisation; participation in an event; and defining the event structure. The creation of WordNet.PT is manual, and its structure

[8] WordNet 3.1 is downloadable through http://wordnet.princeton.edu/wordnet/download/. WordNet 3.1 can be queried online, through http://wordnetweb.princeton.edu/perl/webwn

[9] http://www.princeton.edu/main/

[10] http://vossen.info/docs/2002/EWNGeneral.pdf

[11] http://www.ulisboa.pt/

[12] https://www.instituto-camoes.pt/

is based on the EuroWordNet [Vossen-1997] model, and thus inspired by WordNet. According to the information provided by its website[13], WordNet.PT Global contains a network with 10,000 concepts, including nouns, verbs, and adjectives, their lexicalisations in the different Portuguese variants, and their glosses. The concepts, which are a subset of the WordNet.PT concepts, are integrated in a network with more than 40,000 relation instances of several types. On the current website of the WordNet.PT only definitions of the entries are provided, so we could not assess in general the whole-part relations that may have been encoded in this resource.

MWN.PT - MultiWordNet of Portuguese[14] is the Portuguese branch of the MultiWordNet project [Pianta-et-al-2002]. It is developed by the NLX-Natural Language and Speech Group at the University of Lisbon, and can be consulted on the site, but it can not be downloaded, though it is distributed by ELDA-Evaluation and Language Resources Distribution Agency.

MWN.PT presents the synsets and the semantic relations typical of WordNet ontologies, which can be consulted on the site. A small description is provided below:

> "MWN.PT - MultiWordnet of Portuguese (version 1) spans over 17,200 manually validated concepts/synsets, linked under the semantic relations of hyponymy and hypernymy. These concepts are made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese. They are aligned with the translationally equivalent concepts of the English Princeton WordNet and, transitively, of the MultiWordNets of Italian, Spanish, Hebrew, Romanian and Latin." (information provided from the MWN.PT site).

According to [Santos-et-al-2010], the number of *IS-PART-OF* relations presented in MWN.PT is: 592 for words and 504 for triples. Even though meronymy is claimed to be represented in MWN.PT, we have tested the most common Portuguese *Nbp* (*pescoço* 'neck', *perna* 'leg', *palma da mão* 'palm', *maçã-de-adão* 'Adam's apple') but they did not yield any results. For other nouns (*cabeça* 'head', *garganta* 'throat', *mão* 'hand') the meanings presented by the resource do not correspond to *Nbp*.

Another resource is the thesaurus TeP 2.0[15]: Electronic *Thesaurus* for Brazilian Portuguese [Dias-Da-Silva-and-Moraes-2003], [Maziero-et-al-2008] stores sets of synonym and antonym word forms. To the best of our knowledge, this thesaurus does not directly address the issue of whole-part relations. In the current version of the site, TeP 2.0 just provides the definitions associated to each lexical entry. We have tried several frequent *Nbp*, and in some cases definitions are more complete than for MWN.PT (*garganta* 'throat', *pescoço* 'neck', *maçã-de-adão* 'Adam's apple'), while others are just missing (*cabeça* 'head', *palma da mão* 'palm', *pé* 'foot', *perna* 'leg').

[13] http://www.clul.ul.pt/clg/eng/projectos/wordnetgl.html
[14] http://mwnpt.di.fc.ul.pt/
[15] http://www.nilc.icmc.usp.br/tep2/

2.3.2 PAPEL

PAPEL (Palavras Associadas Porto Editora Linguateca)[16] [Oliveira-et-al-2008] is a lexical resource for NLP of Portuguese. It is based on the (semi)automatic extraction of relations between the words appearing in the definitions of the *Dicionário da Língua Portuguesa* (*DLP*) developed by Porto Editora[17].

Unlike other lexical ontologies for Portuguese, PAPEL is public; *i.e.*, freely available, and open for further improvements by the community.

In order to parse the dictionary definitions, PAPEL uses PEN[18] [Oliveira-and-Gomes-2008], a chart parser freely available, that is a Java implementation of the Earley Algorithm [Early-1970]. PEN parses the text according to a grammar file it receives as input and it can yield several analysis for the same text. PAPEL uses specific different grammars to identify different relations between the defined entities corresponding to words in the dictionary.

PAPEL has explicit description of semantic relations, including whole-part relations. These meronymic relations, totalling 5,491 triples, are split into three types: *part-of* (2,418), *part-of-something-with-property* (3,026) and *property-of-something-part-of* (47).[19] These are defined and illustrated as follows:

1. *Part-of*. A triple (two items connected by a predicate) *a* `PARTE_DE` (*part-of*) *b* indicates that *a* is a part or a constituent of *b*. In the context of PAPEL this relation was established between nouns. Examples of these relations in PAPEL are:

citologia 'cytology' `PARTE_DE` *biologia* 'biology'

chaminé 'chimney' `PARTE_DE` *cachimbo* 'smoking pipe'

núcleo 'nucleus' `PARTE_DE` *cometa* 'comet'

cauda 'tail' `PARTE_DE` *cometa* 'comet'

asa 'wing' `PARTE_DE` *avião* 'airplane'

motor 'motor' `PARTE_DE` *avião* 'airplane'

As we can see from these few examples, the `PARTE_DE` relation includes scientific subdisciplines of a broader discipline (biology), (structural) components of a concrete object (airplane, pipe), parts of celestial bodies (comet), etc.

2. *Part-of-something-with-property*. A triple *a* `PARTE_DE_ALGO_COM_PROPRIEDADE` (*part-of-something-with-property*) *b* indicates that *a* is a part of something that has a property *b*. In the context of PAPEL this relation was established between nouns and adjectives. Examples of these relations in PAPEL are:

tampa 'lid' PARTE_DE_ALGO_COM_PROPRIEDADE *coberto* 'covered'

aptidão 'ability' PARTE_DE_ALGO_COM_PROPRIEDADE *talentoso* 'talented'

pêlo 'hair' PARTE_DE_ALGO_COM_PROPRIEDADE *piloso* 'pilose'

3. *Property-of-something-part-of*. A triple *a* `PROPRIEDADE_DE_ALGO_PARTE_DE` (*property-of-something-part-of*) *b* indicates that the quality *a* is attributable to parts of *b*. In the context of PAPEL this relation was established between adjectives and nouns. There are 47 *property-of-something-part-of* relations in PAPEL,

[16] http://www.linguateca.pt/PAPEL/
[17] http://www.portoeditora.pt/
[18] http://code.google.com/p/pen/
[19] Data from PAPEL v. 3.5 [last update: 23.08.2012].

but the authors are not entirely confident of their accuracy/adequacy.[20] Examples of these relations in PAPEL are:

colonial 'colonial' PROPRIEDADE_DE_ALGO_PARTE_DE *hidrozoário* 'hydrozoan'

carbonosa 'carbonaceous' PROPRIEDADE_DE_ALGO_PARTE_DE *chedite* 'chedite'

The last version (3.5) of PAPEL, in the relations PARTE file, contains 638 triples involving *Nbp*, but if we focus on the relations of the type PARTE_DE (*part-of*), only 165 triples involve *Nbp*. Ignoring all cases, which are the majority of triples here included, where no human *Nbp* relation is involved (*e.g.*, *cabeça* 'head' PARTE_DE *rebite* 'rivet'), some entries are obviously incorrect triples, such as the duplicate entry for *barriga* 'belly':

barriga 'belly' PARTE_DE *barrigudo* 'paunchy'

barriga 'belly' PARTE_DE_ALGO_COM_PROPRIEDADE *barrigudo* 'paunchy'

while other entries are correct and useful relations such as:

colo 'lap' PARTE_DE *corpo* 'body'

colo 'lap' PARTE_DE *intestino* 'intestine'

cólon 'colon' PARTE_DE *intestino* 'intestine'

However, since this resource targets parts of objects that are, for the most part, non-human, it is of little use for our study. Even for the the relation between two *Nbp* such as *unha-pé* 'nail-foot', or *cotovelo-braço* 'elbow-arm' (see section 3.4), many of these obvious pairs are missing.

2.3.3 Onto.PT

Onto.PT[21] [Oliveira-2012] is a lexical ontology for Portuguese. Similarly to PAPEL, Onto.PT is freely available for download. The source is based on Wordnet model: Onto.PT contains synsets - groups of synonymous words, and semantic relations, held between synsets. Onto.PT integrates lexical-semantic knowledge from five lexical resources, more precisely from three dictionaries (Dicionário PRO da Língua Portuguesa (DLP), through PAPEL; Dicionário Aberto (DA)[22]; and Wiktionary.PT[23]) and two thesauri (TeP and OpenThesaurus.PT (OT.PT)). The dictionaries were used for the extraction of semantic relations by using symbolic techniques over the dictionary definitions: semantic relations were extracted by connecting lexical items according to their possible senses. The authors manually encoded a set of semantic patterns, organised in grammars, for processing the dictionaries.

The approach for the acquisition, organisation and integration of lexical-semantic knowledge involves three main automatic steps:

1. Extraction: instances of semantic relations, held between lexical items, are automatically extracted from text, following a pattern based extraction on dictionary definitions.

[20] In fact, the authors inform that "version 2.0 of PAPEL contains 17 occurrences of this relation, all wrong. PAPEL 3.0 has more instances of this relation, but most of them can not be regarded as correct" (our translation, taken from http://www.linguateca.pt/PAPEL/descricao_relacoes_PAPEL.html#PROPRIEDADE_DE_ALGO_PARTE_DE).

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

[22] http://www.dicionario-aberto.net/

[23] https://pt.wiktionary.org/

2. Thesaurus enrichment and clustering: synsets are augmented with the extracted synonymy relations.

3. Ontologisation: the lexical items in the arguments of the non-synonymy relation instances are attached to suitable synsets.

This approach for creating wordnets automatically was baptised as ECO, which stands for Extraction, Clustering and Ontologisation.

The current version of Onto.PT (3.5) contains more than 100,000 synsets and more than 170,000 labelled connections, which represent semantic relations (synonymy, hypernymy, part-of, causation, purpose-of, and manner-of). According to the materials that can be downloaded from the website, there are 1,177 relations of the type `PARTE_DE` (*part-of*); 3,200 of the type `PARTE_DE_ALGO_COM_PROPRIEDADE` (*part-of-something-with-property*); and 44 of the type `PROPRIEDADE_DE_ALGO_PARTE_DE` (*property-of-something-part-of*). The type of relations involving *Nbp* show the same issues like the ones we mentioned about PAPEL.

2.4 Related Work on Whole-Part Relations Extraction in Portuguese

In this review of the state of the art on whole-part relations extraction in Portuguese, we now focus on two well-known parsers for Portuguese: PALAVRAS [Bick-2000] - Visual Interactive Syntax Learning (*VISL*) and LX Semantic Role Labeller [Branco-and-Costa-2010].

In order to test the performance of these parsers, we use a set of testing sentences aimed at capturing different syntactic configurations where whole-part relations are involved: (i) a determinative complement of an *Nbp* object; (ii) a dative complement of a verb with an object *Nbp*; and (iii) an *Nbp* object without any other complement.

2.4.1 PALAVRAS Parser

PALAVRAS[24] [Bick-2000] is a rule-based parser with constraint grammar framework [Karlsson-1990]. In this framework, words are linked through dependencies and there are is no chunking (even if the concept of phrase underlies the dependencies), so that the output of the system is not the usual parsing trees that we are used to see in syntax books, based on generative grammar or immediate constituents analysis. Instead, these parse trees can be read as a graph where each node is a word in the sentence, and the transitions are the syntactic dependencies connecting them up to a *root* node.

The first sentence (1) is a simple case where there is a determinative `PP`, complement *de N* 'of N' of the *Nbp*:

(1) *O Pedro lavou a cara do João* (lit: Pedro washed the face of João) 'Pedro washed João's face'

The output of PALAVRAS parser, using the VISL interface, from sentence (1) is given in Fig. 2.1.

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

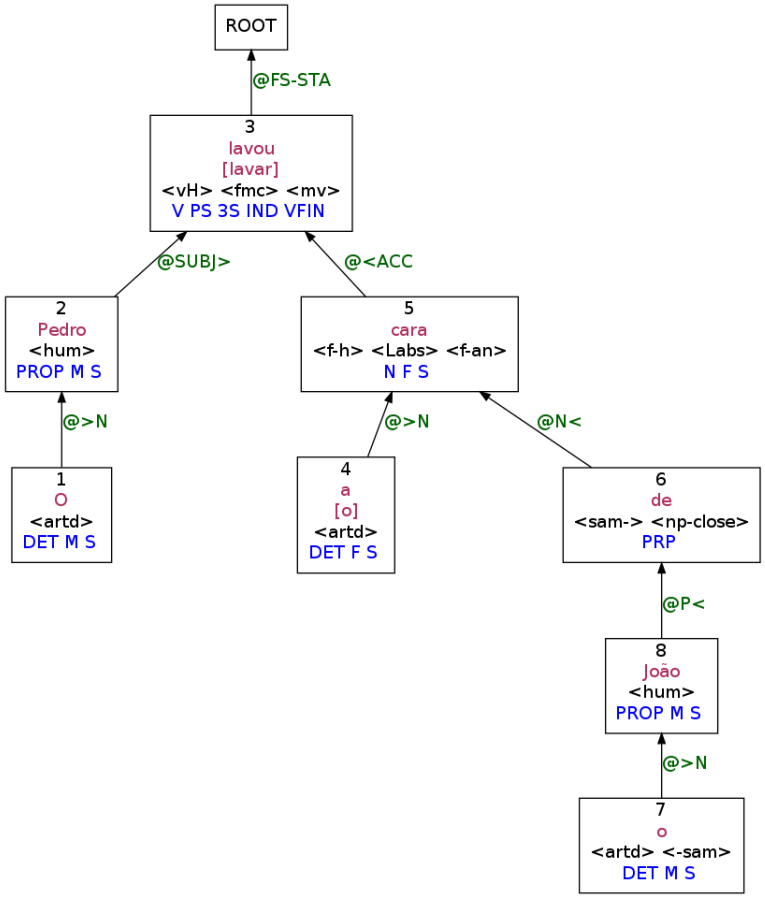


Figure 2.1: Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara do João* (lit: Pedro washed the face of João) 'Pedro washed João's face'.

In this example (Fig. 2.1), the parse is correct. The determinative complement establishes the dependency between *cara* 'face' and *João*. One could say that they are linked, even though there is no explicit semantic relation between the *Nbp* and the human noun.

The next example (2) 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*:

(2) *O Pedro lavou a cara ao João* (lit: Pedro washed the face to João) 'Pedro washed João's face'

The output of PALAVRAS parser on sentence (2) is given in Fig 2.2.

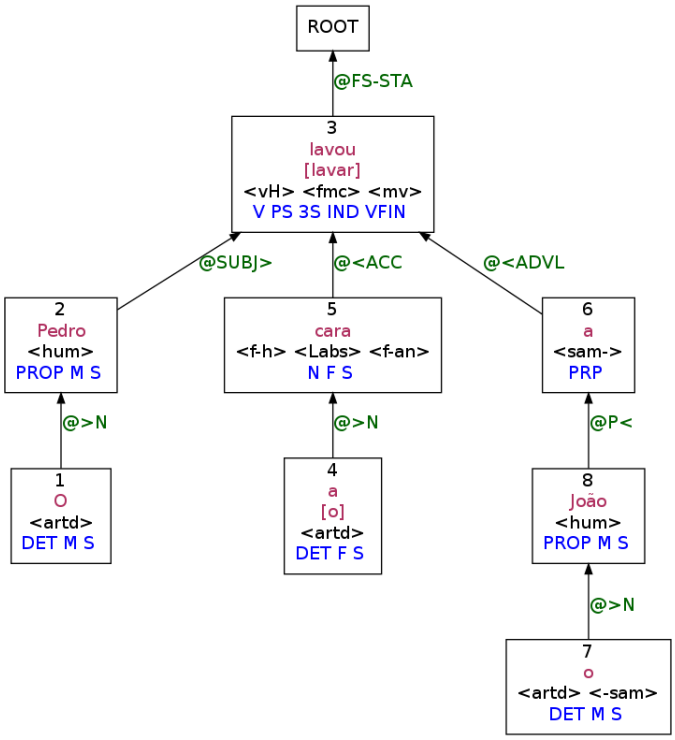


Figure 2.2: Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara ao João* (lit: Pedro washed the face to João) 'Pedro washed João's face'.

The parser correctly splits the sentence into 3 constituents: the subject, the (direct) object, and the prepositional complement. However, the parser incorrectly attributes the syntactic function `ADVL`, which is used for *adverbial adjunct* instead of the *dative complement* dependency (`PIV`).

Finally, the case (3) with just a human subject and an *Nbp* direct object, without any other complement, and where there is meronymy between the human subject and the *Nbp*:

(3) *O Pedro lavou a cara* 'Pedro washed the face'

The output of PALAVRAS parser on sentence (3) is given in Fig. 2.3.

Here again, as we can see, there is no specific element in the graph that establish a semantic relation between *cara* 'face' and the subject of the sentence.

So far, the author does not address the issue, at least in the version available to the public.

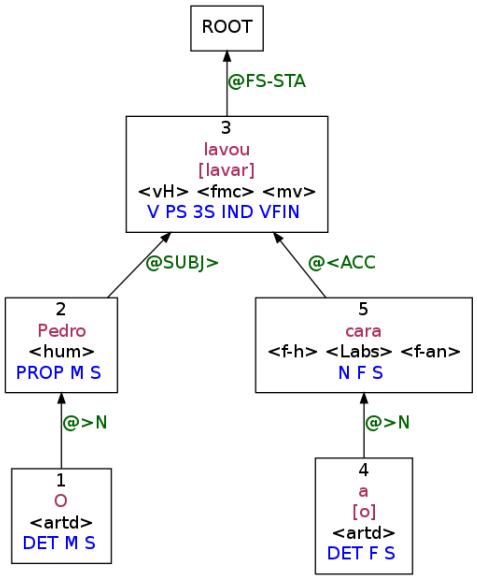


Figure 2.3: Output of PALAVRAS parser on the sentence: *O Pedro lavou a cara* 'Pedro washed the face'.

2.4.2 LX Semantic Role Labeller

LX Semantic Role Labeller[25] [Branco-and-Costa-2010] extracts semantic relations by marking labeling the parse tree nodes with their argument status. The system uses the Berkley Parser [Silva-et-al-2010] and the PHPSyntaxTree Visualizer. The parser uses probabilistic grammars and it is based on the theoretical perspective of X-bar generative syntax theory [Chomsky-1970]. The parser is build using a manually annotated corpus (CINTIL-Corpus Internacional do Português, developed at the University of Lisbon[26]; the corpus currently contains 1 million annotated words[27]) and out-of-the-shelf machine learning tools.

In order to test the performance of this parser, we use the same testing sentences as for testing PALAVRAS parser.

The output of LX Semantic Role Labeller on sentence (1) is given in Fig. 2.4.

In this example (Fig. 2.4), the parse is correct. Concerning semantic roles, two arguments are determined: `ARG1` – the first argument, corresponding to the subject of the verb, and `ARG2` – the second argument, corresponding to the (direct) object of the verb. Nevertheless, we are not sure how to interpret it[28], but as another argument position has been found, in the prepositional phrase, `PP-ARG1`, and this is represented below `ARG2`, maybe there is an underlying relation between *João* and *cara* 'face'.

The output of LX Semantic Role Labeller on sentence (2) is given in Fig. 2.5.

In this example (Fig. 2.5), it is not clear that the parse is completely correct, because the complement *ao João* should be a dative/indirect complement of the verb, and should not be hanging from the noun *cara* 'face' – at least in a traditional immediate constituents analysis.

Unlike the previous case (Fig. 2.4), that had a similar syntactic structure, now the parse tree identifies

[25] http://lxcenter.di.fc.ul.pt/services/en/LXSemanticRoleLabeller.html
[26] http://www.ulisboa.pt/
[27] http://cintil.ul.pt/pt/cintilfeatures.html#corpus
[28] We could not find on the site any relevant documentation of the Parser that could help interpreting these annotations.

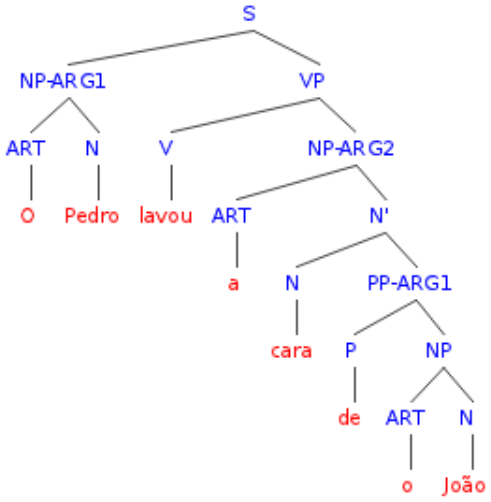


Figure 2.4: Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara do João* (lit: Pedro washed the face of João) 'Pedro washed João's face'.

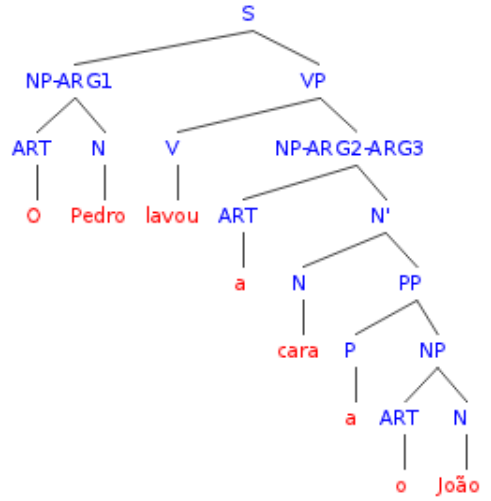


Figure 2.5: Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara ao João* (lit: Pedro washed the face to João) 'Pedro washed João's face'.

3 arguments, placing an ARG3 tag next to the ARG2. If this is correctly interpreted, it may be that the three arguments of the verb *lavar* 'to wash' were identified, though it is unclear why the tag ARG3 is not placed on the corresponding NP node, and two distinct roles were collapsed in the same NP node.

Finally, the output of LX Semantic Role Labeller on sentence (3) is given in Fig. 2.6.

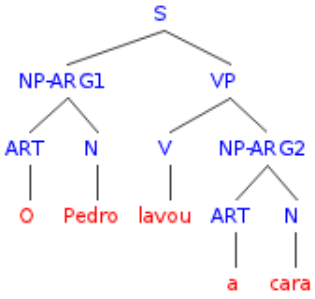


Figure 2.6: Output of LX Semantic Role Labeller on the sentence: *O Pedro lavou a cara* 'Pedro washed the face'.

The sentence is parsed correctly, but there is no explicit semantic relation between *cara* 'face' and *Pedro*.

Thus, judging from the available on-line versions/demos of these systems, apparently, none of these parsers extracts whole-part relations, at least explicitly.

In this chapter, we presented the existing classifications of whole-part relations; an overview of whole-part relations extraction techniques for the English and the Portuguese languages, paying particular attention to existing lexical ontologies for Portuguese and to two well-known parsers for Portuguese: PALAVRAS and LX Semantic Role Labeller.

Chapter 3

Whole-Part Dependencies Extraction Module in STRING

THIS chapter is comprised of six parts: in Section 3.1, the overview of STRING is presented; in Section 3.2, the syntax of the dependency rules used in XIP is briefly described; Section 3.3 describes the way the basic whole-part dependencies involving *Nbp* are extracted in the Portuguese grammar for the XIP parser; Section 3.4 describes the rules for extraction determinative nouns of *Nbp*; Section 3.5 presents the rules that have been made in order to extract complex relations involving derived nouns; Section 3.6 explains the strategy we adopted to deal with the situations where frozen sentences (idioms) containing *Nbp* elements are involved.

3.1 Overview of STRING

STRING [Mamede-et-al-2012] [1] is a fully-fledged NLP chain that 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. 3.1.

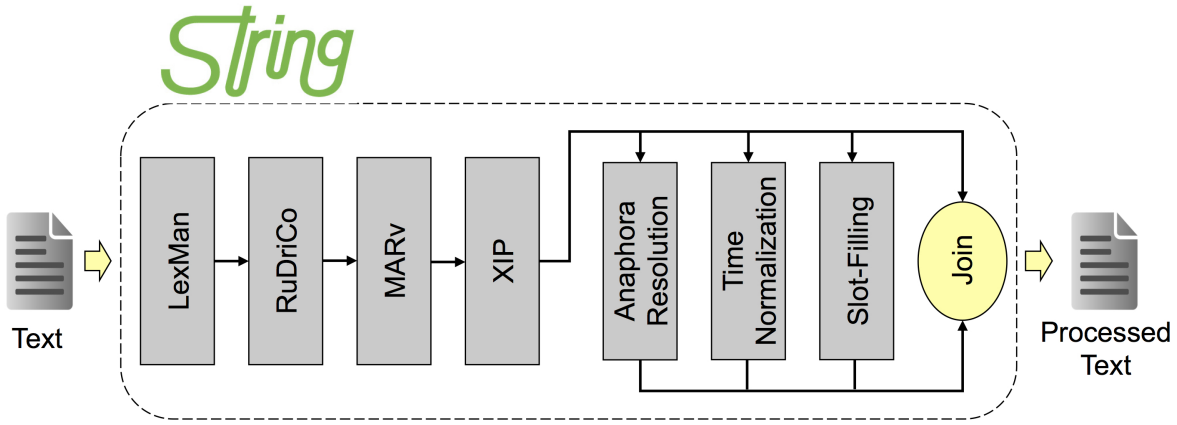


Figure 3.1: STRING Architecture (from [Mamede-et-al-2012]).

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

LexMan [Vicente-2013], the lexical analyzer, is responsible for text segmentation; it divides a text into sentences and these into tokens and assigns tokens all their potential part-of-speech (POS) tags, along with other morphosyntactic features such as gender, number, tense, etc. LexMan is able to identify simple and compound words, abbreviations, email addresses, URLs, punctuation, and other symbols.

Then, RuDriCo [Diniz-2010], a rule-based converter, modifies the segmentation that is done by the former module through declarative rules, which are based on the concept of pattern matching. It executes a series of rules to solve contractions (*e.g.*, *na = em + a* 'in-the'); it also identifies some types of compounds words and joins them as a single token. Furthermore, the module can also be used to solve (or introduce) morphosyntactic ambiguities.

Before the syntactic parsing, a statistical POS disambiguator (MARv) [Ribeiro-2003] is applied, analyzing the POS tags that were attributed to each token in the previous step of the processing chain and then choosing the most likely POS tag for each token. MARv uses a ME (maximum entropy) model [Harremoes-and-Topsoe-2001] based on the Viterbi algorithm [Viterbi-1967] to adequately select the correct POS for a word given its context. The language model is based on second-order (trigram) models, which codify contextual information concerning entities, and unigrams, which codify lexical information. The classification model used by MARv is trained on a 250k words Portuguese corpus, which contains texts from books, journals, and magazines. The corpus has been manually annotated and carefully revised. More recently, this process was repeated and more problematic categories were addressed (*e.g.* personal pronouns), including verb lemma disambiguation (*e.g.*, *ser / ir* 'to be / to go'). This lead to an improvement in the POS-tagging results, that now stand around +98%.

The next step is performed by XIP (Xerox Incremental Parser) [Ait-Mokhtar-et-al-2002]. XIP is a rule-based parser that performs chunking; *i.e.*, the identification of the elementary sentence constituents (`NP`, `PP`, etc.), and extracts syntactic and semantic dependencies between those chunk heads.

After XIP, several post-syntactic modules may come into play to solve specific tasks such as time expression normalization [Mauricio-2011], anaphora resolution [Marques-2013], and slot-filling [Carapinha-2013]. Besides the basic syntactic parsing, XIP also performs some preliminary semantic analysis: it contains a named entity recognition model [Romao-2007], [Loureiro-2007], [Santos-2010], [Oliveira-2010] to identify the main NE categories (PERSON, ORGANIZATION, PLACE, etc.), including time expressions [Hagege-et-al-2008], [Baptista-et-al-2008], [Hagege-et-al-2009], [Hagege-et-al-2010]. Using information from ViPEr [Baptista-2012], a lexicon-grammar of European Portuguese verbal constructions, XIP also performs an hybrid rule-based and statistical word sense disambiguation of verbs [Travanca-2013], assigning each instance to its correct word-sense. Finally, a semantic role labelling model [Talhadas-2014] assigns the arguments and complements of full verbs their corresponding role (from a set of 37 semantic roles: AGENT, PATIENT, etc.).

According to Mamede *et al.* [Mamede-et-al-2012],

> "Since its initial assembly in 2007, the STRING NLP chain has been subject to continuous improvement in several of its modules, and particularly the conversion between them, yielding a 4 ms/word debit. Using the L2F 100 CPU GRID, it is now possible to process the entire CETEMPúblico under 7 hours." [Mamede-et-al-2012, p. 2].

3.2 Dependency Rules in XIP

As part of the parsing process, XIP executes *dependency rules*. Dependency rules extract different types of dependencies between nodes of the sentence chunking tree, namely, the chunk heads (as it will be done in this project). Dependencies can thus be viewed as equivalent to (or representing) the syntactic relations holding between different elements in a sentence. Notice that, conventionally, in all dependencies, the first argument is the governor and the second one is the dependent element. In XIP, the arity of dependencies can be set to zero, one or more arguments, but in most cases dependencies hold between just two arguments.

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)`. Some dependencies can also be seen as auxiliary dependencies and 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 (4):

(4) *O Pedro confia em_o João*[2] (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:
— first, 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`;
— 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`.

[2]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.

This type of (auxiliary) dependency can be useful, for example, for further rules to act upon. To illustrate this interaction between rules and dependencies, consider, for instance, the rule that could now be devised and that, for a sentence such as (5), would extract a complement dependency (and not just an adjunct modifier) between a verb having as a feature the "regency" of preposition *de* 'of' and the PP introduced by that preposition.[3] In other words, if a verb governs a PP introduced by *de* 'of' and there is such a phrase in the sentence already linked to that verb, then extract an (essential) complement of that verb.

(5) *A Ana gostou do meu mais belo livro* 'Ana liked my most beautiful book'

```
|PP#1|
IF ( VDOMAIN (#2,#3[prepDE]) &
     HEAD(#4,#1) & PREPD(#4,#5[lemma:de]) &
     ^MOD(#3,#4) & ~COMPL(#3,#4)
   )
     COMPL(#3,#4)
```

This rule first defines that the main verb #3 selects the preposition *de* 'of' to introduce one of its complement positions (feature prepDE); then, it verifies if a given PP#1 is introduced by that very preposition; to this, the HEAD dependency is used to determine the relation between the PP and its head and the PREPD dependency, for the relation between the preposition and this PP's head; next, the system verifies if a general-purpose MOD dependency has already been extracted between the main verb and that PP's head; this dependency is signaled by a charat symbol '^' to indicate that this dependency will be changed into another one; such condition prevents other PP, if unrelated to the verb, to be affected by the rule; and, finally, the system verifies if no COMPL dependency has been extracted yet, which is marked by the tilde '~' symbol; when all structural and dependency conditions are met, the system extracts the adequate COMPL dependency between the head of the PP and the main verb, irrespective of the length of the PP constituent, or the number of intermediate constituents that may exist between them.

In this section we have presented and illustrated the main features of the dependency rules used in XIP to extract the syntactic relations between the elements of a given sentence. For this project, though whole-part relations are mostly of semantic nature, they rely (and are extracted based) on syntactic dependencies and distributional patterns, so we extract those relations using this same type of dependency rules. In the next section, we present the dependency rules used to extract whole-part relations.

3.3 The Basic Whole-Part Dependencies Involving Body-Part Nouns

This section describes the way the basic whole-part dependencies involving *Nbp* are extracted in the Portuguese grammar for the XIP parser. To this end, a new module of the rule-based grammar was built, which corresponds to a new file (dependencyBodyParts.xip) in the XIP file structure. This file is the first step towards a meronymy extraction module for Portuguese, and it contains most of the rules required for this project.

[3] We use the traditional terminology here. One could also say that the verb subcategorizes a PP introduced by preposition *de* 'of'.

Occasionally, other parts of the grammar and some files in the lexicons had to be adapted, as new features needed to be defined or new lexical entries were required, or some existing entries required adding new features. More rarely, other dependency rules' 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 will target, this section is organized in such a way so that the the more simple cases are illustrated first and then the more complex situations follow. Nevertheless, whenever possible, we tried to keep the order in which processing takes place, so that the reader could get a clearer picture of the topics complexity. Thus, this section is structured as follows: first, the determinative complements are presented (subsection 3.3.1), then the dative complements (3.3.2), followed by subject *Nbp* with determinative complements (3.3.3) and dative clitic pronouns (3.3.4); the possessive pronouns are next (3.3.5), followed by the dative restructuring of subject *Nbp* determinative complements (3.3.6); the section ends with the (apparently) simpler cases of human subject with *Nbp* direct object (3.3.7) and a prepositional phrase with *Nbp* in a sentence with a human subject (3.3.8).

The entire set of rules developed in this dissertation project is presented in Appendix A.

3.3.1 Determinative Complements

The first example (6) is a simple case where there is a determinative `PP`, complement *de* 'of' *N* of the *Nbp*, so that the meronymy is overtly expressed in the text:

(6) *O Pedro partiu o braço do João* 'Pedro broke the arm of João'

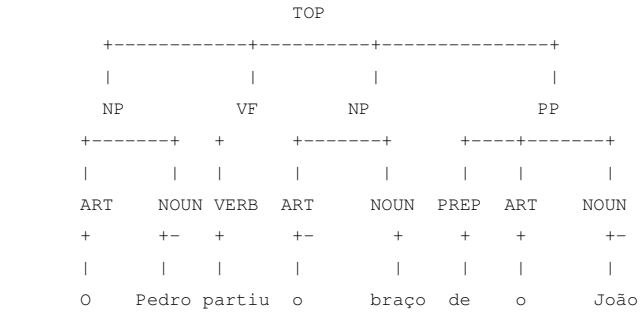
The rule that captures the meronymy relation between *João* and *braço* 'arm':

```
//Example: O Pedro partiu o braço do João. ---> WHOLE-PART(João,braço)
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 first line is a comment, and it is ignored by the parser. In this comment, an example is provided, the same as sentence (6), and the intended output is shown. This has been systematically done to help to build, maintain and correct the rules. The rule itself 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`); these variables are associated to semantic features: the feature `UMB-Anatomical-human` represents all *Nbp* that can be associated to humans, while the feature `human` is a generic feature that designates all nouns that can be assigned human properties. This also applies to named entities referring to people. 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'. Next, the 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; 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 output of the system on sentence (6) is given in Fig. 3.2 (only the relevant dependencies are displayed).




```
PREPD(João,de)
MOD_POST(braço,João)
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
WHOLE-PART(João,braço)}
0>TOP{NP{O Pedro} VF{partiu} NP{o braço} PP{de o João}}
```

Figure 3.2: WHOLE-PART relations for the sentence *O Pedro partiu o braço do João* 'Pedro broke the arm of João'.

The next sentence (7) demonstrates the case where a whole-part dependency should be build between an oblique pronoun determining an *Nbp*. This pronoun is the result of the reduction of a human determinative complement, like the one shown in the previous example.

(7) *O Pedro partiu o braço dele* (lit: Pedro broke the arm of him) 'Pedro broke his arm'

The rule that captures the meronymy relation between *ele* 'he' and *braço* 'arm' in sentence (7):

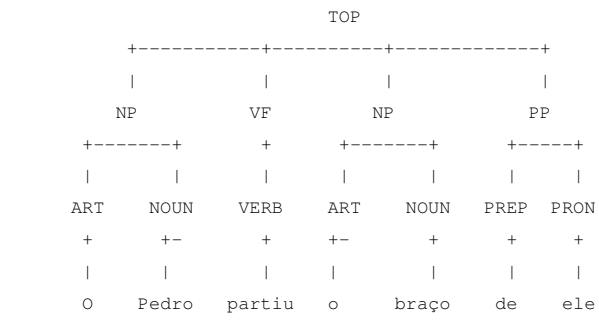
```
IF( MOD[POST](#2[UMB-Anatomical-human],#1[obl,3p]) &
    PREPD(#1,?[lemma:de]) &
    CDIR[POST](#3,#2) &
    ~WHOLE-PART(#1,#2)
  )
    WHOLE-PART(#1,#2)
```

This rule verifies, first, if there is a [MOD]ifier dependency between an *Nbp* and an *oblique* (obl), third-person (3p) pronoun (in this example *ele* 'he'[4]), which must be introduced by the preposition *de* 'of' (PREPD); then, similarly to example (6), a constraint is defined that the *Nbp* must be a direct object (CDIR) of a given verb (variable #3); and, finally, if there is still no WHOLE-PART dependency between the pronoun and the *Nbp*; then, the parser builds this dependency.

The output of the system on sentence (7) is given in Fig. 3.3.

Notice that, in theory, the subject NP of sentence (7) could also function as the antecedent of the oblique pronoun. This interpretation is grammatically valid, though a bit redundant, as the *Nbp* occurs

[4]Other person oblique forms are not allowed in Portuguese. Instead, a possessive pronoun is used.




```
MOD_POST(braço,ele)
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
WHOLE-PART(ele,braço)
0>TOP{NP{O Pedro} VF{partiu} NP{o braço} PP{de ele}}
```

Figure 3.3: WHOLE-PART relations for the sentence *O Pedro partiu o braço dele* (lit: Pedro broke the arm of him) 'Pedro broke his arm'.

in direct object position. In this case, we have decided to ignore it, and *force* the coreference between the *Nbp* and the oblique pronoun, which corresponds to a *preferable* (*i.e.*, more likely) interpretation of the sentence. By doing this, we postpone the Anaphora Resolution (AR) step, which, in fact, takes place after whole-part dependencies are extracted [Marques-2013]. The AR module can then take into account whether or not the presence of an explicit subject (among other factors), influences the anaphoric interpretation of the oblique pronoun.

More rarely, the *Nbp* can occur as a determinative complement of a human noun, as in *a rapariga de olhos azuis* 'the girl with blue eyes', *a senhora de pernas elegantes* 'the lady with elegant legs', *o homem de ventre rotundo* 'the man with round belly', etc. A single rule was also built to capture this cases:

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

The output for the sentence (8) is given in Fig. 3.4.

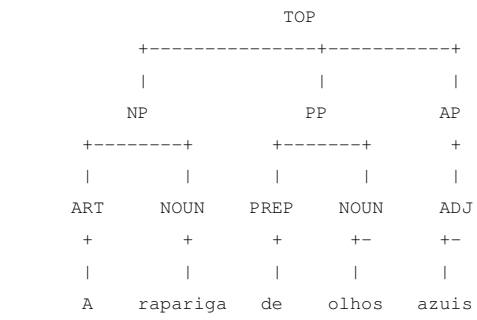
(8) *A rapariga de olhos azuis* 'The girl with blue eyes'

3.3.2 Dative Complements

The next example (9) 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*.[5]

(9) *O Pedro partiu o braço ao João* 'Pedro broke the arm to João'

[5]Syntactically, this dative complement can be analysed as the result from the dative restructuring ([Leclere-1995], [Baptista-1997a]) of the *Nbp de Nhum* base phrase.




```
PREPD(olhos,de)
MOD_POST(olhos,azuis)
MOD_POST(rapariga,olhos)
WHOLE-PART(rapariga,olhos)
0>TOP{NP{A rapariga} PP{de olhos} AP{azuis}}
```

Figure 3.4: `WHOLE-PART` relations for the sentence *A rapariga de olhos azuis* 'The girl with blue eyes'.

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

```
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)
```

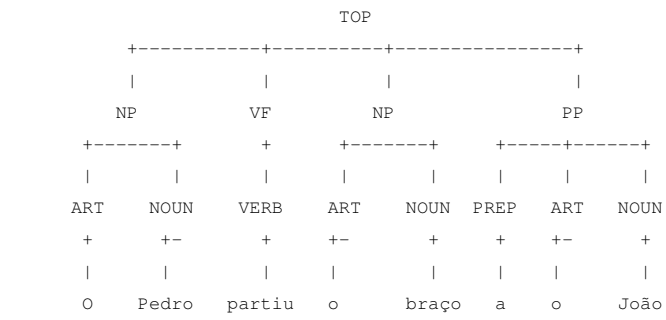
This rule reads as follows: first, the default `MOD[ifier]` dependency between the verb and the prepositional phrase *a Nhum* 'to Nhum' has to be changed (and it is, thus, preceded by the symbol '^') into an indirect complement (`CINDIR`) dependency; to do this, the system verifies if there is a syntactic 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` dpendencies; 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 output of the system from sentence (9) is given in Fig. 3.5.

3.3.3 Subject *Nbp* and Determinative Complements

In the previous cases, the *Npb* was the direct object, which is by far the most frequent situation in texts. However, an *Nbp* can also be placed as the subject of a verb and, so, a similar set of rules is necessary to capture this situation. In the next sentence (10), the meronymy holds between the *Nbp* and a determinative complement with a human noun.

(10) *O braço do Pedro está partido* (lit: The arm of Pedro is broken) 'Pedro's arm is broken'




```
PREPD(João,a)
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
CINDIR_POST(partiu,João)
WHOLE-PART(João,braço)
0>TOP{NP{O Pedro} VF{partiu} NP{o braço} PP{a o João}}
```

Figure 3.5: `WHOLE-PART` relations for the sentence *O Pedro partiu o braço ao João* 'Pedro broke the arm to João'.

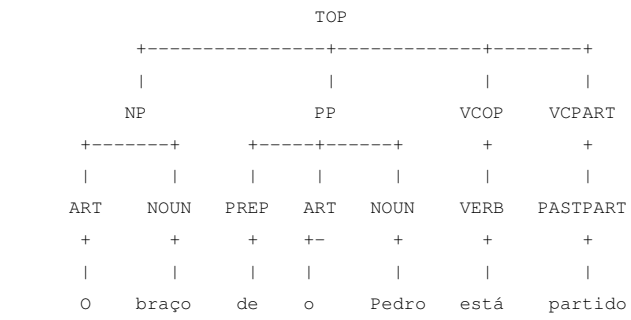
The general rule, below, is sufficient to capture this relation. However, notice that this rule must only be applied *after* the rule accompanying example (6) has taken place, as it makes no reference to the subject position of the *Nbp*. In other words, direct object *Nbp* must first be captured, in order to prevent incorrect extraction of whole-part relations. As other rule-based systems, rule order is one of the features of the XIP parser that can be used to simplify the building of the grammar. Still, in the rule below, we ensure that no `WHOLE-PART` dependency has been previously extracted, not only between the *Nbp* and its human determinative complement, but also between that *Nbp* and any other syntactic node (variable `#4`), or between the human noun and any other *Npb* (variable `#3`):

```
IF( MOD[POST](#2[UMB-Anatomical-human],#1[human]) &
    PREPD(#1,?[lemma:de]) &
    ~WHOLE-PART(#1,#2) &
    ~WHOLE-PART(#1,#3) &
    ~WHOLE-PART(#4,#2)
  )
    WHOLE-PART(#1,#2)
```

The output of the system from sentence (10) is given in Fig. 3.6.

A similar rule, below, is made for the case (example (11)) of a subject *Nbp* with an oblique determinative complement, as in example (10). This rule is almost the same as the one given for the sentence (7), but since it takes place at a later step of the analysis, namely, after the case of a direct object *Nbp* has been taken care of, the rule can be simpler.

(11) *O braço dele está partido* (lit: The arm of him is broken) 'His arm is broken'

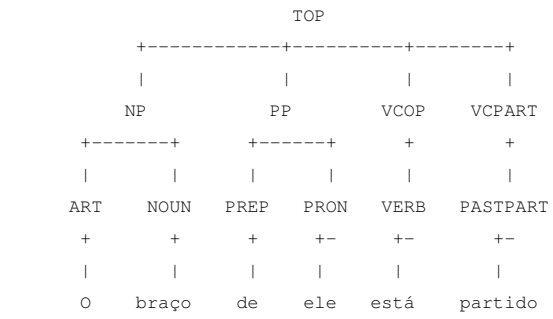



```
VLINK(está,partido)
VDOMAIN(está,partido)
MOD_POST(braço,Pedro)
SUBJ_PRE(partido,braço)
WHOLE-PART(Pedro,braço)
0>TOP{NP{O braço} PP{de o Pedro} VCOP{está} VCPART{partido}}
```

Figure 3.6: `WHOLE-PART` relations for the sentence *O braço do Pedro está partido* (lit: The arm of Pedro is broken) 'Pedro's arm is broken'.

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

The output of the system from sentence (11) is given in Fig. 3.7.




```
PREPD(ele,de)
VLINK(está,partido)
VDOMAIN(está,partido)
MOD_POST(braço,ele)
SUBJ_PRE(partido,braço)
WHOLE-PART(ele,braço)
0>TOP{NP{O braço} PP{de ele} VCOP{está} VCPART{partido}}
```

Figure 3.7: `WHOLE-PART` relations for the sentence *O braço dele está partido* (lit: The arm of him is broken) 'His arm is broken'.

3.3.4 Dative Pronouns

In the next example (12), the dative complement is pronominalized by the dative clitic pronoun *-lhe* 'him'.

(12) *O Pedro partiu-lhe o braço* 'Pedro broke him the arm'

In Portuguese, the dative pronoun incorporates the preposition *a* 'to' that introduces indirect objects. When the *Nbp* is the direct object of the main verb, there is no ambiguity regarding the meronymy relation between the *Nbp* and the dative pronoun. However, at this stage of the parsing, no indirect object has been built yet, due to the fact that dative pronouns can fulfil other syntactic-semantic functions (benefactive or "politeness" datives). Because of this, the dative pronoun is provisory parsed as a special type of `[MOD]`ifier, with a `[DAT]`ive flag. In these cases, the system captures the `WHOLE-PART` relation and changes the `MOD[DAT]` into an indirect complement `CINDIR`, as in example (9). This is carried out by the following rule[6]:

```
IF( ^MOD[DAT](#3,#1[dat,cli]) &
    SUBJ[PRE] (#3,#4) &
    CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    ~SUBJ[elips](#3,#5) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    CINDIR[DAT=~](#3,#1),
    WHOLE-PART(#1,#2)
```

This rule verifies that there is a `MOD[DAT]` involving a `[cli]`tic `[dat]`ive pronoun and that there is an *Nbp* as direct object; if no `CINDIR` dependency has been calculated yet for the pronoun, nor any `WHOLE-PART` relation involving the *Nbp* and the pronoun, then these two dependencies are built. The `DAT` flag, which makes sense in the parsing process to signal these special, yet-unsolved, dative modifier is also zeroed. Two supplementary constraints were added, to enforce the presence of an explicit subject of the verb, as long as this is not a dummy pronoun, that the parser introduces for elliptic subjects (see examples (17a)-(17b), below).

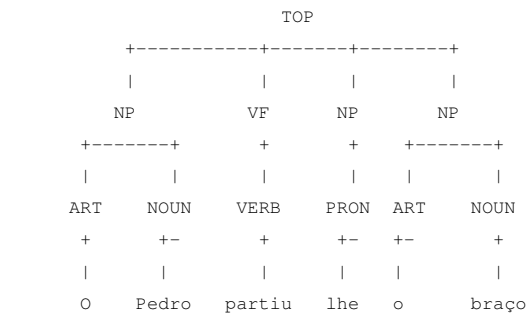
The output of the system on sentence (12) is given in Fig. 3.8.

Since dative pronouns are *clitic* pronouns, in Portuguese, they can be fronted to the left of the verb, like in example (13) under several syntactic conditions (subordinate clauses, negation, etc.).

(13) *O Pedro não lhe partiu o braço* (lit: Pedro did_not to-him broke the arm)
'Pedro did not break his arm'

The fronted clitic pronoun is previously captured by another auxiliary dependency `CLITIC` with the flag `PRE`. The rule that captures this fronted dative pronoun is, otherwise, similar to the previous one,

[6]The condition `~PREPD(#5,#7[lemma:de]) & ~MOD(#2,#5)` has been added during the error analysis.




```
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
CINDIR(partiu,lhe)
CLITIC_POST(partiu,lhe)
WHOLE-PART(lhe,braço)
0>TOP{NP{O Pedro} VF{partiu} NP{lhe} NP{o braço}}
```

Figure 3.8: `WHOLE-PART` relations for the sentence *O Pedro partiu-lhe o braço* 'Pedro broke him the arm'.

as it is shown below[7]:

```
IF( CLITIC[PRE](#3,#1[dat]) &
    CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    SUBJ[PRE](#3,#4) &
    ~SUBJ[elips](#3,#5) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    CINDIR(#3,#1),
    WHOLE-PART(#1,#2)
```

The output of the system on sentence (13) is given in Fig. 3.9.

3.3.5 Possessive Pronouns

Though determinative possessive 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.'

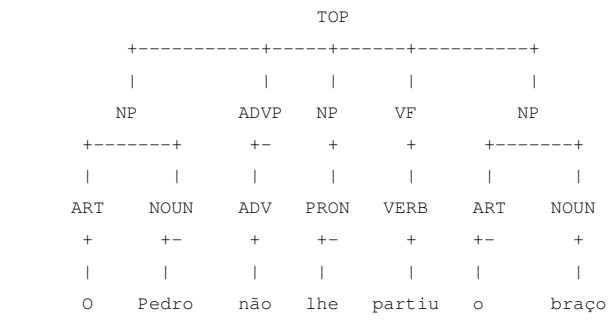
os teus braços 'your_2nd-sg.masc.pl. arm_masc.pl.'

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:

(14) *O Pedro partiu o seu braço* 'Pedro broke his arm'

[7] The condition `~PREPD(#6,#7[lemma:de]) & ~MOD(#2,#6)` has been added during the error analysis.




```
MOD_PRE_NEG(partiu,não)
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
CINDIR(partiu,lhe)
CLITIC_PRE(partiu,lhe)
WHOLE-PART(lhe,braço)
0>TOP{NP{O Pedro} ADVP{não} NP{lhe} VF{partiu} NP{o braço}}
```

Figure 3.9: `WHOLE-PART` relations for the sentence *O Pedro não lhe partiu o braço* (lit: Pedro did_not to-him break the arm) 'Pedro did not break his arm'.

The rule that captures the meronymy relation between the possessive pronoun *seu* 'his' and *braço* 'arm', sentence (14):

```
IF( POSS[PRE](#2[UMB-Anatomical-human],#1[poss]) &
    ~WHOLE-PART(#1,#2) &
  )
    WHOLE-PART(#1,#2)
```

This rule reads as follows: if there is a `[POSS]`essive dependency between an *Nbp* and a possessive pronoun, in this case, the possessive is *seu* 'his'; and, if there is still no `WHOLE-PART` dependency between the possessive pronoun and the *Nbp*; then, the parser builds this dependency.

The output of the system on sentence (14) is given in Fig. 3.10.

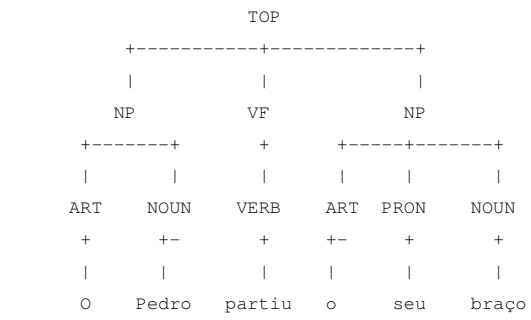
3.3.6 Complex Dative Restructuring with Subject *Nbp*

The next case constitute a complex situation involving the dative restructuring of determinative complements (see section 3.3.4, above). In Portuguese, certain verbs, like *doer* 'hurt', select a subject *Nbp* and its determinative human complement is normally restructured into a dative pronoun (hence the dubious acceptability of sentences (15a)-(15b)).

(15a) *?Os braços do Pedro doem* (lit: The arms of Pedro hurt) 'Pedro's arms hurt'

(15b) *?Os braços doem ao Pedro* (lit: The arms hurt to Pedro) 'Pedro's arms hurt'

(15c) *Os braços doem-lhe* (lit: The arms hurt him) 'His arms are hurting'



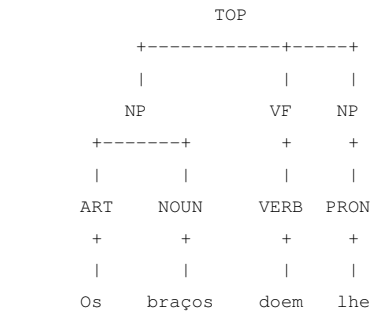

```
POSS_PRE(braço,seu)
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
WHOLE-PART(seu,braço)
0>TOP{NP{O Pedro} VF{partiu} NP{o seu braço}}
```

Figure 3.10: `WHOLE-PART` relations for the sentence *O Pedro partiu o seu braço* 'Pedro broke his arm'.

As the *Nbp* is the subject, the coreference between the dative pronoun is captured by the rule illustrated below:

```
IF( ^MOD[DAT](#3,#1[dat,cli]) &
    SUBJ[PRE](#3,#2[UMB-Anatomical-human]) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    CINDIR[DAT=~](#3,#1),
    WHOLE-PART(#1,#2)
```

The output of this rule is presented in Fig. 3.11.




```
SUBJ_PRE(doem,braços)
CINDIR(doem,lhe)
CLITIC_POST(doem,lhe)
WHOLE-PART(me,braços)
0>TOP{NP{Os braços} VF{doem} NP{lhe}}
```

Figure 3.11: `WHOLE-PART` relations for the sentence *Os braços doem-lhe* (lit: The arms hurt him) 'His arms are hurting'.

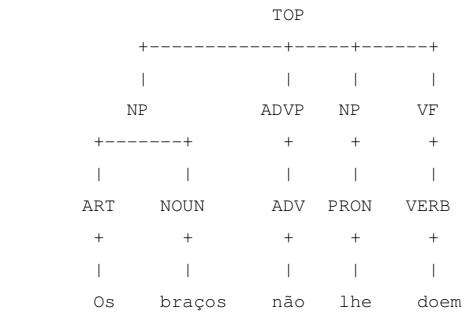
Naturally, dative clitic pronoun fronting has also to be taken into consideration as in sentence (16), in much the same way as it was done before in section 3.3.4).

(16) *Os braços não lhe doem* (lit: The arms do_not to-him hurt) 'His arms are not hurting'

The rule that capture this relation:

```
IF( ^CLITIC[PRE](#3,#1[dat]) &
    SUBJ[PRE](#3,#2[UMB-Anatomical-human]) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    CINDIR(#3,#1),
    WHOLE-PART(#1,#2)
```

The output of this rule is presented in Fig. 3.12.




```
MOD_PRE_NEG(doem,não)
SUBJ_PRE(doem,braços)
CINDIR_PRE(doem,lhe)
WHOLE-PART(lhe,braços)
0>TOP{NP{Os braços} ADVP{não} NP{lhe} VF{doem}}
```

Figure 3.12: `WHOLE-PART` relations for the sentence *Os braços não lhe doem* (lit: The arms do_not to-him hurt) 'His arms are not hurting'.

However, in this type of sentences, a subject inversion can also take place, like in examples (17a)-(17b).

(17a) *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'

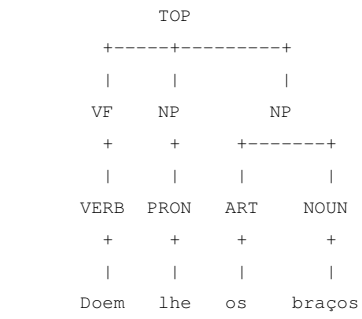
(17b) *Não lhe doem os braços* (lit: Not to-him are_hurting the arms) 'His arms are not hurting'

This yields another new, not previously considered, syntactic configuration, which is captured by the following set of rules:

```
(i)
IF( ^MOD[DAT](#3,#1[dat,cli]) &
    CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    SUBJ[ELIPS](#3,#4) &
    ~SUBJ(#3,#2) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    CINDIR[DAT=~](#3,#1),
    SUBJ[POST=+](#3,#2),
    WHOLE-PART(#1,#2)

(ii)
IF( CINDIR(#3,#1) &
    ^CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    SUBJ[POST](#3,#2) &
    WHOLE-PART(#1,#2)
  )
    ~
```

In these two rules, numbered (i) and (ii), the system, first, matches an initial, incorrect parse of the sentence (17a), and then it proceeds to *correct* the dependencies that were inadequately extracted, until the final, adequate parse is achieved. However, as XIP can only *modify* one dependency per rule, this process involves splitting the corrections into several steps. To better understand the process, let us consider the initial, incorrect parse of sentence (17a) shown in Fig. 3.13




```
MOD_DAT(Doem,lhe)
SUBJ_ELIPS(Doem,Eles)
CDIR_POST(Doem,braços)
CLITIC_POST(Doem,lhe)
0>TOP{VF{Doem} NP{lhe} NP{os braços}}
```

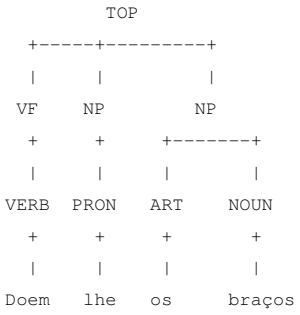
Figure 3.13: Initial, incorrect parse for the sentence: *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'.

Based on the auxiliary dependency `CLITIC` a `MOD` dependency is extracted between the verb and the dative pronoun *lhe* 'to-him', and this is given the `DAT` feature. Later on, this `MOD_DAT` dependency will be changed into a `CINDIR` dependency (indirect object). Since there is no explicit subject, an elliptic subject (`SUB_ELIPS`) is first calculated and a dummy nominative pronoun *eles* 'they' is inserted; the *Nbp* is then wrongly parsed as a direct object (`CDIR`) of the verb.

From this initial parse, rule (i), firstly matches all the dependencies above, and verifies if there is still no SUBJ dependency between the verb and the *Nbp*, nor a CINDIR between the verb and the dative pronoun, nor a WHOLE-PART relation between the dative pronoun and the *Nbp*; once all these verifications are done, the rule proceeds to correct the MOD dependency into a CINDIR, extract a new SUBJECT dependency between the verb and the *Nbp*, and establish the WHOLE-PART relation between the pronoun and the *Nbp*. Notice that at this stage there are two SUBJ dependencies, one for the elliptic subject with the dummy pronoun and this new one, with the *Nbp*. This duplication is solved by removing the elliptic subject using a general rule based on word-order:

```
IF ( ^SUBJ(#1,#2) & SUBJ(#1,#3) & #2 < #3 & ~(COORD(#4,#2) & COORD(#4,#3)))
~
```

This rule is interpreted as follows: if there are two SUBJ dependencies on the same verb, and if the first subject appears before the second one (and there is no coordination between the two), then the first SUBJ dependency is deleted. The outcome of this parsing step is shown in Fig. 3.14.

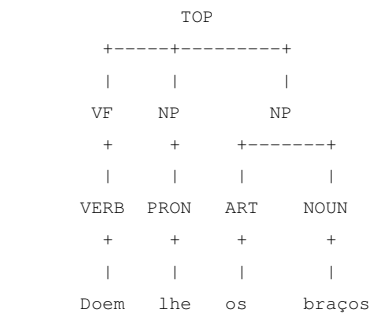



```
SUBJ_POST(Doem,braços)
CDIR_POST(Doem,braços)
CINDIR(Doem,lhe)
CLITIC_POST(Doem,lhe)
WHOLE-PART(lhe,braços)
0>TOP{VF{Doem} NP{lhe} NP{os braços}}
```

Figure 3.14: First step of the parsing for the sentence *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'.

As one can see, SUBJ[ELIPS] has been removed at this stage, but the *Nbp* is still parsed as a direct object (CDIR). This is where the rule (ii) comes into play: it removes the CDIR[POST] dependency between the verb and the *Nbp*, as long as there is a SUBJ between them and a WHOLE-PART dependency has already been extracted for the *Nbp*. The output is now the correct parse of the sentence, and it is shown in Fig. 3.15.

A similar rule has to be done for the sentence (17b), where the negation entails the fronting of the dative pronoun. This rule, which captures the meronymy relation between *lhe* 'to-him' and *braços* 'arms',



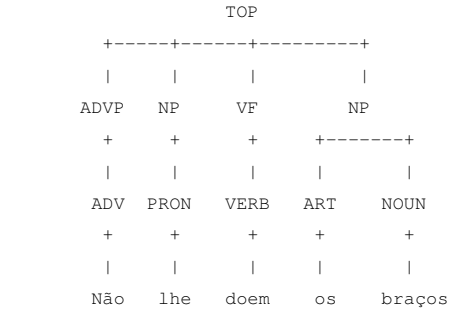

```
SUBJ_POST(Doem,braços)
CINDIR(Doem,lhe)
CLITIC_POST(Doem,lhe)
WHOLE-PART(lhe,braços)
0>TOP{VF{Doem} NP{lhe} NP{os braços}}
```

Figure 3.15: Correct parsing for the sentence *Doem-lhe os braços* (lit: Are_hurting to-him the arms) 'His arms are hurting'.

is given below:

```
IF( CLITIC[PRE](#3,#1[dat]) &
    ^CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    ~CINDIR(#3,#1) &
    ~WHOLE-PART(#1,#2)
  )
    SUBJ[POST=+](#3,#2),
    CINDIR(#3,#1),
    WHOLE-PART(#1,#2)
```

The correct parse is provided in Fig. 3.16.




```
MOD_PRE_NEG(doem,Não)
SUBJ_POST(doem,braços)
CINDIR(doem,lhe)
CLITIC_PRE(doem,lhe)
WHOLE-PART(lhe,braços)
0>TOP{ADVP{Não} NP{lhe} VF{doem} NP{os braços}}
```

Figure 3.16: `WHOLE-PART` relations for the sentence *Não lhe doem os braços* (lit: Not to-him are_hurting the arms) 'His arms are not hurting'.

3.3.7 Subject *Nhum* and Direct Object *Nbp*

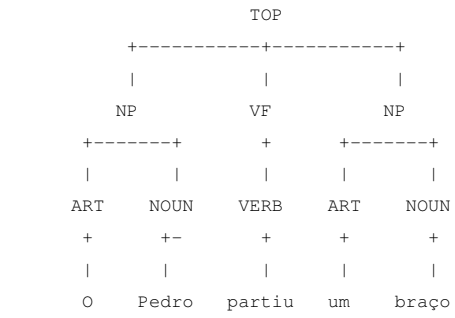
In example (18), we present the (apparently) more simple case of a sentence with just a human subject and an *Nbp* direct object:

(18) *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 (18) are preferably interpreted as holding a whole-part relation between the human subject and the object *Nbp*. Notice that the negative conditions stated above imply that the rule to process this case can only be fired after all the previous rules were tested, hence, this rule appears after all the others in the corresponding grammar file. Such rule is, after all, rather simple[8]:

```
IF( SUBJ[PRE](#3,#1[human]) &
    CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    ~WHOLE-PART(#1,#2) &
    ~WHOLE-PART(#4,#2)
  )
    WHOLE-PART(#1,#2)
```

This rule reads: if there is a subject and a direct complement dependency 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 result of this rule is shown in Fig. 3.17.




```
SUBJ_PRE(partiu,Pedro)
CDIR_POST(partiu,braço)
WHOLE-PART(Pedro,braço)
0>TOP{NP{O Pedro} VF{partiu} NP{um braço}}
```

Figure 3.17: `WHOLE-PART` relations for the sentence *O Pedro partiu um braço* 'Pedro broke an arm'.

3.3.8 Subject *Nhum* and Prepositional Phrase with *Nbp*

In this subsection we deal with the cases where an *Nbp* is in a prepositional phrase in a sentence with a human subject.

[8] The condition `~PREPD(#5,#7[lemma:de]) & ~MOD(#2,#5)` has been added during the error analysis.

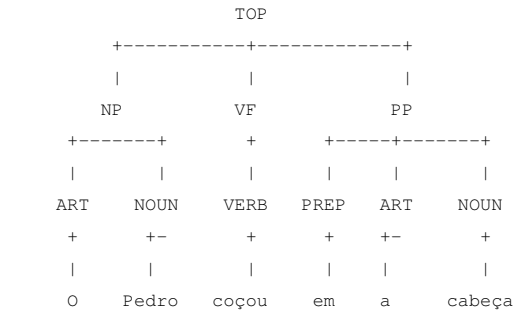
In example (19), there is no other complement the *Nbp* can be related to, so a meronymy relation should be established between the human subject and the *Nbp*. Because of the very constrained context, the corresponding rule has to explicitly state all the possible constituents that must not occur to allow the rule to be fired.

(19) *O Pedro coçou na cabeça* (lit: Pedro scratched on the head) 'Pedro scratched the head'

The rule that captures the meronymy relation between *Pedro* and *cabeça* 'head', sentence (19), is the following[9]:

```
IF( MOD[post](#1,#2[UMB-Anatomical-human]) &
    SUBJ[pre](#1,#3[human]) &
    ~WHOLE-PART(#3,#2) &
    ~POSS[pre](#2,#4[poss]) &
    ~MOD[post](#2,#5[human]) & ~PREPD(#5,#6[lemma:de]) &
    ~CDIR(#1,#7[human]) &
    ~CDIR(#1,#8[acc]) &
    ~CINDIR(#1,#9) &
    ~MOD[dat](#1,#10)
  )
    WHOLE-PART(#3,#2)
```

The output of the system is shown in Fig. 3.18.




```
PREPD(cabeça,em)
MOD_POST(coçou,cabeça)
SUBJ_PRE(coçou,Pedro)
WHOLE-PART(Pedro,cabeça)
0>TOP{NP{O Pedro} VF{coçou} PP{em a cabeça}}
```

Figure 3.18: `WHOLE-PART` relations for the sentence *O Pedro coçou na cabeça* (lit: Pedro scratched on the head) 'Pedro scratched the head'.

In the next case (example (20)), the sentence shows a (dative) prepositional phrase, with a human noun, a situation that had not yet been captured in any of the previous rules.

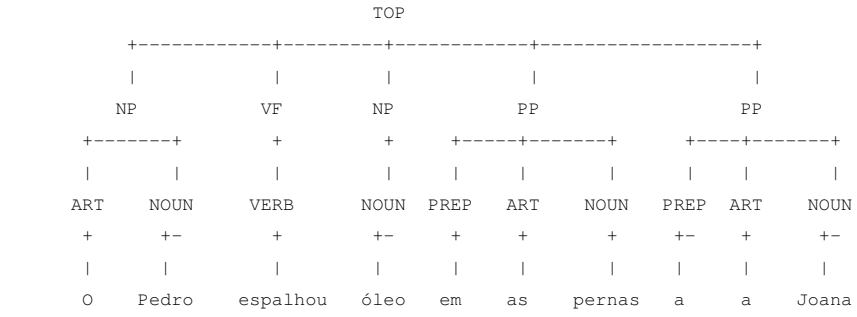
(20) *O Pedro espalhou óleo nas pernas à Joana* 'Pedro spread oil on the legs of Joana'

[9]During the error analysis, the line `~MOD[post](#2,#5[human]) & ~PREPD(#5,#6[lemma:de])` has been changed to `(~MOD[post](#2,#5[human]) || ~PREPD(#5,#6[lemma:de]) )`.

The rule that captures the meronymy relation between *Joana* and *pernas* 'legs', sentence (20):

```
IF( MOD[post](#1,#2[UMB-Anatomical-human]) & PREPD(#2,#5[lemma:em]) &
    MOD[post](#1,#3[human]) & PREPD(#3,#6[lemma:a]) &
    SUBJ[pre](#1,#4[human]) &
    ~WHOLE-PART(#3,#2) &
    ~POSS[pre](#2,#7[poss]) &
    ~CDIR(#1,#10[human]) &
    ~CINDIR(#1,#11)
  )
    WHOLE-PART(#3,#2)
```

The output of the system is presented in Fig. 3.19.




```
PREPD(pernas,em)
PREPD(Joana,a)
MOD_POST(espalhou,pernas)
MOD_POST(espalhou,Joana)
SUBJ_PRE(espalhou,Pedro)
CDIR_POST(espalhou,óleo)
WHOLE-PART(Joana,pernas)
0>TOP{NP{O Pedro} VF{espalhou} NP{óleo} PP{em as pernas} PP{a a Joana}}
```

Figure 3.19: WHOLE-PART relations for the sentence *O Pedro espalhou óleo nas pernas à Joana* 'Pedro spread oil on the legs of Joana'.

The other cases with an *Nbp* in a prepositional phrase involve clitics, usually a dative pronoun issued from the restructuring of a determinative complement of the *Nbp*. However, the clitic can also be a reflexive pronoun, if the action of the subject falls upon itself.

This situation is complicated by the fact that in Portuguese the accusative, dative, and reflexive pronouns are only different in the 3^{rd} person (accusative: *o* 'him', *a* 'her', *os* 'them', *as* 'them'; dative: *lhe* 'him/her', *lhes* 'them'; reflexive: *se* 'himself/herself/itself'); while the 1^{st} and the 2^{nd} person have the same form (1^{st}-sg. *me* 'me', 2^{nd}-sg. *te* 'you', 1^{st}-pl. *nos* 'us', 2^{nd}-pl. *vos* 'you'). In view of this, a statistical disambiguation module was developed in STRING specifically to deal with this 4 ambiguous forms. Precision of this module is very high, so at the stage of processing where the meronymy module comes into play, we consider that the disambiguation issue is solved. We first deal with the reflexive clitic pronoun *-se* 'himself' (example (21)).

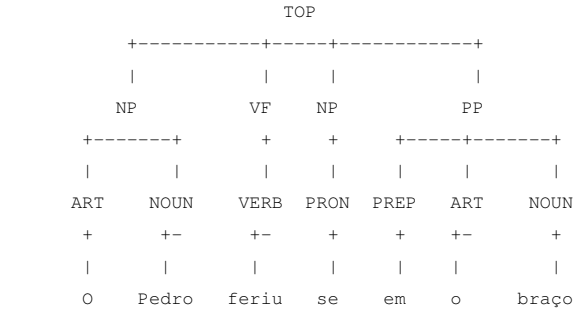
(21) *O Pedro feriu-se no braço* (lit: Pedro wounded himself in the arm) 'Pedro wounded his arm'

The rule that captures the meronymy relation between *Pedro* and *cabeça* 'arm', sentence (21), is the

following[10]:

```
IF(  CLITIC(#3,#1[cli,ref]) &
     SUBJ[PRE](#3,#6) &
     MOD[POST](#3,#2[UMB-Anatomical-human]) &
     PREPD(#2,#4[lemma:em]) &
     ~WHOLE-PART(#6,#2)
  )
     WHOLE-PART(#6,#2)
```

The output of the system is presented in Fig. 3.20.




```
PREPD(braço,em)
MOD_POST(feriu,braço)
SUBJ_PRE(feriu,Pedro)
CDIR_POST(feriu,se)
CLITIC_POST(feriu,se)
WHOLE-PART(Pedro,braço)
0>TOP{NP{O Pedro} VF{feriu} NP{se} PP{em o braço}}
```

Figure 3.20: `WHOLE-PART` relations for the sentence *O Pedro feriu-se no braço* (lit: Pedro wounded himself in the arm) 'Pedro wounded his arm'.

A similar rule has been built for all the non-reflexive pronouns (example (22)).

(22) *O Pedro bateu-me nas pernas* (lit: Pedro hit me in the legs) 'Pedro hit my legs'

The rule that captures the meronymy relation between *me* 'me' and *pernas* 'legs' in sentence (22) is given below[11]:

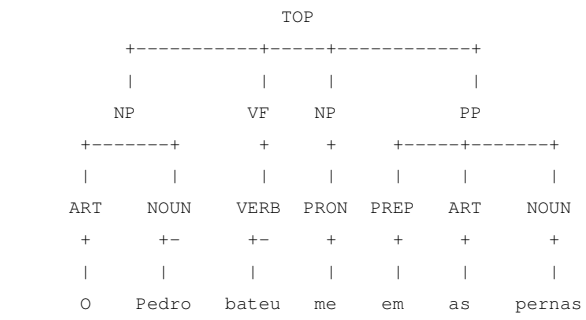
```
IF(  CLITIC(#3,#1[cli,ref:~]) &
     SUBJ[PRE](#3,#6) &
     MOD[POST](#3,#2[UMB-Anatomical-human]) &
     PREPD(#2,#4[lemma:em]) &
     ~WHOLE-PART(#1,#2)
  )
     WHOLE-PART(#1,#2)
```

The output of the system is presented in Fig. 3.21.

The *Nbp* can also appear in a `PP` in sentences with copula or support-verbs, which entail a different set of dependencies (`PREDSUBJ`) (example (23)).

(23) *O Pedro andava de braços cruzados* 'Pedro walked with arms crossed'

[10]The condition `~PREPD(#6,#7[lemma:de]) & ~MOD(#2,#6)` has been added during the error analysis.
[11]The condition `~PREPD(#6,#7[lemma:de]) & ~MOD(#2,#6)` has been added during the error analysis.



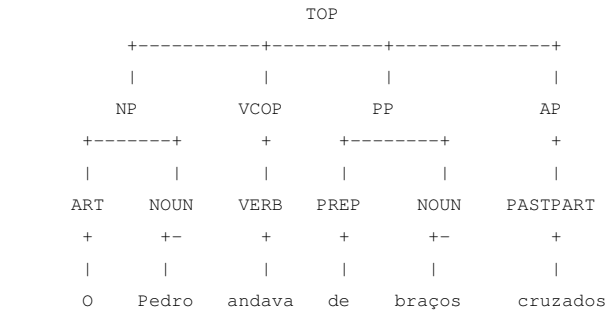

```
PREPD(pernas,em)
MOD_POST(bateu,pernas)
SUBJ_PRE(bateu,Pedro)
CDIR_POST(bateu,me)
CINDIR_POST(bateu,me)
CLITIC_POST(bateu,me)
WHOLE-PART(me,pernas)
0>TOP{NP{O Pedro} VF{bateu} NP{me} PP{em as pernas}}
```

Figure 3.21: `WHOLE-PART` relations for the sentence *O Pedro bateu-me nas pernas* (lit: Pedro hit me in the legs) 'Pedro hit my legs'.

In sentence (23), the verb *andar* 'to walk' has been parsed as a copula (`VCOP`), and for the `PP` with the *Nbp* head the `PREDSUBJ` dependency was extracted. A similar parse would be obtained for support verb construction with *ser* 'to be' and *estar* 'to be'. However, as support verbs are still not captured by the system at this time, only the copula case is addressed here. This type of sentences are matched by following rule:

```
IF(  VDOMAIN(#1,#2[cop]) &
     SUBJ(#2,#3) &
     PREDSUBJ(#2,#4[UMB-Anatomical-human]) &
     MOD[POST](#5[prep],#4) &
     ~WHOLE-PART(#3,#4)
  )
     WHOLE-PART(#3,#4)
```

The output of the system is presented in Fig. 3.22.




```
PREPD(braços,de)
PREDSUBJ(andava,braços)
PREDSUBJ(andava,de)
MOD_POST(de,braços)
MOD_POST(braços,cruzados)
SUBJ_PRE(andava,Pedro)
WHOLE-PART(Pedro,braços)
0>TOP{NP{O Pedro} VCOP{andava} PP{de braços} AP{cruzados}}
```

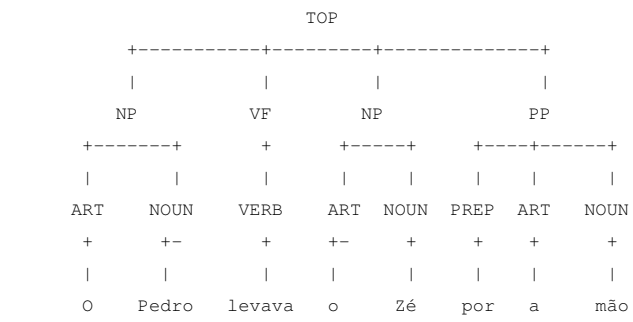
Figure 3.22: `WHOLE-PART` relations for the sentence *O Pedro andava de braços cruzados* 'Pedro walked with arms crossed'.

Finally, a heuristic rule, below, captures all cases where there is a human direct object and a PP with an *Nbp*, like in example (24).

(24) *O Pedro levava o Zé pela mão* 'Pedro led Ze by the hand'

```
IF ( VDOMAIN(#1,#2) &
     CDIR(#2,#3[human]) &
     MOD[post](#2,#4[UMB-Anatomical-human]) &
     ~WHOLE-PART(?,#4) &
     ~WHOLE-PART(#3,#4)
   )
     WHOLE-PART(#3,#4)
```

The output of the system is presented in Fig. 3.23.




```
PREPD(mão,por)
MOD_POST(levava,mão)
SUBJ_PRE(levava,Pedro)
CDIR_POST(levava,Zé)
WHOLE-PART(Zé,mão)
0>TOP{NP{O Pedro} VF{levava} NP{o Zé} PP{por a mão}}
```

Figure 3.23: WHOLE-PART relations for the sentence *O Pedro levava o Zé pela mão* 'Pedro led Ze by the hand'.

This section presented the main cases of whole-part relations in Portuguese, and the rules built to extract them from real texts. The next section addresses the issue of longer sequences of *Nbp* in sentences.

3.4 Determinative Nouns of *Nbp*

3.4.1 Relations between *Nbp*

There may be a relation within the same sentence between different *Nbp*, like in example (25). In this case, the WHOLE-PART relation should be established not only between the subject of the sentence and the *Nbp*, but also between *Nbp* in the sentence.

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

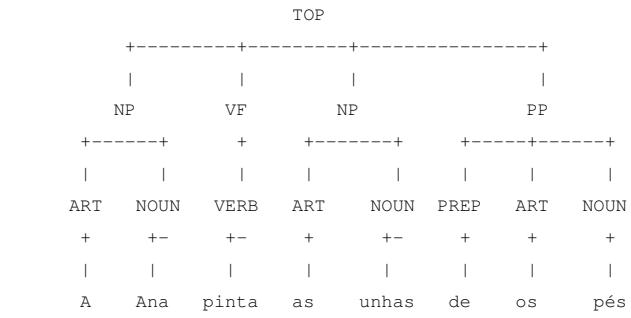
In example (25), there is a meronymic relation between *Ana* and *unhas* 'nails', but also between *pés* 'feet' and *unhas* 'nails', so that two WHOLE-PART relations should be extracted.

The rule that extracts the WHOLE-PART relation between the subject of the sentence and the *Nbp* has already been explained in example (18).

The next rule captures the WHOLE-PART relation between the two *Nbp*, based on the [MOD]ifier dependency among them, and the preposition introducing the complement *Nbp*:

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

The result of this rule is given in Fig. 3.24.




```
PREPD(pés,de)
MOD_POST(unhas,pés)
SUBJ_PRE(pinta,Ana)
CDIR_POST(pinta,unhas)
WHOLE-PART(Ana,unhas)
WHOLE-PART(pés,unhas)
0>TOP{NP{A Ana} VF{pinta} NP{as unhas} PP{de os pés}}
```

Figure 3.24: WHOLE-PART relations for the sentence *A Ana pinta as unhas dos pés* (lit: Ana paints the nails of the feet) 'Ana paints the toenails'.

3.4.2 Relation between *Nbp* and Parts of *Nbp*

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* (*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 (26) illustrates this situation.

(26) *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)` - correct; `WHOLE-PART(língua,ponta)` - correct;

`WHOLE-PART(Pedro,ponta)` - 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*. This can done by the following rule, before the chunking stage:

```
noun[lemma:planta,npart=+], prep[lemma:de], art[lemma:o], noun[lemma:pé].
```

So far, 54 rules were built to associate the *Nbp* with their parts. (Appendix B.1).

As the context that fires these rules is lexically and syntactically defined, it can be further used to narrow down the ambiguity of some adjacent lexical items. For example, the preposition *de* 'of, from' in this context is just a connector, so the locative feature `preplocsource` that was given to it at the lexical tagging phase, in the initial steps of the parsing, can be removed. This can done by the following rule, also before the chunking stage:

```
noun[lemma:peito,npart=+,sem-an=~,sem-am=~], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
```

Other specific rules were also built for the cases where *npart* is involved. For the sake of brevity, these rules will not be fully explained in this document. This rules cover the patterns that have already been presented for *Nbp* in previous sections; for example, prepositional phrases, presents of dative complements, possessive determiners, etc. The list of all rules can be found in the Appendix A. Examples that illustrate the cases where *npart* is involved, and the `WHOLE-PART` relations that are thus extracted, are shown below ((27)-(30)):

(27) *O Pedro roeu os seus cantos das unhas* (lit: Pedro gnawed his corners of the nails)

'Pedro gnawed the corners of his nails'

```
WHOLE-PART(unhas,cantos)
WHOLE-PART(seus,unhas)
```

(28) *O Pedro roeu o canto da unha* 'Pedro gnawed the corner of the nail'

```
WHOLE-PART(unha,canto)
WHOLE-PART(Pedro,unha)
```

(29) *O canto da sua unha infetou* 'The corner of his nail was infected'

```
WHOLE-PART(unha,canto)
WHOLE-PART(sua,unha)
```

(30) *O Pedro esgravatou no canto da unha* 'Pedro scratched the corner of the nail'

```
WHOLE-PART(Pedro,unha)
WHOLE-PART(unha,canto)
```

In this section, we have seen different cases that involve a noun designating a part of *Nbp*, the different patterns in which they co-occur, and the adaptations that were necessary in order to capture them adequately.

The next section will move to more complex relations that involve derived nouns associated to *Nbp*.

3.5 Complex Relations Involving Derived Nouns

As we have mentioned before, in some cases, a whole-part relation is only implicit, and though *Nbp* are involved, they are not mentioned directly (*gastritis*-'stomach'). In these cases, we decided that, nevertheless, a whole-part relation between the human entity and the "hidden" *Nbp* should be established.

At this time, we focus on predicative nouns designating *diseases*. 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 (31):

(31) *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])
```

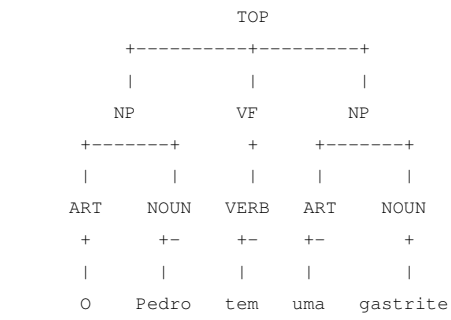
The rule itself reads as follows: first, the system checks if the disease noun (in this case, *gastrite* 'gastritis') is the direct object (`CDIR`) of the verb *ter* 'have' (variable `#1`); secondly, the system verifies if there is an explicit subject (variable `#3`) for the verb; and if there is still 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 output of the system on sentence (31) is given in Fig. 3.25.

The next type of sentences (example (32)) involves the support verb *estar com* 'be with' (more punctual aspect than *ter* 'have'):

(32) *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. The rule that captures the



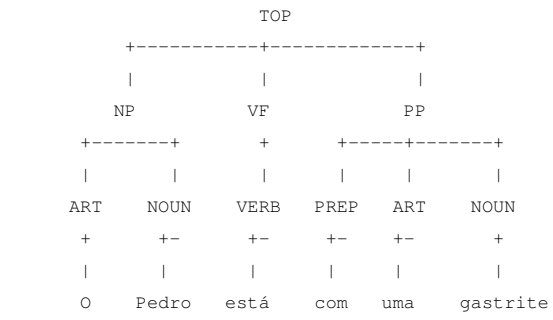

```
SUBJ_PRE(tem,Pedro)
CDIR_POST(tem,gastrite)
WHOLE-PART_HIDDEN(Pedro,estômago)
0>TOP{NP{O Pedro} VF{tem} NP{uma gastrite}}
```

Figure 3.25: WHOLE-PART relations for the sentence *O Pedro tem uma gastrite* 'Pedro has gastritis'.

meronymy relation between *Pedro* and *estômago* 'stomach':

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

The output of the system on sentence (32) is given in Fig. 3.26.




```
PREPD(gastrite,com)
MOD_POST(está,gastrite)
SUBJ_PRE(está,Pedro)
WHOLE-PART_HIDDEN(Pedro,estômago)
0>TOP{NP{O Pedro} VF{está} PP{com uma gastrite}}
```

Figure 3.26: WHOLE-PART relations for the sentence *O Pedro está com uma gastrite* (lit: Pedro is with a gastritis) 'Pedro has gastritis'.

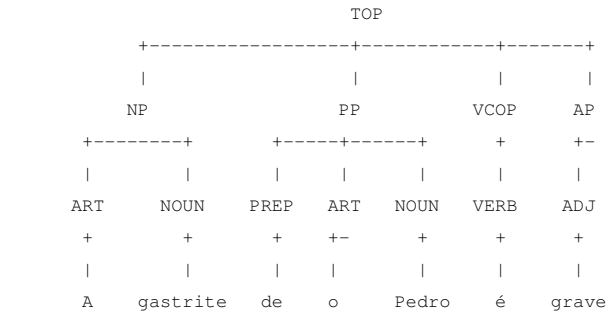
Finally, many support verbs and predicative nouns' constructions can be reduced to complex NPs, where the predicative noun is the head of the NP and its subject becomes a determinative *de N* 'of N' complement (eventually followed by any other complement of the predicative noun), as in sentence (33).

(33) *A gastrite do Pedro é grave* 'Pedro's gastritis is severe'

The rule that captures the meronymy relation between *Pedro* and *estômago* 'stomach' in these complex noun phrases is very similar to the previous ones, and it is shown below:

```
IF( MOD[POST](#2[lemma:gastrite],#3[human]) &
    PREPD(#3,?[lemma:de]) &
    ~WHOLE-PART(#3,?)
  )
    WHOLE-PART[hidden=+](#3,##noun#[surface:estômago,lemma:estômago])
```

The output of the system on sentence (33) is given in Fig. 3.27.




```
PREPD(Pedro,de)
MOD_POST(gastrite,Pedro)
SUBJ_PRE(é,gastrite)
WHOLE-PART_HIDDEN(Pedro,estômago)
0>TOP{NP{A gastrite} PP{de o Pedro} VCOP{é} AP{grave}}
```

Figure 3.27: `WHOLE-PART` relations for the sentence *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.[12]

3.6 Frozen Sentences (idioms) and Exclusion of Whole-Part Relations

There are many frozen sentences (or idioms) that involve *Nbp*, but for the overall meaning of these expressions the whole-part relation is often irrelevant, as in example (34).

(34) *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 (34) 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'

[12]Because of the XIP's syntax, it is not possible to merge the three rules of each (*predicative noun*, *Nbp*) pair into a single one, nor to make just 3 rules and keep the pairings.

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 the example (34) above, this corresponds to the dependency: FIXED(perdeu,cabeça), 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* 'loose', and a post-positioned direct complement whose head is the surface form *cabeça* 'head'.

Rules for identifying idioms and extracting the corresponding FIXED dependency are semi-automatically build from the lexicon-grammar tables of European Portuguese idioms [Baptista-et-al-2004], [Baptista-et-al-2005], [Baptista-et-al-2014]. In order to capture the idioms involving *Nbp*, we built about 400 of such rules, from 10 formal classes of idioms.

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 later, 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 (35) of idiom:

(35) *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)

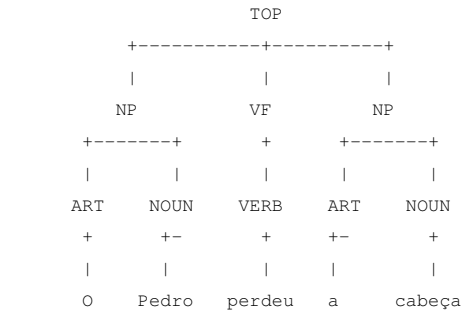
while another rule also captures the WHOLE-PART dependency between the subject and the *Nbp cabeça* 'head':

WHOLE-PART(Pedro,cabeça)

This is when the "cleaning" rule above takes place. It, first, verifies if both FIXED and WHOLE-PART dependencies are present and signals the later ('^') to be removed (1^{st} line); then it checks if they have common arguments (2^{nd} 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 (3^{rd} 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 2^{nd} and the 3^{rd} line are in disjunction: if at least one of the conditions match, the rule fires and removes the WHOLE-PART dependency.

Thus, considering the example (35) and the corresponding (provisory) dependencies, above, the 1^{st} line conditions are matched, but none of the 2^{nd} 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.

Similar rules had to be made to `FIXED` dependencies involving a smaller number of arguments (from 2 up to 7 elements). Returning to our example, the output of STRING for the idiom *perder a cabeça* 'loose the head' is given in Fig. 3.28




```
SUBJ_PRE(perdeu,Pedro)
CDIR_POST(perdeu,cabeça)
FIXED(perdeu,cabeça)
0>TOP{NP{O Pedro} VF{perdeu} NP{a cabeça}}
```

Figure 3.28: Frozen sentences (idioms) and exclusion of whole-part relations.

As one can see, no `WHOLE-PART` dependency was extracted and the `FIXED` dependency identifies the idiom.

In the case of idioms that involve *Nbp*, example (36):

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

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 (37) (which is signaled by '^{o}', below):

(37) ^{o}*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)` that would preclude the meronymy rules to be triggered, only the correct dependencies are extracted: `WHOLE-PART(João,cara)` and `WHOLE-PART(cara,lado)`. Similar rules were necessary for `FIXED`

dependencies with 3 or more arguments.

In this chapter, we presented the overview of STRING; the syntax of the dependency rules used in XIP; and the general rules addressing the most relevant syntactic constructions triggering whole-part relations in Portuguese; the chapter also addressed situations involving determinative nouns of *Nbp*, complex relations involving nouns derived from *Nbp*, and the way frozen sentences (idioms) containing *Nbp* elements were parsed.

Chapter 4

Evaluation

In this chapter, we present how the evaluation of the meronymy extraction module was performed: in Section 4.1, we describe how the evaluation corpus was produced; Sections 4.2 and 4.3 illustrate the organization of the annotation campaign, and the evaluation of the inter-annotator agreement; Section 4.4 presents the evaluation of the whole-part dependencies extraction involving *Nbp* and *Nsick*; in Section 4.5, we describe the error analysis, focusing on false-positive and false-negative cases; as a result of the error analysis, we provide, in Section 4.6, a second evaluation of the system's performance, once some of those problems were corrected.

4.1 Evaluation Corpus

The 1^{st} fragment of the CETEMPúblico corpus [Rocha-and-Santos-2000] 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* was adapted to the DELA format to be used within the UNITEX corpus processor [Paumier-2003],[Paumier-2014] 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. Also, the following nouns that are mostly non-human *Nbp* but can figuratively be applied to humans, in a pejorative way, were excluded: *asa* 'wing', *bico* 'nozzle', *casco* 'hoof', *cauda* 'tail', *cerne* 'core', *corno* 'horn'. 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). The distribution of the 10 most frequent *Nbp* is shown in Table 4.1; *Nsick* nouns are shown in Table 4.2. The full table of the *Nbp* in alphabetic order is presented in Appendix C.

Table 4.1: 10 most frequent *Nbp*.

Nbp			
Lemma	**Count 1 (in the corpus)**	**%**	**Count 2 (selected)**
mão 'hand'	1,525	12.05	120
face 'face'	1,362	10.76	107
corpo 'body'	1,116	8.82	88
cabeça 'head'	970	7.66	76
pé 'foot'	721	5.70	56
língua 'tongue'	683	5.40	53
olho 'eye'	655	5.17	51
braço 'arm'	420	3.32	33
coração 'heart'	416	3.29	32
cara 'face'	396	3.13	31
Total:	**8,264**	**65.28**	**647**

Table 4.2: Number of *Nsick*.

Nsick			
Lemma	**Count 1 (in the corpus)**	**%**	**Count 2 (selected)**
artrite 'arthritis'	7	8.86	6
bronquite 'bronchitis'	3	3.80	1
diabetes 'diabetes'	36	45.57	7
faringite 'pharyngitis'	1	1.27	0
hepatite 'hepatitis'	28	35.44	3
osteoporose 'osteoporosis'	4	5.06	3
Total:	**79**	**100**	**20**

A total of 17 sentences with *Nsick* were randomly collected from the 79 occurrences in the corpus; however, from the distribution of these nouns shown in Table 4.2, one can see that there were 20 occurrences of them in 17 sentences. This was due to the fact that some sentences featured more than one *Nsick*.

A separate exercise of annotation was done to this small class of nouns (see subsection 4.4.4).

On the date 29.01.2014, the rules were integrated in the system, and the corpus was parsed. For each sentence the WHOLE-PART relations were extracted (or not). The output of a sentence looks like this:

```
WHOLE-PART(sua,boca)
45>TOP{Mas , PP{em a sua boca} , NP{a palavra} NP{democratização} VF{tem} NP{o sentido inverso} PP{a o invocado} PP{por Smith} .}

WHOLE-PART(Carmen,corpo)
218>TOP{ADVP{Então} , VGER{obstaculizando} PP{com o seu corpo} PP{a marcha} PP{de Carmen} , NP{Jesus} VF{cravou} NP{lhe} PP{a navalha} .}
```

In the first example, the WHOLE-PART relation was correctly extracted, while in the second it was not, for the whole argument should also be the possessive pronoun *seu* 'his'.

4.2 Annotation Campaign

The output sentences were then divided into 4 subsets of 225 sentences each, and a common set of 100 sentences was added to each subset in order to assess inter-annotator agreement.

A set of annotation guidelines (Appendix D) was prepared for the annotators, in order to ensure uniformity in the process.

The four annotators involved in the task:

Annotator 1 holds a PhD in Linguistics and is quite familiar with the topic being described.

Annotator 2 holds a MSc in Marine Biology and a BA in Language and Communication; while previously unfamiliar with the topic at hand, she has basic notions of corpus annotation and semantic relations.

Annotator 3 holds an MA in Linguistics, and she was also previously unfamiliar with the topic and with corpora annotation tasks.

Annotator 4 has an incomplete BA degree in Organizational Communication, and she was also previously unfamiliar with the topic and with corpora annotation tasks.

The age of the annotators varied, from 45 (Annotator 1 and 2) to 25 (annotator 3) and 23 (annotator 4).

While annotators 1 and 2 were both native European Portuguese speakers; annotators 3 and 4 were both native Brazilian Portuguese speakers. These two last annotators have both been living in Portugal for at least 6 months. The fact that two annotators were native speakers of the non-European variety was deemed to be irrelevant for the nature of the task.

None of the annotators is mutually acquainted and the annotation process was done separately, all contacts being done through e-mail. While the possibility existed for clarifying any eventual doubts, no annotator contacted us to that purpose.

Annotator 1 reviewed annotator 2 and 3 for formatting mistakes, namely, the insertion of FIXED (removed) and the use of the determiners instead of the head nouns in the WHOLE-PART dependency. Annotator 4 has also consulted the author on the issue of removing/correcting FIXED dependencies,

apparently not made sufficiently clear in the annotation guidelines. Furthermore, annotator 3 raised the issue, also not explicit in the guidelines about choosing the closer "whole" antecedent for the body-part, when this is a pronoun, even if the antecedent of that pronoun is in the same sentence, like in example (38):

(38) *Quando o João o atacou, o Pedro partiu-lhe o braço*

'When João attacked him, Pedro broke him[=João] the arm'

4.3 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 [Fleiss-1971], and the Cohen's Kappa coefficient of inter-annotator agreement [Cohen-1960] using ReCal3: Reliability Calculator [Freelon-2010], for 3 or more annotators.[1]

The raw data provided by annotators was converted into a tabular format, adopting the following convention, comparing the changes introduced (or not) by the annotators against the output of the system:

[**0**] The annotator **did not change** the output of the system.

[**1**] The annotator **removed** the `WHOLE-PART` relation.

[**20**] The annotator **added** a `WHOLE-PART` relation.

[**21**], [**22**], etc[2]. The annotator **added** a `WHOLE-PART` relation, but a different one from another annotator.

[**31**] The annotator **changed** the `WHOLE-PART` of the system output (only the **whole** was changed).

[**32**] The annotator **changed** the `WHOLE-PART` of the system output (only the **part** was changed).

[**33**] The annotator **changed** the `WHOLE-PART` of the system output (both the **whole** and **part** were changed).

Table 4.3 describes the distribution of the different types of interventions the annotators made in the corpus.

As one can see, in most cases, the annotators did not change the output of the system [0]. The second most frequent case is an annotator added a `WHOLE-PART` dependency [20]. Finally, the third most frequent situation is the removal of the semantic relation [1]. As for the partial changes in the dependencies, only those affecting the hole were observed.

In some cases, different annotators added different sets of whole-part dependencies. For example, for sentence:

(39) `4>TOP{NP{NOUN{Abdel Rahman}} , NP{55 anos} , SC{que VCOP{é}} AP{cego} e VF{sofre} PP{de diabetes} , VF{sentia} NP{se} ADVP{" bastante bem "} , [...]`

'Abdel Rahman, 55 years-old, who is blind and suffers from diabetes, felt himself "very well" ...'

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

[2] Since the number of the annotators is four, there may be up to four different annotations for a given instance.

Table 4.3: Distribution of the annotations in the corpus.

	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Total (%)
[0]	81	78	86	83	328 (75.9%)
[1]	7	7	7	7	28 (6.5%)
[20]	17	17	11	15	60 (13.9%)
[21], [22], ...	0	4	1	1	6 (1.4%)
[31]	3	2	3	2	10 (2.3%)
[32]	0	0	0	0	0 (0%)
[33]	0	0	0	0	0 (0%)
Total instances:	**108**	**108**	**108**	**108**	**432 (100%)**

In this case, annotator 2 added two `WHOLE-PART` relations:

`WHOLE-PART_HIDDEN(Abdel Rahman,olhos)`

`WHOLE-PART_HIDDEN(Abdel Rahman,pâncreas)`

The first one, probably, because of the adjective *cego* 'blind', was incorrectly added, since no disease noun – this would be *cegeira* 'blindness' – is involved, which was the task at hand; the second one is correct, as diabetes relates to the *Nbp pâncreas* 'pancreas'. On the other hand, annotator 4 only added the second, correct dependency. In order to calculate agreement, in these cases, we treated this sentence as two instances of annotation, one where both annotator agreed, and another with the off-mark annotation.

Another case happened in the following sentence:

(40) `88>TOP{NP{Os budistas} e NP{adeptos} PP{de o NOUN{" candomblé "}} VF{indicaram} SC{que VF{receberão}} NP{NOUN{João Paulo II}} PP{de braços} AP{abertos} .}`
'Budists and adepts of "candomblé" stated that they would welcome João Paulo II with open arms'

For which the system incorrectly extracted the dependencies:

`WHOLE-PART(João Paulo II,braços)`

`WHOLE-PART(adeptos,braços)`

Having failed to identify the ambiguous idiomatic adverb *de braços abertos* 'with open arms'. In this case, all 4 annotators correctly removed both `WHOLE-PART` dependencies, so we consider that they have agreed twice, and duplicated the corresponding annotation instance.

A more complex case took place with the following sentence, where the system produced no output:

(41) `42>TOP{NP{NOUN{Marjorie Wallace}} , SC{quando NP{as} VF{viu}} PP{por a primeira vez} PP{em o julgamento} , VF{escreveu} SC{que VF{eram}} NP{dois seres} " AP{pequenos} e AP{vulneráveis} , e ADVP{não} VF{abriam} PP{a boca} VINF{a ADVP{não} VINF{ser}} SC{para VINF{emitir}} NP{uns murmúrios} SC{que NP{o tribunal} VF{interpretou}} como NP{sinais} AP{evidentes} PP{de culpabilidade} " .}`
'Marjorie Wallace, when she saw them(fp) both for the first time in the trial, wrote that they were two

small and vulnerable beings, and that they did open their mouths unless to utter some mumblings that the court interpreted as evident signs of guilty'

While annotators 1 and 2 correctly added the dependency:

`WHOLE-PART(seres,boca)`

annotator 4 has inadequately picked up the adjectival modifier for the "whole" argument:

`WHOLE-PART(pequenos,boca)`

Finally, annotator 3 considered that the expression was idiomatic and added a `FIXED` dependency, something that had not been required by the guidelines (and perhaps it should have been made more clear that it was not supposed to be done):

`FIXED(abriam,boca)`

The idiomatic nature of the expression is unclear for much of the literal meaning of the elements involved is still there, so it is only natural that annotators could adopt either perspective on the expression status.

Finally, another interesting and similar case occurred with the next sentence:

(42) `34>TOP{SC{Para NP{o} VINF{conseguir}} , NP{os dirigentes} PP{de o PSD} VF{ouviram} PP{de a boca} PP{de o líder} PP{de o partido} PP{a argumentação} AP{necessária} SC{para VF{convencerem}} NP{o eleitorado} PP{até Dezembro} .}`
'To achieve this, the leaders of the PSD (political party) heard from the mouth of the Party's leader the arguments needed to convince the electorate until December'

where the system produced the following, obviously incorrect output:

`WHOLE-PART(dirigentes,boca)`

While annotators 1, 3 and 4 changed it into:

`WHOLE-PART(líder,boca)`

This case is interesting because it depends on how one analyses the expression *(ouvir) da boca de Nhum*: it can be considered an adverbial idiom, meaning 'receive the information directly from someone', but it still has much of the literal meaning of the elements involved, so it could be interpreted by our annotators as a valid target for a `WHOLE-PART` dependency extraction. Now, to make matters even more complicated, annotator 2 changed the dependency into:

`WHOLE-PART(portistas,boca)`

Notice that the noun *portistas* 'fans of Porto football club' does not even appear in this sentence, but in another sentence that happened to appear nearby, so this is an obvious mistake of the annotator. Considering that this last notation was intended to produce a similar result as the others, we encoded it in a similar way.

Because of these different solutions, instead of 100 sentences, in the end there were 108 annotation instances to be compared and the number of decisions was 432.

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 4.4.

Table 4.4: Average Pairwise Percent Agreement.

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

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 4.5. Fleiss' Kappa[3]:

> "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."

Table 4.5: 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 4.6.

Table 4.6: Average Pairwise Cohen's Kappa (CK).

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

Cohen's Kappa coefficient[4] is defined as:

> "a statistical measure of inter-rater or inter-annotator agreement for qualitative (categorical) items [. . .]. The equation for k is:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

[3] http://en.wikipedia.org/wiki/Fleiss'_kappa
[4] http://en.wikipedia.org/wiki/Cohen's_kappa

where *Pr(a)* is the relative observed agreement among raters, and *Pr(e)* is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category."

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 [Landis-and-Koch-1977] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [Fleiss-1981], these results correspond to an inter-annotator agreement halfway between "fair" and "good".

In view of these results, we can assume as a reasonable expectation that the remaining, independent and non-overlapping annotation of the corpus by the four annotators is sufficiently consistent, and will use it for the evaluation of the system output, in the way described in the next section.

4.4 Evaluation of the Whole-Part Dependencies Involving *Nbp* and *Nsick*

In order to evaluate the output of the system we need to produce a *golden standard*, that is, a correctly annotated corpus. The first 100 sentences of the corpus that were annotated by 4 different native speakers were compared among themselves, and the majority decision of the annotators was chosen as the correct solution or the golden standard (Appendix E). This also allowed us to evaluate the inter-annotator agreement. For the remainder of the corpus' sentences, we rely on the relatively high inter-annotator agreement to consider them as a golden standard, in order to confront it against the system's output. Nevertheless, in this section, results for each segment of the corpus will always be presented separately.

4.4.1 Definition of Evaluation Measures

For the calculation of the evaluation measures of *Precision* (P), *Recall* (R), *F-Measure* (F), and *Accuracy* (A) we adopted the following definitions:

$$\text{Precision} = \frac{\text{number of correctly extracted whole-part dependencies}}{\text{total number of extracted whole-part dependencies}}$$

$$\text{Recall} = \frac{\text{number of correctly extracted whole-part dependencies}}{\text{total number of whole-part dependencies in the corpus}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{total number of correctly extracted dependencies} + \text{total number of true-negative cases}}{\text{total number of instances}}$$

True-negative (TN) cases correspond to the instances where there is an *Nbp* in the sentence but no whole-part relation can be extracted, either because it is an idiom or because the whole is not mentioned, or some other reason.

As we will see in the next paragraphs, these calculations are not without some problems.

4.4.2 Problematic Cases

There were 5 cases, in the first 100 sentences, where two annotators disagreed with the other two. In these cases, as it was impossible to identify a majority vote, the decision had to be made by us. For example, in the sentence below (already shown, above, but here repeated for clarity):

(43) `4>TOP{NP{NOUN{Abdel Rahman}} , NP{55 anos} , SC{que VCOP{é}} AP{cego} e VF{sofre} PP{de diabetes} , VF{sentia} NP{se} ADVP{" bastante bem "} , [...]`
'Abdel Rahman, 55 years-old, who is blind and suffers from diabetes, felt himself "very well" ...'

we consider that annotator 4 made the correct decision by extracting the dependency:

`WHOLE-PART_HIDDEN(Abdel Rahman,pâncreas)`

while annotators 1 and 3 have failed to spot any relation. Besides, in this case, annotator 2 added two `WHOLE-PART` relations, the first one espurious:

`WHOLE-PART_HIDDEN(Abdel Rahman,olhos)`

`WHOLE-PART_HIDDEN(Abdel Rahman,pâncreas)`

Thus, we followed the solutions where both annotators 2 and 4 agreed (the second dependency), and discarded the other dependency, which was, in fact, incorrect. The sentenced was assessed as a *false-negative*, that is, there is an annotation that the system should have made, but it did not.

In the next example, the system also did not extract any `WHOLE-PART` dependency:

(44) `26>TOP{NP{A outra} VF{mostra} NP{um judeu} , NP{ultra-ortodoxo} , AP{identificado} como NP{tal} PP{por a farta barba} e NP{a NOUN{" kippa "}} , NP{a mitra} .}`
'The other half shows a jew, ultra-orthodox, identified as such by the abundant beard and the "kippa", the traditional small round cover for the head'

Annotators 3 and 4 did not add any `WHOLE-PART` dependency, whereas annotators 1 and 2 decided to add a `WHOLE-PART` dependency:

`WHOLE-PART(judeu,barba)`

In this case, we consider that annotators 1 and 2 made a correct decision. Thus, it is also a case of a *false-negative*.

In the next example, the system extracted a `WHOLE-PART` dependency:

`WHOLE-PART(santos,bocas)`

(45) `53>TOP{PP{Em dois templos} VCOP{foram} VCPART{destruídos} NP{sacrários} e NP{as hóstias} AP{colocadas} PP{em as bocas} PP{de as imagens} PP{de santos} , mas NP{as caixas de esmolas} ADVP{não} VCOP{foram} VCPART{assaltadas} .}`
'In two temples, the shrines were destroyed, and the communion wafers placed in the mouths of the images of the saints, but the poor boxes were not robbed'

Annotators 2 and 4 considered the output of the system to be correct; but annotators 1 and 3 changed the *whole* argument in the extracted dependency from *santos* 'saints' to *imagens* 'images':

`WHOLE-PART(imagens,bocas)`

We consider that annotators 2 and 4 were correct, as *santos* 'saints' is a determinative complement of the head noun *imagens* 'images', which, in its stead, a determinative complement of the *Nbp*. Thus, the correct notation is treated as a *true-positive*, that is, a notation added by the system that should in fact be extracted.

In the other cases, where an annotator changed correctly the *whole* **or** the *part* argument of WHOLE-PART dependency, we decided to count it only as half of *true-positive* case, as the system already extracted one part of a dependency correctly.

In the next example, the system did not extract any WHOLE-PART dependency:

(46) 77>TOP{VF{Confirmou} ADVP{assim} NP{a versão} PP{de o antigo comandante} PP{de o NOUN{posto de a GNR de Sacavém}} que , quando PP{de o início} PP{de o julgamento} , VF{explicou} PP{a o colectivo} NP{o movimento} SC{que VF{fez}} PP{com o braço} - - PP{em o sentido ascendente} - - e SC{que VF{provocou}} NP{o disparo} (VF{dito} AP{acidental}) .}

'[He] confirmed then the version of the former commanding officer of the Sacavém GNR police station, who, at the beginning of the trial, had explained to the the collective of judges the mouvement he did with the[=his] arm – in as ascending way – and which caused the (so-called accidental) shot'

Annotators 3 and 4 did not add any WHOLE-PART dependency either. Annotator 1 added a new WHOLE-PART dependency:

WHOLE-PART(comandante,braço)

annotator 2 added a WHOLE-PART dependency different from annotator 1:

WHOLE-PART(ele,braço)

Notice that there is no pronoun *ele* 'he' in the sentence, so the annotator reconstructed the elliptic subject of the sentence. Furthermore, the sentence is ambiguous, and there is not enough evidence in it to decide who is the author of the movement, and hence the "owner" of the arm. In this case, the majority vote (no dependency extracted) is incorrect, but the two other annotators partially got the dependency right (the part argument), though they disagree with about the *whole* argument. As either one is partially correct, we considered this case as a *false-negative*.

Finally, we give the example where the system extracted a WHOLE-PART dependency:

WHOLE-PART(Jorge Soares,cabeça)

(47) 99>TOP{ADVP{Ainda} PP{em o mesmo jogo} , NP{destaque} PP{para o golo} PP{de NOUN{João Pinto}} , NP{outro tiro} PP{de fora de a área} , e NP{o primeiro} PP{de NOUN{Paulo Nunes}} , AP{acrobático} , PP{depois de dois toques} PP{de cabeça} PP{de NOUN{Jorge Soares}} e NP{Gamarra} .}

'Still in the same match, notice that the goal made by João Pinto, another shot from outside of the area, and the first one, from Paulo Nunes, acrobatic, after two touches of head from Jorge Soares and Gamarra'

Annotators 2 and 4 did not change the output of the system. Annotators 1 and 3 added one more WHOLE-PART dependency:

WHOLE-PART(Gamarra,cabeça)

In this case, we consider that annotators 1 and 3 were right as the "two touches" can be read distributively, one from each player. As the system extracted one WHOLE-PART dependency correctly but it did not extract the second WHOLE-PART dependency, the output of the system is assessed as one *true-positive* and one *false-negative*.

4.4.3 Evaluation of the System's Overall Performance

Next, the system performance was evaluated using the usual evaluation metrics of Precision, Recall, F-measure and Accuracy, explained in section 4.4.1, with the remarks of section 4.4.2. The results are shown in Table 4.7, where TP=*true-positives*; TN=*true-negatives*; FP=*false-positives*; FN=*false-negatives*.

Table 4.7: 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

The number of instances (TP, TN, FP and FN) is higher than the number of sentences, as one sentence may involve several instances, like in the example described above, where the sentence is assessed as one *true-positive* and one *false-negative*. The relative percentages of the TP, TN, FP and FN instances are similar between the 100 and the 900 set of sentences. This explains the similarity of the evaluation results and seems to confirm our decision to use the remaining 900 sentences' set as a golden standard for the evaluation of the system's output with enough confidence. 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 nouns and the target *Nbp*. 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 not any whole-part relation to be extracted, even if there is an *Nbp* in the sentence.

4.4.4 Evaluation of the System Performance for *Nsick*

In the same way, we then compared the automatically produced subcorpus of 70 sentences with *Nsick* against a golden standard that was manually annotated by a linguist. Again, we evaluated the system's performance using the usual evaluation metrics of Precision, Recall, F-measure and Accuracy. Results are shown in Table 4.8, below:

Notice that there are 79 *Nsick* in 70 sentences, but the sum of TP, TN, FP and FN is 80 because in 1 sentence with 1 *Nsick* the system incorrectly extracted 2 WHOLE-PART relations:

WHOLE-PART(pessoas,mama)

WHOLE-PART(pessoas,cólon)

Table 4.8: System's performance for *Nsick*.

Nsick	TP	TN	FP	FN	Precision	Recall	F-measure	Accuracy
80	2	59	2	17	0.5	0.11	0.17	0.76

(48) *Na região, os acidentes de viação matam mais **pessoas** do que as doenças como os diabetes e os tumores malignos da **mama** e do **cólon***
'In this region, car accidents kill more **people** than diseases like diabetes or malign tumors of the **breast** and of the **colon**'

The accuracy is high because there are many sentences where the disease is just mentioned, and there is no human noun who could be interpreted as affected by that disease, like in the next example:

(49) *As histórias da poluição do rio Grande correm toda a região, desde o aparecimento de cadáveres de animais na sua foz até ao boato de um surto de **hepatite B** que no ano passado afastou centenas de veraneantes.*
'The stories about pollution in the rio Grande spread out through the entire region, since the appearance of animals' corpses at the river mouth and even the rumor of a **hepatitis B** outbreak that last year drove off hundreds of summer tourists'

Cases like these are treated as *true-negatives*, and from the previous table one can see that they constitute the majority of the sentences in this small subcorpus. A more detailed error analysis will be given in the next section.

4.5 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 a detailed analysis of the false-positives and then move on to the false-negatives.

4.5.1 False-positives

Rules' Correction

To begin with, we tackled the situation where the system incorrectly extracted the whole-part relation between the subject of the sentence and a direct complement *Nbp* when this is further modified by a `PP` introduced by preposition *de* 'of', as in sentence (50):

(50) *Os cientistas não encontraram o crânio do animal*
'The scientists have not found the cranium of the animal'

```
WHOLE-PART(cientistas,crânio)
```

In this case, we restricted the general rule by precluding the whole-part relation extraction if there is a [MOD]ifier relation between the *Nbp* and another noun introduced by preposition *de* 'of':

```
IF ( SUBJ[PRE](#3,#1[human]) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     ~PREPD(#5,#6[lemma:de]) & // line added during the error analysis
     ~MOD(#2,#5) & // line added during the error analysis
     ~WHOLE-PART(#1,#2) &
     ~WHOLE-PART(#4,#2)
   )
     WHOLE-PART(#1,#2)
```

In the same way, we modified 4 other rules in order to avoid the whole-part relation extraction in these situations, that had not been previously taken into consideration in the grammar.

Disambiguation of *Nbp* in Context

An interesting number of cases occurred 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 meanining of an anatomical part, we adopted one of the following strategies:

(i) we removed the *Nbp* (sem-anmov) feature from the nouns lexical set of features; this is carried out by the following rules, which are applied before the chunking stage, in a similar way as we had done in 3.4.2:

— 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=+].
```

Still, this solution does cover many the instances where *língua* 'language' is not an *Nbp*:

(51) *O futuro do Zaire talvez comece este fim-de-semana num navio de 167 metros de comprimento auspiciosamente chamado "Outeniqua", o que à letra – na língua de um dos povos sul-africanos – significa "transportador de mel"*
'The future of Zaire may start this week-end in a 167-meter long ship, auspiciously named "Outeniqua", which literally - in the language of one of the South African people - means "carrier of honey"'

WHOLE-PART(povos,língua)

A finer-grained word-sense disambiguation is, thus, necessary.

— 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 (Appendix B.2), 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=+].
```

(ii) Certain word combinations would be better described, maybe, as compound nouns: *dicionário de língua* 'language dictionary', *professor de línguas* 'language teacher'; others are not so clearly compounds: *ensino de línguas* 'language teaching'. In these cases, if the sequence is followed by a gentilic adjective, the word *língua* 'language' is already disambiguated (see above); otherwise, we did not want to enforce the compound noun analysis, so a disambiguation rule was also devised; only the most frequent combinations were considered.

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

On the other hand, the compound noun *escola de línguas* 'language school' was dealt with as a new compound.

Certain compound prepositions and adverbs were absent from the lexicon, so we added them: *de dedo em riste* 'with his finger pointed', *na/à cabeça de* 'at the head of <a group of people>'. The later requires a plural or a collective noun as its argument.

Another interesting case involving compounds also occurs in:

(52) *Os campeões portugueses começaram bem a partida, com dois lançamentos triplos de Carlos Lisboa, mas não conseguiram repetir a vitória de a* ***primeira mão*** *em Israel*
'The Portuguese champions started the match well, with two triple launches by Carlos Lisboa, but could not repeat the victory of the *first match* [lit: first hand] in Israel'

```
WHOLE-PART(campeões,mão)
```

where the compound *primeira mão* (lit: first hand) 'the first match between two teams, in a football championship' had not been identified. This has to do with the ambiguous status of this word combination, that also appears in many other frozen or idiomatic combinations.

Some idioms have not been captured because they had not been encoded in the lexicon yet. Therefore, we completed the existing list of rules for FIXED expressions, in order to encompass those missing cases:

- *ser de boa boca* (lit: to be of good mouth) 'to have sound appetite, to eat everything';
- *estar/ver-se a braços com* 'having to deal with some problem';
- *estar/ficar de braços cruzados* (lit: to cross one's arms) 'to do nothing';
- *(não) passar pela cabeça de Hum* 'not to come to one's mind';
- *morder as canelas de/a Hum* 'to trick/betray Hum';

- *abrir o coração a* 'to open one's heart to sb., to speak openly';
- *fazer face a* 'to deal with';
- *deixar N de mãos atadas, estar de mãos atadas* 'to leave someone / to be with one's hands tied';
- *sair da (sua) mão* 'when driving, move to the opposite lane of the trafic'.

Some idioms correspond to support verb constructions ([Gross-1981], [Ranchhod-1990], [Baptista-1997b]), so that they may have to receive further attention in the future, when this type of expressions becomes integrated in STRING:

- *dar uma/a mão a* 'give a hand to' [class DR, [Baptista-1997b]];
- *estar em as mãos de* 'to be in one's hands' [class EPCQ0, [Ranchhod-1990]].

In all, 22 new rules had to be devised, tested, and finally added to the lexicon-grammar of idioms.

Difficult Cases

Finally, a certain number of cases were found where the use of the *Nbp* is clearly figurative, but it is not neither an idiom nor a compound word, so we were unable to devise any strategy to avoid capturing the whole-part relation:

(53) *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)`

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'.

A similar figurative use of the noun *face* (*id.*) is found in:

(54) *Além disso, a nova* ***face*** *desta Igreja chilena não se forjou na luta contra o comunismo, mas na defesa dos direitos humanos contra a barbárie, durante a ditadura militar de Pinochet*
'Moreover, the new face of this Chilean Church was not forged in the struggle against communism, but in defense of human rights against barbarism, during the military dictatorship of Pinochet'

`WHOLE-PART(igreja,face)`

In this case, the figurative use of *face* (*id.*) is similar to the one in the English translation. A more explicit, predicative metaphor using a synonym of this noun, *rosto* 'rostrum', is found in:

(55) *No Malecón, a enorme marginal da cidade, que é, segundo Vivian Corona, "o seu rosto", os belos edifícios de colunas foram pintados há uma meia dúzia de anos de cores vivas*
'On the Malecon, the huge seaside walk of town, which is, according to Vivian Corona, "its face", the beautiful buildings of columns were painted there are a half dozen years of vivid'

`WHOLE-PART(seu,rosto)`

Even More Difficult Cases

As the whole-part dependency extraction is being carried out at the final stages of parsing, any problems in the preceding steps accumulate, and can often hinder the correct extraction.

Errors can be derived right from the sentence-splitting stage, one of the first processing steps in the STRING chain, as in the sentence below:

(56) *"É um vírus muito frágil e, nas condições em que os corpos se devem encontrar congelados, quase de certeza que foi destruído", disse ao PÚBLICO este investigador do Instituto de Patologia das Forças Armadas, em Washington D.C. Houve quem ficasse tempo sem fim deslumbrado a ligar o interruptor que apagava e acendia uma lâmpada fluorescente, acompanhando com movimentos do corpo os "estremecimentos" luminosos da lâmpada*

'"It is a very fragile virus and in the conditions in which bodies must be now, that is, frozen, almost certainly it has been destroyed", said to the PÚBLICO this researcher from the Armed Forces Pathology Institute in Washington D.C. There were people who remain dazzled an endless time, flipping the switch that extinguished and lit a fluorescent lamp, accompanying with their body movements the bright "shivers" of the lamp'

`WHOLE-PART(investigador,corpo)`

In this case, the sentence-splitter did not recognized the abbreviation mark of D.C., which is also the end of that sentence. Therefore, this was considered as only one sentence, and naturally, the remainder of the parsing becomes problematic. If only the second sentence is parsed, no whole-part dependency is extracted. Still, it could be argued that there is a whole-part relation between the interrogative pronoun *quem* 'who' and the *Nbp corpo* 'body', but the guidelines we defined did not refer this situation, which prompts to its future improvement.

Complex continuents are particularly difficult to parsing as it happens, for example, in the following sentence:

(57) *A sua mulher, Elizabeth, e seus filhos Philip and Chislaine acompanharam a transladação do corpo, num jacto particular, desde as Ilhas Canárias até Israel*

'His wife, Elisabeth, and his sons Philip and Ghislaine accompanied the body's relocation in a private jet, from the Canary Islands to Israel'

`WHOLE-PART(filhos Philip,corpo)`

The coordination of two proper nouns that are in appostion to *filhos* 'sons', but which are themselves coordinated to *mulher* 'wife', this noun also with an apposition (Elizabeth), makes this a too complex `NP` to be correctly parsed at this stage by the system. Nevertheless, the parser was able to extract as the verb's subjects *filhos* 'sons' and *Ghislaine*. The whole-part relation between *filhos* 'sons' and *corpo* 'body' was incorrectly captured, however, due lack of the semantic information on the construction of *acompanhar* 'accompany' with an object as *corpo* 'body', which precludes correference between the subject and the *Nbp* (the deceased).

A somewhat similar case occurs with the follwing examples:

(58) *Os árabes chamavam-lhe, por causa da sua forma, dedo* 'The Arabs called it, because of its shape, finger'

`WHOLE-PART(árabes,dedo)`

(59) *Uma das últimas vezes foi quando um amigo lhe pediu para que falasse perante um congresso de médicos no problema das glândulas supra-renais*
'One of the last times was when a friend asked him to speak before a congress of medical doctors about the problem of the adrenal glands'

`WHOLE-PART(médicos,glândulas)`
`WHOLE-PART(amigo,glândulas)`

As no syntactic-semantic information derived from the verb construction is being used in the meronymy module, the rules are unaware of the specific syntactic function and the corresponding semantic role of the verb's arguments. In the examples above, the fact that the verb *chamar* 'to call' and *falar* 'to speak' have been disambiguated as ViPEr verbs [Baptista-2012] from classes 39 and 41, respectively, could be used to remove the incorrect whole-part dependencies, as the semantic roles of *dedo* 'finger' and *problema* 'problem' with these verbs are incompatible (or at least difficult to conceive) with a meronymy relation.

In the next case, the parser incorrectly extracted whole-part relations for elements very distant from each other:

(60) *São as gémeas Jane e Louise Wilson que apresentam uma obra construída a partir do segredo impartilhável da duplicidade-unidade unovolar: uma sala vazia destruída por lutas de violência indescritível e um duplo vídeo onde as artistas se fazem figurar nesse espaço assumindo a impureza do corpo performativo*
'It were the twin sisters Jane and Louise Wilson who are presenting a work constructed from the unsharable secret of the unovolar unicity-duplicity: an empty room destroyed by struggles of indescribable violence and a double video where the artists present themselves in that space assuming the impurity of the performative body'

`WHOLE-PART(artistas,corpo)`
`WHOLE-PART(gémeas Jane,corpo)`
`WHOLE-PART(Louise Wilson,corpo)`

There is also a complex subject `NP`, with the proper names in apposition to the noun *gémeas* 'twin sisters', however, the coordination between the two `NP`s was captured, hence there are two (anaphoric) subjects for the verb *apresentam* 'present' in the relative clause. However, here, the incorrect extraction of whole-part relation has two different causes: first, the sentence after the colon (:) can be viewed as a description of the noun *obra* 'artistic work'; it should be a new syntactic unit, but the parser does not treat the colon as a sentence separator; secondly, the original rule did not enforce a relation between the modifier *Nbp* and the verb with a human subject. Therefore, the system captured any previously occurring subject, including the coordinated `NP`s as the "whole" of a `PP` with an *Nbp* head noun and introduced by preposition *de* 'of', even if they were syntactically unrelated.

The original rule was corrected and a new condition added, `?(#2,#6)`, making sure that at the verb and the element the *Nbp* depends on are syntactically related:

```
IF ( VDOMAIN(#1,#2) &
     SUBJ(2,#3[human]) &
     ?(#2,#6) &
     MOD(#6,#4[UMB-Anatomical-human]) & PREPD (#4,#7[lemma:de]) &
     ( ~MOD(#4,#5[human]) || ~CINDIR(#2,#5) ) &
     ~WHOLE-PART(#3,#4) &
     ~WHOLE-PART(#8,#4)
   )
     WHOLE-PART(#3,#4)
```

Now, the rule yields `WHOLE-PART(artistas,corpo)`, which is not altogether wrong, though the (poetic?) description may allow for a generic (and non-correferent) interpretation of *corpo* 'body'.

A more obvious, generic use of this *Nbp*, *corpo* 'body', can be found in:

(61) *Os escapes dos automóveis, das camionetas e dos autocarros que, constantemente, fumigam as ruas e as pessoas, os muitos lixos e os seus receptáculos, o odor dos corpos comprimidos nos transportes públicos, quase fazem esquecer os cheiros agradáveis da nossa cidade*
'The exhausts of the cars, the vans and the buses that constantly fumigate the streets and the people, the many wastes and their containers, the smell of the bodies compressed inside the public transportation, almost make you forget the pleasant smells of our city'

`WHOLE-PART(pessoas,corpos)`
In this case, the definite article used in a generic way: *os corpos das pessoas* 'the bodies of the people', but there is no syntactic relation, unlike the extracted dependency might suggest, between the previous instance of *pessoas* 'people' and the later occurring *Nbp* *corpos* 'bodies'.

In the next case (which in fact occurred twice), several problems arised:

(62) `TOP{NP{Iniciativa} PP{de a sociedade} AP{civil} PP{de os países} NP{promotores}, NP{o encontro} VF{pretendeu} VINF{ser} NP{um degrau} ADVP{mais} PP{para a formalização} PP{de a Comunidade} PP{de os Povos} PP{de Língua} VF{Portuguesa} .}`
'As an initiative of the civil society from the promoting countries, the meeting was intended to be a step further towards the formalization of the Community of Portuguese-Speaking Peoples'

`WHOLE-PART(Comunidade,Língua)`
On the one hand, the POS tagging failed to recognized *Portuguesa* as an adjective 'Portuguese' and treated as a verb *portuguesar* 'to render Portuguese', or 'Portuguese-like'; this situation was corrected at the pre-parsing stage. On the other hand, there is a multiword named entity that was absent from the lexicons, and we added it after the fact. However, even if the named entity had not been identified, the rules involving gentilic adjectives would have removed the *Nbp* sense from the noun *língua* 'language', if it were not for the POS initial error.

Some of the errors derived from the fact that at this stage of the processing chain no anaphora resolution has been carried out yet:

(63) *No regresso dos arguidos à sala de audiências, instalou-se a confusão, com dois deles, José Freitas e Filipe Moreira (este com uma das pernas engessada e apontando uma muleta na direcção do colectivo de juízes) a levantarem a voz, afirmando-se "ameaçados pela segurança"*

'In the defendants' return to the courtroom, some confusion occurred, with two of them, José Freitas and Filipe Moreira (this one with a leg in a cast and pointing a crutch towards the panel of judges) raising theirs voices, claiming that they had been "threatened by the security officers"'

```
WHOLE-PART(José Freitas,pernas)
WHOLE-PART(Filipe Moreira,pernas)
```

In this case, there is a bracketed insertion with the demonstrative pronoun *este* 'this' that refers to the last named entity, Filipe Moreira; however, as this is coordinated with another entity, José Freitas, the whole-part relation was inadequately percolated to the first named entity. There is no way to solve this type of errors at this time.

4.5.2 False-negatives

New Rules

Several situations had not been considered in the first stage of development of the rules, and were only detected during this phase of error analysis. Some, like the following case, are similar to cases we had already described, for example the meronymy with a dative pronoun:

(64) *Com um lenço de várias cores a cobrir-lhe os cabelos*

'With a scarf of many colors covering him the hairs = covering her hair'

The existing rules required the presence of a subject; a new, more general, rule was produced and placed at the end of the meronymy module, so that it will function as an heuristic to capture this type of cases.

```
IF( MOD[DAT](#3,#1[dat,cli]) &
    CDIR[POST](#3,#2[UMB-Anatomical-human]) &
    ~WHOLE-PART(#1,#2)
  )
    WHOLE-PART(#1,#2)
```

While in the previous case the *Nbp* was the direct object, a similar rule was required for the cases when the *Nbp* was a prepositional complement (MOD) of a subjectless verb:

(65) *Os dois homens, com idades compreendidas entre os 25 e os 30 anos, aproximaram-se de um passageiro e, encostando-lhe uma pistola ao corpo, obrigaram-no a entregar a carteira, que continha cerca de dez contos em dinheiro*

'Two men, aged between 25 and 30, approached a passenger and, putting a gun to his body, forced him to give them his wallet, which contained about ten thousand in cash'

```
IF ( MOD[DAT](#1,#2[dat,cli]) &
     MOD[POST](#1,#3[UMB-Anatomical-human]) & PREPD(#3,?) &
     ~WHOLE-PART(#2,#3)
   )
     WHOLE-PART(#2,#3)
```

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:

(66) `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'. The meronymy module did not contemplate these complex `NP`s, as most of the rules always involved a verb argument. This will have to be taken into consideration in future work.

The next case is even more complex: a `PP` with an *Nbp* depends on a human noun and not on a verb; however, this `PP` is also coordinated with an `AP` modifier of the same human noun. In this case, though the chunking is correct, the coordination rules fail to capture the coordination of `AP` and `PP`:

(67) *Rapazolas atléticos e de cabelo **preso** servem às mesas, onde se sentam os filhos daqueles que fazem de um estaleiro de obras local de férias e exemplares adulterados da etnia africana*
'Athletic young boys and with [their] hair **stuck** are serving at the tables, where the children of those who make from a construction site a vacation place and adulterated specimens of African ethnicity are sitted'

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:

(68) *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 were not triggered because the human entity is expressed by a personal pronoun and this category is not marked with the human feature: in Portuguese, 3^{rd} nominative person pronouns can refer both to humans and non-human entities.

(69) *E quando lhe digo que «em princípio» a culpa por este estado de coisas se deve aos autarcas, **ele** logo retorque, abanando a cabeça: "Sim, mas Portugal também é um todo ..."*
'And when I say 'in principle' the blame for this state of affairs is upon the mayors and town oficials, **he** quickly replies, shaking his head: "Yes, but Portugal is also a whole ... "'

If the information on the ViPEr verb class 09 of the verb *retorquir* 'retort' was used, it would be possible to assign the pronoun *ele* 'he' that feature, in order to make way for the whole-part rules to be triggered.

A similar case occurs with relative pronouns. In the next sentence, the system failed to establish the whole-part relation because it can not ascribe the human feature to the relative pronoun *que* 'who' that is the subject of the relative clause.

(70) *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)
```

According to [Marques-2013], 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

(71) *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 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*.

They do not solve all the cases, naturally, since the distinction between the determiner and the *Nbp* can not yet be done, as it would require a previous word sense disambiguation module.

Ambiguous `FIXED` Expressions, Incorrectly Captured

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

(72) *Ele arrancava-me os cabelos todos* 'He pulled out all my hair'

`FIXED(arrancava,cabelos)`

In this case, the correct relation should be: `WHOLE-PART(me,cabelos)`

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):

(73) *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 à lareira, o que poderá 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)`, which is beyond the scope of our current parser. Eventually, a bag-of-words machine learning approach could overcome this difficulty, which can not be done by this rule-based approach.

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':

(74) *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, naturally.

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

(75) *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, *cabelo(s) branco(s)* 'white hair(s)' is a compound noun, and it has not been given the *Nbp* feature; therefore, the system did not capture this element. Even so, the problem is in the apposition, since no dependency exists between the subject and the apposite; however as the subject also is

a pronoun, hence, no human feature is there to trigger the rules. Even if the compound was given the *Nbp* feature, this might not be entirely adequate, for the compound has a predicative function (it may be considered as a predicative noun), because of its idiomatic nature; *e.g.*, *O Pedro tem cabelos brancos* 'Peter has white hair', *Tu não respeitas os meus cabelos brancos* 'You do not respect my white hairs'.

In the next case, *corte de cabelo* 'haircut' is a compound noun, and though it involves an *Nbp* element, it is not clear whether a whole-part relation should be extracted or not. The original annotation directives were silent about such cases, and one of the annotators decided to consider a whole-part relation. Notice that this compound is largely synonym of *penteado* (*id*), but the word is derived from the instrument noun *pente* 'comb'.

(76) *Decididamente um tipo de suspensórios, com um corte de cabelo e corte de calça à maneira e um BMW, não podia ser visto a transaccionar pesos em público*
'[He was] definitely a guy with suspenders, with a haircut and very fashionable trousers and a BMW, so he could not be seen in public trading 'pesos' (currency)'

Upon reflexion, the golden standard was changed and the directives adapted to exclude explicitly all cases where a compound word involves an *Nbp*. Naturally, it presupposes that annotators know what a compound is, which is not obvious.

Typos

In the next case, there is a typo in the corpus, *v.g.*, *antigia* instead of *atingia* 'strike'.

(77) *Momentos depois, antigia mortalmente na cabeça um seu vizinho, José Maria Soares, agricultor de 77 anos, a trabalhar à porta de casa*
'Moments later, [he] fatally struck in the head one of his neighbors, José Maria Soares, a 77 years-old farmer, working at the doorstep'

`WHOLE-PART(vizinho,cabeça)`
If the typo was corrected, the system would have extracted the whole-part relation, as the annotator did; however, we decided not to change the corpus (using a spell-checker prior to the processing).

4.6 Post-Evaluation

Ones all the corrections were taking into consideration, we ran the system again in order to carry out the second evaluation of the system's performance. The results are shown in Table 4.9, where `TP`=*true-positives*; `TN`=*true-negatives*; `FP`=*false-positives*; `FN`=*false-negatives*.

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 (so we do not repeat them here). Since only some the errors detected 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 per-

Table 4.9: 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

formance will be achieved in future work.

In this chapter, we described in some detail the evaluation of the meronymy extraction module: the development of the corpus for the evaluation of whole-part relations extraction; the organization of the annotation campaign; the assessment of the inter-annotator agreement and of the whole-part dependencies extraction involving *Nbp* and *Nsick*; we also described how the error analysis was carried out and provided the results from a second evaluation of the system's performance.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This work addressed the problem of extraction of whole-part relations (*meronymy*), that is, a semantic relation between an entity that is perceived as a constituent part of another entity, or a member of a set. 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, and 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 [Mamede-et-al-2012].

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 did they handle the whole-part relations extraction: the PALAVRAS parser [Bick-2000], consulted using the Visual Interactive Syntax Learning (VISL) environment, and LX Semantic Role Labeller [Branco-and-Costa-2010]. 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 [Abreu-et-al-2013], no other mentions on whole-part relations extraction for Portuguese have been identified.

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 (two rules were added during the error analysis) 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*. 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. 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 themselves *Nbp* (*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 [Rocha-and-Santos-2000] (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 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 [Freelon-2010], 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 0.63. According to Landis and Koch [Landis-and-Koch-1977] this figures correspond to the lower bound of the "substantial" agreement; however, according to Fleiss [Fleiss-1981], 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 output.

After confronting the produced golden standard against the system's output, the results for *Nbp* show 0.57 precision, 0.38 recall, 0.46 F-measure, and 0.81 accuracy. 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; nevertheless, 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 presented in the detailed error analysis made after the systems' 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: some part of this task should be also performed after anaphora resolution, certainly producing better results.

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 0.5 precision, 0.11 recall, 0.17 F-measure, and 0.76 accuracy. A detailed error analysis was performed to determine the most relevant cases for these results, which led to some situations being implemented.

A second evaluation of the system's performance was carried out, and it 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.

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.

5.2 Future Work

In future work, the extraction of other types of whole-part relations will be addressed such as component-integral object (*pedal - bicycle*), member-collection (*player - team*), place-area (*grove - forest*), and others [Winston-et-al-1987]. The intention is also to use the list of *Nbp* provided by Cláudia Freitas [Freitas-2014] in order to complete the existing *Nbp* lexicon in STRING. As it was mentioned in section 2, Ittoo and Bouma [Ittoo-and-Bouma-2010] reported that focusing on particular type of whole-part relations in information extraction tasks gives more stable results than using general sets of whole-part relations as seeds for machine-learning algorithms. Follow this suggestion, other types of whole-part relations will be tackled, using already existing lexical sets in the STRING system (vehicles, human collective nouns, place-botanic, place-human building, place-geographic, tools, plants, animals, etc.). However, it is not obvious that for some of these classes of objects the strategy used here will be adequate; eventually, other strategies must be adopted such as a machine learning approach that will capture words associated to this lexical classes in patterns that are prone to be interpreted in this way.

Another line of future work will be the improvement of the 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, as it is currently being attempted at the L^2F group at INESC-ID Lisboa. 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 here built. 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.

Bibliography

[Abreu-et-al-2013] S. Abreu, T. Bonamigo, and R. Vieira. A review on Relation Extraction with an eye on Portuguese. *J. Braz. Comp. Soc.*, 19(4):553–571, 2013.

[Agirre-et-al-2009a] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings Human Language Technologies: 2009 Annual Conference of the North American Chapter of ACL (NAACL-HLT)*, pages 19–27. Stroudsburg, PA, USA. ACL Press, 2009.

[Agirre-et-al-2009b] E. Agirre, O. Lacalle, and A. Soroa. Knowledge-based WSD on specificcdomains: performing better than generic supervised WSD. In *In Proceedings of 21st International Joint Conference on Artificial Intelligence*, IJCAI 2009, pages 1501–1506. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 2009.

[Ait-Mokhtar-et-al-2002] S. Ait-Mokhtar, J. Chanod, and C. Roux. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8(2/3):121–144, 2002.

[Banerjee-and-Pedersen-2002] S. Banerjee and T. Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, volume 2276 of *LNCS*, pages 136–145. London, UK. Springer, 2002.

[Baptista-1997a] J. Baptista. Conversão, nomes parte-do-corpo e restruturação dativa. In I. Castro, editor, *Actas do XII Encontro da Associação Portuguesa de Linguística*, volume I – Linguística, pages 51–59, 1997.

[Baptista-1997b] J. Baptista. Sermão, tareia e facada: Uma classificação das construções conversas dar – levar. *Seminários de Linguística*, 1:5–37, 1997.

[Baptista-2012] J. Baptista. ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In J. Radimsky, editor, *Proceedings of the 31st International Conference on Lexis and Grammar*, pages 10–16. Università degli Studi di Salerno (Italy)/University of South Bohemia in Nové Hrady (Czech Republic), 2012.

[Baptista-et-al-2004] J. Baptista, A. Correia, and G. Fernandes. Frozen Sentences of Portuguese: Formal Descriptions for NLP. In *Workshop on Multiword Expressions: Integrating Processing. Interna-*

tional Conference of the European Chapter of the Association for Computational Linguistics, pages 72–79, Barcelona, Spain, 2004.

[Baptista-et-al-2005] J. Baptista, A. Correia, and G. Fernandes. Léxico Gramática das Frases Fixas do Portugués Europeo. *Cadernos de Fraseoloxía Galega*, 7:41–53, 2005.

[Baptista-et-al-2008] J. Baptista, C. Hagège, and N. Mamede. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. *Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro*, (Chapter 2):33–54, 2008.

[Baptista-et-al-2012a] J. Baptista, V. Cabarrão, and N. Mamede. Classification directives for Events and Relations Extraction between Named Entities in Portuguese texts. Technical report, Instituto Superior Técnico, Universidade do Algarve.

[Baptista-et-al-2012b] J. Baptista, N. Mamede, C. Hagège, and A. Maurício. Time Expressions in Portuguese. Guidelines for Identification, Classification and Normalization. Technical report, Universidade do Algarve, Instituto Superior Técnico, Xerox Research Centre Europe.

[Baptista-et-al-2014] J. Baptista, N. Mamede, and I. Markov. Integrating verbal idioms into an NLP system. In J. Baptista, N. Mamede, S. Candeias, I. Paraboni, T. Pardo, and M. Nunes, editors, *Computational Processing of Portuguese Language, PROPOR 2014*, LNAI/LNCS, São Carlos, SP, Brazil, 2014. Springer.

[Bellare-et-al-2004] K. Bellare, A. Sharma, N. Loiwal, and P. Bhattacharyya. Generic text summarization using WordNet. In *Proceedings of 4th International Conference on Language Resources and Evaluation*, LREC 2004, pages 691–694. Barcelona, Spain. ELRA, 2004.

[Berland-and-Charniak-1999] M. Berland and E. Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Morristown, NJ, USA. Association for Computational Linguistics, 1999.

[Bick-2000] E. Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University. Aarhus, Denmark: Aarhus University Press, 2000.

[Branco-and-Costa-2010] A. Branco and F. Costa. A Deep Linguistic Processing Grammar for Portuguese. In T. Pardo, A. Branco, A. Klautau, R. Vieira, and V. Lima, editors, *Computational Processing of Portuguese, PROPOR 2010*, LNAI/LNCS 6001, pages 86–89. Springer, 2010.

[Bruckschen-et-al-2008] M. Bruckschen, J. Guilherme Camargo de Souza, R. Vieira, and S. Rigo. Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas. In C. Mota and D. Santo, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, pages 247–260, 2008.

[Cabrita-et-al-2013] V. Cabrita, J. Baptista, and N. Mamede. Diretivas de classificação e anotação de corpora para a extração de relações entre eventos. Technical report, Instituto Superior Técnico.

[Carapinha-2013] F. Carapinha. Extração Automática de Conteúdos Documentais. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2013.

[Chomsky-1970] N. Chomsky. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Reading in English Transformational Grammar*, pages 184–221. Waltham: Ginn, 1970.

[Clark-et-al-2008] P. Clark, C. Fellbaum, and J. Hobbs. Using and extending WordNet to support question-answering. In *Proceedings of 4th Global WordNet Conference*, GWC 2008, pages 111–119. Szeged, Hungary, 2008.

[Cohen-1960] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[Dias-Da-Silva-and-Moraes-2003] B. Dias-Da-Silva and H. Moraes. A construção de um thesaurus eletrônico para o português do Brasil. *ALFA*, 47(2):101–115, 2003.

[Diniz-2010] C. Diniz. Um Conversor baseado em regras de transformação declarativas. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2010.

[Early-1970] J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[Elberrichi-et-al-2006] Z. Elberrichi, A. Rahmoun, and M. Bentaalah. Using WordNet for text categorization. *International Arab Journal of Information Technology*, 5(1):3–37, 2006.

[Esuli-and-Sebastiani-2007] A. Esuli and F. Sebastiani. PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of 45th Annual Meeting of the Association of Computational Linguistics*, ACL'07, pages 424–431. ACL Press, 2007.

[Fellbaum-1998] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT, Cambridge, 1998.

[Fellbaum-2010] C. Fellbaum. WordNet. In *Theory and Applications of Ontology: Computer Applications*, chapter 10, pages 231–243. Springer, 2010.

[Fleiss-1971] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[Fleiss-1981] J. Fleiss. *Statistical methods for rates and proportions*. New York: John Wiley, Heidelberg, second edition, 1981.

[Freelon-2010] D. Freelon. ReCal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science*, 5(1):20–33, 2010.

[Freitas-2014] C. Freitas. ESQUELETO - ANOTAÇÃO das palavras do corpo humano. Technical Report Versão 5: 20.05.2014.

URL http://www.linguateca.pt/acesso/Esqueleto.pdf

[Gerstl-and-Pribbenow-1995] P. Gerstl and S. Pribbenow. Midwinters, end games, and body parts: a classification of part-whole relations. *International Journal of Human Computer Studies*, 43:865–890, 1995.

[Girju-et-al-2003] R. Girju, A. Badulescu, and D. Moldovan. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In *Proceedings of HLT-NAACL*, volume 3, pages 80–87, 2003.

[Girju-et-al-2006] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 21(1):83–135, 2006.

[Gomes-et-al-2003] P. Gomes, F. Pereira, P. Paiva, N. Seco, P. Carreiro, J. Ferreira, and C. Bento. Noun Sense Disambiguation with WordNet for Software Design Retrieval. In *Proceedings of Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence*, pages 537–543. Halifax, Canada.

[Gross-1981] M. Gross. Les bases empiriques de la notion de prédicat sémantique. *Langages*, (63):7–52, 1981.

[Hagege-et-al-2008] C. Hagège, J. Baptista, and N. Mamede. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. *Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o Segundo HAREM*, pages 289–308, 2008.

[Hagege-et-al-2009] C. Hagège, J. Baptista, and N. Mamede. Portuguese Temporal Expressions Recognition: from TE characterization to an effective TER module implementation. *The 7th Brazilian Symposium in Information and Human Language Technology*, 2009.

[Hagege-et-al-2010] C. Hagège, J. Baptista, and N. Mamede. Caracterização e Processamento de Expressões Temporais em Português. *Linguamática*, 2(1):63–76, 2010.

[Harremoes-and-Topsoe-2001] P. Harremoës and F. Topsøe. Maximum Entropy Fundamentals. *Entropy*, 3(3):191–226, 2001.

[Hearst-1992] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics*, volume 2 of *COLING 92*, pages 539–545. Association for Computational Linguistics Morristown, NJ, USA, 1992.

[Hemayati-et-al-2007] R. Hemayati, W. Meng, and C. Yu. Semantic-based grouping of search engine results using WordNet. In *Proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management Conference on Advances in Data and Web Management*, APWeb/WAIM'07, pages 678–686. Springer, 2007.

[Hirst-2004] G. Hirst. Ontology and the lexicon. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 209–230. Springer, 2004.

[Iris-et-al-1988] M. Iris, B. Litowitz, and M. Evens. Problems of the Part-Whole Relation. In M. Evens, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, pages 261–288. Cambridge University Press, 1988.

[Ittoo-and-Bouma-2010] A. Ittoo and G. Bouma. On Learning Subtypes of the Part-Whole Relation: Do Not Mix your Seeds. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1328–1336. University of Groningen, 2010.

[Karlsson-1990] F. Karlsson. Constraint Grammar as a Framework for Parsing Unrestricted Text. In H. Karlgren, editor, *Proceedings of the 13th International Conference of Computational Linguistics*, volume 3, pages 168–173, Helsinki 1990, 1990.

[Keet-and-Artale-2008] M. Keet and A. Artale. Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110, 2008.

[Khoo-2006] C. Khoo and J. Na. Semantic Relations in Information Science. *Annual Review of Information Science and Technology*, 40:157–229, 2006.

[Landis-and-Koch-1977] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[Leclere-1995] C. Leclère. Sur une restructuration dative. *Language Research*, 31-1:179–198, 1995.

[Loureiro-2007] J. Loureiro. 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, 2007.

[Mamede-et-al-2012] N. Mamede, J. Baptista, C. Diniz, and V. Cabarrão. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *Computational Processing of Portuguese, PROPOR 2012*, volume Demo Session, Paper available at `http://www.propor2012.org/demos/DemoSTRING.pdf`, 2012.

[Marques-2013] J. Marques. Anaphora Resolution. Master's thesis, University of Lisbon/IST and INESC-ID Lisboa/L2F, 2013.

[Marrafa-2001] P. Marrafa. *WordNet do Português: uma base de dados de conhecimento linguístico*. Instituto Camões, 2001.

[Marrafa-2002] P. Marrafa. Portuguese WordNet: general architecture and internal semantic relations. *DELTA*, 18:131–146, 2002.

[Marrafa-et-al-2011] P. Marrafa, R. Amaro, and S. Mendes. WordNet.PT Global – extending WordNet.PT to Portuguese varieties. In *Proceedings of the 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 70–74, Edinburgh, Scotland. ACL Press, 2011.

[Mauricio-2011] A. Maurício. Identificação, Classificação e Normalização de Expressões Temporais. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2011.

[Maziero-et-al-2008] E. Maziero, T. Pardo, A. Felippo, and B. Dias da Silva. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392, 2008.

[Miller-1995] G. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[Miller-et-al-1993] G. Miller, C. Leacock, R. Tengi, and R. Bunker. A Semantic Concordance. In *Proceeding HLT '93 Proceedings of the workshop on Human Language Technology*, pages 303–308, 1993.

[Navigli-and-Velardi-2003] R. Navigli and P. Velardi. An analysis of ontology-based query expansion strategies. In *Proceedings of the ECML 2003 Workshop on Adaptive Text Extraction and Mining (ATEM) in the 14th European Conference on Machine Learning*, pages 42–49. Cavtat-Dubrovnik, Croatia, 2003.

[Odell-1994] J. Odell. Six different kinds of composition. *Journal of Object-Oriented Programming*, 5(8):10–15, 1994.

[Oliveira-2010] D. Oliveira. Extraction and Classification of Named Entities. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2010.

[Oliveira-2012] H. Oliveira. *Onto.PT: Towards the Automatic Construction of a Lexical Ontology for Portuguese*. PhD thesis, University of Coimbra/FST, 2012.

[Oliveira-and-Gomes-2008] H. Oliveira and P. Gomes. Utilização do (analisador sintáctico) PEN para extracção de informação das deniçõ es de um dicionário. Technical report, Linguateca, pólo de Coimbra, DEI - FCTUC, CISUC, 2008.

[Oliveira-et-al-2008] H. Oliveira, P. Gomes, D. Santos, and N. Seco. PAPEL: A Dictionary-based Lexical Ontology for Portuguese. In A. Teixeira, V. Lima, L. Oliveira, and P. Quaresm, editors, *Computational Processing of the Portuguese Language, PROPOR 2008*, volume 5190 of *LNAI/LNCS 5190*, pages 31–40, Aveiro, Portugal, 2008. Springer.

[Pantel-and-Pennacchiotti-2006] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113–120. Sydney, Australia, 2006.

[Pasca-and-Harabagiu-2001] M. Pasca and S. Harabagiu. The informative role of WordNet in open-domain question answering. In *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pages 138–143. Pittsburgh, USA, 2001.

[Paumier-2003] S. Paumier. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. PhD thesis, Université de Marne-la-Vallée, 2000.

[Paumier-2014] S. Paumier. *Unitex 3.1beta, User Manual*. Univ. Paris-Est Marne-la-Vallée, 2014.

[Pianta-et-al-2002] E. Pianta, L. Bentivogli, and C. Girardi. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the 1st International WordNet Conference*, pages 293–302, Mysore, India, 2002.

[Plaza-et-al-2010] L. Plaza, A. Díaz, and P. Gervás. Automatic summarization of news using WordNet concept graphs. *International Journal on Computer Science and Information System (IADIS)*, V:45–57, 2010.

[Prevot-et-al-2010] L. Prévot, C. Huang, N. Calzolari, A. Gangemi, A. Lenci, and A. Oltramari. Ontology and the lexicon: a multi-disciplinary perspective (introduction). In C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prévot, editors, *Ontology and the Lexicon: A Natural Language Processing Perspective*, Studies in Natural Language Processing, chapter 1, pages 3–24. Cambridge University Press, 2010.

[Ranchhod-1990] E. Ranchhod. *Sintaxe dos predicados nominais com Estar*. Lisboa: INIC - Instituto Nacional de Investigação Científica, 1990.

[Resnik-1995] P. Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of 3rd Workshop on Very Large Corpora*, pages 54–68. Cambridge, MA, USA, 1995.

[Ribeiro-2003] R. Ribeiro. Anotação Morfossintáctica Desambiguada do Português. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2003.

[Richardson-et-al-1998] S. Richardson, W. Dolan, and L. Vanderwende. MindNet: Acquiring and structuring semantic information from text. In *Proceedings of 17th International Conference on Computational Linguistics*, pages 1098–1102. COLING'98, 1998.

[Rocha-and-Santos-2000] P. Rocha and D. Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In M. Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR 2000*, pages 131–140. São Paulo: ICMC/USP, 2000.

[Romao-2007] L. Romão. Reconhecimento de Entidades Mencionadas em Língua Portuguesa: Locais, Pessoas, Organizações e Acontecimentos. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2007.

[Rosso-et-al-2004] P. Rosso, E. Ferretti, D. Jiménez, and V. Vidal. Text categorization and information retrieval using WordNet senses. In *Proceedings of 2nd Global WordNet Conference*, GWC 2004, pages 299–304, 2004.

[Santos-2010] D. Santos. Extracção de relações entre entidades mencionadas. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2010.

[Santos-et-al-2010] D. Santos, A. Barreiro, C. Freitas, H. Oliveira, J. Medeiros, L. Costa, P. Gomes, and R. Silva. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística*, pages 681–700. APL, Lisboa, Portugal, 2010.

[Seco-et-al-2004] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in WordNet. In *Proceedings of 16th European Conference on Artificial Intelligence*, ECAI 2004, pages 1089–1090. Valencia, Spain. IOS Press, 2004.

[Silva-et-al-2010] J. Silva, A. Branco, S. Castro, and R. Reis. Out-of-the-box robust parsing of Portuguese. In *Computational Processing of the Portuguese Language, PROPOR 2010*, pages 75–85. Porto Alegre, RS, Brazil, 2010.

[Talhadas-2014] R. Talhadas. Semantic Role Labelling in European Portuguese. Master's thesis, Universidade do Algarve/FCHS, 2014.

[Travanca-2013] T. Travanca. Verb Sense Disambiguation. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2013.

[Van-Hage-et-al-2006] W. Van Hage, H. Kolb, and G. Schreiber. A method for learning part-whole relations. *The Semantic Web - ISWC 2006, LNAI/LNCS*, 4273:723–725, 2006.

[Vanderwende-1995] L. Vanderwende. Ambiguity in the acquisition of lexical information. In *Proceedings of the AAAI 1995 Spring Symposium, Working notes of the symposium on representation and acquisition of lexical knowledge*, pages 174–179, 1995.

[Vanderwende-et-al-2005] L. Vanderwende, G. Kacmarcik, H. Suzuki, and A. Menezes. MindNet: An Automatically-Created Lexical Resource. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 8–9. Association for Computational Linguistics, 2005.

[Vicente-2013] A. Vicente. LexMan: um Segmentador e Analisador Morfológico com transdutores. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, 2013.

[Viterbi-1967] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260—-269, 1967.

[Voorhees-1998] E. Voorhees. Using WordNet for Text Retrieval. In *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, pages 285–303. The MIT Press, 1998.

[Vossen-1997] P. Vossen. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of DELOS workshop on Cross-Language Information Retrieval*, pages 5–7, Zurich, 1997.

[Williams-and-Anand-2009] G. Williams and S. Anand. Predicting the Polarity Strength of Adjectives Using WordNet. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, ICWSM 2009, pages 346–349. San Jose, California, USA. AAAI Press, 2009.

[Winston-et-al-1987] M. Winston, R. Chaffin, and D. Herrmann. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11:417–444, 1987.

[Zhang-et-al-2010] L. Zhang, B. Liu, S. Hwan Lim, and E. O'Brien-Strain. Extracting and ranking product features in opinion documents. In *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1462–1470, Stroudsburg, PA, USA, 2010.

Appendix A

Nbp Whole-Part Extraction Rules

A.1 General Rules

```
//1. Example: O Pedro roeu os seus cantos das unhas.
//---> WHOLE-PART(seus,unhas)
//---> WHOLE-PART(unha,cantos)
IF ( MOD[POST](#1[npart],#2[UMB-Anatomical-human]) &
     PREPD(#2,?[lemma:de]) &
     ^POSS[PRE](#1[npart],#4[poss]) &
     ~WHOLE-PART(#2,#1) &
     ~WHOLE-PART(#4,#1) &
   )
     POSS[pre](#2,#4),
     WHOLE-PART(#2,#1),
     WHOLE-PART(#4,#2)

//2. Example: O Pedro roeu o canto da unha.
//---> WHOLE-PART(Pedro,unha)
//---> WHOLE-PART(unha,canto)
IF ( MOD[POST](#1[npart],#2[UMB-Anatomical-human]) &
     PREPD(#2,?[lemma:de]) &
     ~WHOLE-PART(#2,#1) & ~POSS[PRE](#1,#4[poss]) &
     ~FIXED(#5,#3) &
     ^CDIR(#3,#1)
   )
     CDIR(#3,#2),
     WHOLE-PART(#2,#1)

//3. Example: O canto da sua unha infetou.
//---> WHOLE-PART(sua,unha)
//---> WHOLE-PART(unha,canto)
IF ( MOD[POST](#1[npart],#2[UMB-Anatomical-human]) &
     PREPD(#2,?[lemma:de]) &
     ~WHOLE-PART(#2,#1) &
     ^SUBJ(#3,#1)
   )
     SUBJ(#3,#2),
     WHOLE-PART(#2,#1)

//4. Example: O Pedro esgravatou no canto da unha
//---> MOD_POST(esgravatou,unha)
//---> MOD_POST(unha,canto)
```

```
//---> WHOLE-PART(Pedro,unha)
//---> WHOLE-PART(unha,canto)
IF ( VDOMAIN(#1,#2) &
     SUBJ(#2,#7) &
     MOD(#2,#3) &
     MOD[POST](#3[npart],#4[UMB-Anatomical-human]) &
     PREPD(#4,?[lemma:de]) &
     ~WHOLE-PART(#4,#3) &
     ~POSS[PRE](#4,#5[poss]) &
     ~CINDIR(#2,#6) &
   )
     MOD(#2,#4),
     WHOLE-PART[POST=~](#7,#4),
     WHOLE-PART(#4,#3)

//4a.
IF ( VDOMAIN(#1,#2) &
     SUBJ(#2,#7) &
     ^MOD(#2,#3) &
     MOD[POST](#3[npart],#4[UMB-Anatomical-human]) & PREPD(#4,?[lemma:de]) &
     WHOLE-PART(#4,#3) & ~POSS[PRE](#4,#5[poss]) & ~CINDIR(#2,#6) &
   )
     ~

//5. Example: O Pedro esgravatou no canto da unha
// This is a general rule to change the MOD of an NP de NP sequence
// involving a [npart] and a Nbp
//---> WHOLE-PART(Pedro,unha)
//---> WHOLE-PART(unha,canto)
IF ( MOD[POST](#1[npart],#2[UMB-Anatomical-human]) &
     PREPD(#2,?[lemma:de]) &
     ~WHOLE-PART(#2,#1) & ~POSS[PRE](#1,#4[poss]) &
     ^MOD(#5,#1)
   )
     MOD(#5,#2),
     WHOLE-PART(#2,#1)

//6. Example: O Pedro partiu o braço ao João. ---> WHOLE-PART(João,braço)
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)

//7. Example: O Pedro partiu o braço do João. ---> WHOLE-PART(João,braço)
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)

//8. Example: O Pedro partiu o braço dele. ---> WHOLE-PART(ele,braço)
IF ( MOD[POST](#2[UMB-Anatomical-human],#1[obl,3p]) &
     PREPD(#1,?[lemma:de]) &
```

```
     CDIR[POST](#3,#2) &
     ~WHOLE-PART(#1,#2)
   )
     WHOLE-PART(#1,#2)

//9. Example: O Pedro partiu o seu braço. ---> WHOLE-PART(seu,braço)
IF ( POSS[PRE](#2[UMB-Anatomical-human],#1[poss]) &
     ~WHOLE-PART(#1,#2) &
   )
     WHOLE-PART(#1,#2)

//10. Example: O Pedro partiu-lhe o braço. ---> WHOLE-PART(lhe,braço)
IF ( ^MOD[DAT](#3,#1[dat,cli]) &
     SUBJ[PRE] (#3,#6) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     ~PREPD(#5,#7[lemma:de]) &
     ~MOD(#2,#5) &
     ~SUBJ[elips](#3,#4) &
     ~CINDIR(#3,#1) &
     ~WHOLE-PART(#1,#2)
   )
     CINDIR[DAT=~](#3,#1),
     WHOLE-PART(#1,#2)

//11. Example: O Pedro não lhe partiu o braço. ---> WHOLE-PART(lhe,braço)
// CINDIR(partiu,lhe)
// WHOLE-PART(lhe,braços)
// There must be a subject that is not an elipsis,
//so that we can inforce the SUBJ[elips] later and zero it.
IF ( CLITIC[PRE](#3,#1[dat]) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     SUBJ[PRE](#3,#4) &
     ~SUBJ[elips](#3,#5) &
     ~CINDIR(#3,#1) &
     ~PREPD(#6,#7[lemma:de]) &
     ~MOD(#2,#6) &
     ~WHOLE-PART(#1,#2)
   )
     CINDIR(#3,#1),
     WHOLE-PART(#1,#2)

//12. Example: O braço do João está partido. ---> WHOLE-PART(João,braço)
IF ( MOD[POST](#2[UMB-Anatomical-human],#1[human]) &
     PREPD(#1,?[lemma:de]) &
     ~WHOLE-PART(#1,#2) &
     ~WHOLE-PART(#1,#3) &
     ~WHOLE-PART(#4,#2)
   )
     WHOLE-PART(#1,#2)

//13. Example: O braço dele está partido. ---> WHOLE-PART(ele,braço)
IF ( MOD[POST](#2[UMB-Anatomical-human],#1[obl,3p]) &
     PREPD(#1,?[lemma:de]) &
     ~WHOLE-PART(#1,#2) &
     ~WHOLE-PART(#3,#2)
   )
     WHOLE-PART(#1,#2)
```

```
//14. Example: Os braços doem-me. ---> WHOLE-PART(me,braços)
IF ( ^MOD[DAT](#3,#1[dat,cli]) &
     SUBJ[PRE](#3,#2[UMB-Anatomical-human]) &
     ~CINDIR(#3,#1) &
     ~WHOLE-PART(#1,#2)
   )
     CINDIR[DAT=~](#3,#1),
     WHOLE-PART(#1,#2)

//15. Example: Doem-me os braços. ---> WHOLE-PART(me,braços)
//    CINDIR_POST(Doem,me)
//    SUBJ_POST(Doem,braços) (note: the SUBJ_EPLIPS is to be zeroed)
//    WHOLE-PART(braços,me)
IF ( ^MOD[DAT](#3,#1[dat,cli]) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     SUBJ[ELIPS](#3,#4) &
     ~SUBJ(#3,#2) &
     ~CINDIR(#3,#1) &
     ~WHOLE-PART(#1,#2)
   )
     CINDIR[DAT=~](#3,#1),
     SUBJ[POST=+](#3,#2),
     WHOLE-PART(#1,#2)

//15a.
IF ( CINDIR(#3,#1) &
     ^CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     SUBJ[POST](#3,#2) &
     WHOLE-PART(#1,#2)
   )
     ~

//16. Os braços não me doem. ---> WHOLE-PART(me,braços)
IF ( ^CLITIC[PRE](#3,#1[dat]) &
     SUBJ[PRE](#3,#2[UMB-Anatomical-human]) &
     ~CINDIR(#3,#1) &
     ~WHOLE-PART(#1,#2)
   )
     CINDIR(#3,#1),
     WHOLE-PART(#1,#2)

//17. Não me doem os braços. ---> WHOLE-PART(me,braços)
//    CINDIR_POST(doem,me)
//    SUBJ_POST(doem,braços)
//    WHOLE-PART(braços,me)
IF ( CLITIC[PRE](#3,#1[dat]) &
     ^CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     ~CINDIR(#3,#1) &
     ~WHOLE-PART(#1,#2)
   )
     SUBJ[POST=+](#3,#2),
     CINDIR(#3,#1),
     WHOLE-PART(#1,#2)

//18. Example: O Pedro partiu o braço. ---> WHOLE-PART(Pedro,braço)
IF ( SUBJ[PRE](#3,#1[human]) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     ~PREPD(#5,#6[lemma:de]) &
```

```
      ~MOD(#2,#5) &
      ~WHOLE-PART(#1,#2) &
      ~WHOLE-PART(#4,#2)
   )
      WHOLE-PART(#1,#2)

//19. Example: Este brasileiro de pernas altas. ---> WHOLE-PART(brasileiro,pernas)
IF ( MOD[POST](#1[human],#2[UMB-Anatomical-human]) &
      PREPD(#2,?[lemma:de]) &
      ~WHOLE-PART(#1,#2)
   )
      WHOLE-PART(#1,#2)

//20. O Pedro feriu-se no braço ---> WHOLE-PART(se,braço)
IF ( CLITIC(#3,#1[cli,ref]) &
      SUBJ[PRE](#3,#6) &
      MOD[POST](#3,#2[UMB-Anatomical-human]) &
      PREPD(#2,#4[lemma:em]) &
      ~PREPD(#5,#7[lemma:de]) &
      ~MOD(#2,#5) &
      ~WHOLE-PART(#6,#2)
   )
      WHOLE-PART(#6,#2)

//21. Example: O Pedro bateu-me nas pernas. ---> WHOLE-PART(me,pernas)
IF ( CLITIC(#3,#1[cli,ref:~]) &
      SUBJ[PRE](#3,#6) &
      MOD[POST](#3,#2[UMB-Anatomical-human]) &
      PREPD(#2,#4[lemma:em]) &
      ~PREPD(#5,#7[lemma:de]) &
      ~MOD(#2,#5) &
      ~WHOLE-PART(#1,#2)
   )
      WHOLE-PART(#1,#2)

//22. Example: O Zé andava de cabeça erguida ---> WHOLE-PART(Zé,cabeça)
IF ( VDOMAIN(#1,#2[cop]) &
      SUBJ(#2,#3) &
      PREDSUBJ(#2,#4[UMB-Anatomical-human]) &
      MOD[POST](#5[prep],#4) &
      ~WHOLE-PART(#3,#4)
   )
      WHOLE-PART(#3,#4)

//23. Example: O Pedro levava o Zé pela mão. ---> WHOLE-PART(Zé,mão)
IF ( VDOMAIN(#1,#2) &
      CDIR(#2,#3[human]) &
      MOD[post](#2,#4[UMB-Anatomical-human]) &
      ~WHOLE-PART(?,#4) &
      ~WHOLE-PART(#3,#4)
   )
      WHOLE-PART(#3,#4)

//24. Example: A Ana pinta as unhas dos pés. ---> WHOLE-PART(pés,unhas)
If ( MOD(#1[UMB-Anatomical-human],#2[UMB-Anatomical-human]) &
      PREPD(#2,#3[lemma:de]) &
      ~WHOLE-PART(#2,#1)
   )
```

```
     WHOLE-PART(#2,#1)

//25. O Pedro comparou o comprimento da mão direita com o da mão esquerda.
//---> WHOLE-PART(Pedro,mão direita)
//---> WHOLE-PART(Pedro,mão esquerda)
IF ( VDOMAIN(#1,#2) &
     SUBJ(2,#3[human]) &
     ?(#2,#6) &
     MOD(#6,#4[UMB-Anatomical-human]) &
     PREPD (#4,#7[lemma:de]) &
     ( ~MOD(#4,#5[human]) || ~CINDIR(#2,#5) ) &
     ~WHOLE-PART(#3,#4) &
     ~WHOLE-PART(#8,#4)
   )
     WHOLE-PART(#3,#4)

//26. Example: O Pedro coçou na cabeça
//---> WHOLE-PART (Pedro, cabeça)
IF ( MOD[post](#1,#2[UMB-Anatomical-human]) &
     SUBJ[pre](#1,#3[human]) &
     ~WHOLE-PART(#3,#2) &
     ~POSS[pre](#2,#4[poss]) &
     ( ~MOD[post](#2,#5[human]) || ~PREPD(#5,#6[lemma:de]) ) &
     ~CDIR(#1,#7[human]) &
     ~CDIR(#1,#8[acc]) &
     ~CINDIR(#1,#9) &
     ~MOD[dat](#1,#10)
   )
     WHOLE-PART(#3,#2)

//27. Example: O Pedro espalhou óleo nas pernas à Joana
//---> WHOLE-PART(João,pernas)
IF ( MOD[post](#1,#2[UMB-Anatomical-human]) &
     PREPD(#2,#5[lemma:em]) &
     MOD[post](#1,#3[human]) &
     PREPD(#3,#6[lemma:a]) &
     SUBJ[pre](#1,#4[human]) &
     ~WHOLE-PART(#3,#2) &
     ~POSS[pre](#2,#7[poss]) &
     ~CDIR(#1,#10[human]) &
     ~CINDIR(#1,#11)
   )
     WHOLE-PART(#3,#2)

//28. Example: Com um lenço de várias cores a cobrir-lhe os cabelos
IF ( MOD[DAT](#3,#1[dat,cli]) &
     CDIR[POST](#3,#2[UMB-Anatomical-human]) &
     ~WHOLE-PART(#1,#2)
   )
     WHOLE-PART(#1,#2)

//29. Example: O Pedro encostou-lhe uma pistola ao corpo
IF ( MOD[DAT](#1,#2[dat,cli]) &
     MOD[POST](#1,#3[UMB-Anatomical-human]) & PREPD(#3,?) &
     ~WHOLE-PART(#2,#3)
   )
     WHOLE-PART(#2,#3)
```

A.2 Disease Nouns

A set of 87 rules has been build for the 29 disease nouns and their corresponding hidden *Nbp* (*e.g.*, *gastrite* 'gastritis' - *estômago* 'stomach'). Only the rules for *gastrite* 'gastritis' are shown below. The list of the disease nouns and their corresponding hidden *Nbp* is presented afterwards.

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

//2. Example: O Pedro está com uma gastrite.
IF( MOD[POST](#1[lemma:estar],#2[lemma:gastrite]) &
    PREPD(#2,?[lemma:com]) &
    SUBJ[PRE](#1,#3) &
    ~WHOLE-PART(#3,?)
  )
    WHOLE-PART[hidden=+](#3,##noun#[surface:estômago,lemma:estômago])

//3. Example: A gastrite do Pedro é grave.
IF( MOD[POST](#2[lemma:gastrite],#3[human]) &
    PREPD(#3,?[lemma:de]) &
    ~WHOLE-PART(#3,?)
  )
    WHOLE-PART[hidden=+](#3,##noun#[surface:estômago,lemma:estômago])
```

Artrite - articulação; bronquite - brônquio; cardiosclerose - coração; cistite - bexiga; colecistite - vesícula; conjuntivite - olho; dermatite - pele; diabetes - pâncreas; endarterite - artéria; faringite - faringe; gastrite - estômago; glomerulonefrite - rim; hemorróidas - ânus; hepatite - fígado; miastenia - músculo; neurite - nervo; nevrite - nervo; osteocondrose - osso; osteomielite - osso; osteoporose - osso; otite - ouvido; pancreatite - pâncreas; periodontite - periodonto; pielonefrite - rim; pleurisia - pleura; prostatite - próstata; rinite - nariz; tonsilite - amígdala; traqueíte - traquéia.

Appendix B

Nbp Lexicon

B.1 Parts of *Nbp*

```
1> noun[lemma:alto,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cabeça].
1> noun[lemma:alto,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:alto,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
1> noun[lemma:ápice,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:asa,npart=+], (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:nariz].
1> noun[lemma:asa,npart=+], (adj[lemma:direito]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:nariz].
2> noun[lemma:barriga,npart=+,sem-anorg=~], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:perna].
2> noun[lemma:base,npart=+,sfazer=~], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:coluna].
2> noun[lemma:base,npart=+,sfazer=~], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pescoço].
1> noun[lemma:cana,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:nariz].
1> noun[lemma:canto,npart=+], (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:boca].
1> noun[lemma:canto,npart=+], (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:olho].
1> noun[lemma:canto,npart=+], (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:unha].
1> noun[lemma:canto,npart=+], (adj[lemma:direito]), prep[lemma:de], art[lemma:o],  (pron[poss]), noun[lemma:boca].
1> noun[lemma:canto,npart=+], (adj[lemma:direito]), prep[lemma:de], art[lemma:o],  (pron[poss]), noun[lemma:olho].
1> noun[lemma:canto,npart=+], (adj[lemma:direito]), prep[lemma:de], art[lemma:o],  (pron[poss]), noun[lemma:unha].
1> noun[lemma:canto,npart=+], (adj[lemma:interno]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:boca].
1> noun[lemma:canto,npart=+], (adj[lemma:interno]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:olho].
1> noun[lemma:canto,npart=+], (adj[lemma:interno]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:unha].
1> noun[lemma:canto,npart=+], (adj[lemma:externo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:boca].
1> noun[lemma:canto,npart=+], (adj[lemma:externo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:olho].
1> noun[lemma:canto,npart=+], (adj[lemma:externo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:unha].
2> noun[lemma:coroa,npart=+,sem-percep-w=~,sem-mon=~,sem-cloh-hat=~,sem-currency=~], prep[lemma:de], art[lemma:o],
(pron[poss]), noun[lemma:língua].
1> noun[lemma:costas,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:mão].
1> noun[lemma:coto,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:perna].
1> noun[lemma:cova,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
1> noun[surface:covinha,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:queixo].
1> noun[surface:covinha,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:rosto].
1> noun[lemma:dorso,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:dorso,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:mão].
1> noun[lemma:dorso,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
2> noun[lemma:face,npart=+,sem-anmov=~,prep1a=~,n1nhum=~,n0hum=~,sfazer=~,nn1=~], (adj[lemma:externo]),
prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:coxa].
2> noun[lemma:face,npart=+,sem-anmov=~], (adj[lemma:interno]), prep[lemma:de], art[lemma:o], (pron[poss]),
noun[lemma:coxa].
1> noun[lemma:freio,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:lado,npart=+],  (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cabeça].
1> noun[lemma:lado,npart=+],  (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cara].
```

```
1> noun[lemma:lado,npart=+],  (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:lado,npart=+],  (adj[lemma:esquerdo]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:tronco].
1> noun[lemma:lado,npart=+],  (adj[lemma:direito]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cabeça].
1> noun[lemma:lado,npart=+],  (adj[lemma:direito]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cara].
1> noun[lemma:lado,npart=+],  (adj[lemma:direito]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:lado,npart=+],  (adj[lemma:direito]), prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:tronco].
1> noun[lemma:lóbulo,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:orelha].
1> noun[lemma:palma,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:mão].
2> noun[lemma:peito,npart=+,sem-an=~,sem-am=~], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
1> noun[lemma:planta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
1> noun[lemma:ponta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cabelo].
1> noun[lemma:ponta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:dedo].
1> noun[lemma:ponta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:língua].
1> noun[lemma:ponta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:nariz].
1> noun[lemma:ponta,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:pé].
1> noun[lemma:rabo,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:olho].
1> noun[lemma:raiz,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:cabelo].
1> noun[lemma:sabugo,npart=+], prep[lemma:de], art[lemma:o], (pron[poss]), noun[lemma:unha].
```

B.2 *Nbp* Disambiguation

```
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Molière].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Cervantes].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Goethe].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Dante].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Racine].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Rimbaud].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Cícero].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Virgílio].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Bocaccio].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Tolstoi].
2> noun[lemma:língua,sem-anmov=~], prep[lemma:de], noun[lemma:Jesus].
```

Appendix C

Distribution of *Nbp*

Table C.1: Distribution of *Nbp*.

Nbp			
Lemma	**Count 1 (in the corpus)**	**%**	**Count 2 (selected)**
alvéolo pulmonar 'pulmonary alveoli'	1	0.01	0
anca 'hip'	14	0.11	1
aparelho circulatório 'circulatory system'	1	0.01	0
aparelho digestivo 'digestive system'	1	0.01	0
aparelho urinário 'urinary tract'	1	0.01	0
artéria 'artery'	73	0.58	5
baço 'spleen'	12	0.09	1
barba 'beard'	70	0.55	5
barriga 'belly'	38	0.30	3
bexiga 'bladder'	16	0.13	1
boca 'mouth'	282	2.23	22
braço 'arm'	420	3.32	33
brônquio 'bronchus'	2	0.02	1
cabeça 'head'	970	7.66	76
cabelo 'hair'	180	1.42	14
calcanhar 'heel'	26	0.21	2
canela 'shin'	27	0.21	2
cara 'face'	396	3.13	31
cérebro 'brain'	97	0.77	7
cintura 'waist'	66	0.52	5
clitóris 'clitoris'	1	0.01	0
colo 'lap'	37	0.29	2

Continued on next page

Table C.1 – *Continued from previous page*

Lemma	Count 1 (in the corpus)	%	Count 2 (selected)
cólon 'colon'	10	0.08	1
coluna 'spine'	140	1.11	11
coração 'heart'	416	3.29	32
corpo 'body'	1,116	8.82	88
costas 'back'	286	2.26	22
costela 'rib'	16	0.13	1
cotovelo 'elbow'	10	0.08	1
coxa 'thigh'	24	0.19	1
crânio 'skull'	22	0.17	1
dedo 'finger'	168	1.33	13
dedo indicador 'forefinger'	2	0.02	1
dedo médio 'middle finger'	1	0.01	0
dedo polegar 'thumb'	1	0.01	0
dente 'tooth'	91	0.72	7
derme 'derm'	1	0.01	0
duodeno 'duodenum'	2	0.02	1
esófago 'esophagus'	6	0.05	1
espinha 'spine'	23	0.18	1
esqueleto 'skeleton'	40	0.32	3
estômago 'stomach'	42	0.33	3
face 'face'	1,362	10.76	107
fígado 'liver'	28	0.22	2
garganta 'throat'	49	0.39	3
glândula 'gland'	3	0.02	1
joelho 'knee'	77	0.61	6
lábio 'lip'	47	0.37	3
laringe 'larynx'	3	0.02	1
língua 'tongue'	683	5.40	53
mama 'breast'	40	0.32	3
mamilo 'nipple'	2	0.02	1
mandíbula 'mandible'	2	0.02	1
mão 'hand'	1,525	12.05	120
mão direita 'right hand'	24	0.19	1
mão esquerda 'left hand'	16	0.13	1
maxilar 'jaw'	7	0.06	1

Continued on next page

Table C.1 – *Continued from previous page*

Lemma	Count 1 (in the corpus)	%	Count 2 (selected)
membrana 'membrane'	10	0.08	1
músculo 'muscle'	42	0.33	3
nariz 'nose'	74	0.58	5
nervo 'nerve'	63	0.50	4
olho 'eye'	655	5.17	51
ombro 'shoulder'	89	0.70	7
orelha 'ear'	60	0.47	4
osso 'bone'	107	0.85	8
ouvido 'ear'	340	2.69	26
ovário 'ovary'	2	0.02	1
pâncreas 'pancreas'	10	0.08	1
pé 'foot'	721	5.70	56
peito 'chest'	88	0.70	6
pele 'skin'	200	1.58	15
pelo púbico 'pubes'	1	0.01	0
pénis 'penis'	23	0.18	1
perna 'leg'	202	1.60	15
pescoço 'neck'	59	0.47	4
pestana 'eyelash'	52	0.41	4
prepúcio 'foreskin'	1	0.01	0
próstata 'prostate'	12	0.09	1
pulmão 'lung'	71	0.56	5
pulso 'pulse'	36	0.28	2
punho 'fist'	50	0.39	3
queixo 'chin'	14	0.11	1
reto 'rectum'	2	0.02	1
rim 'kidney'	10	0.08	1
rosto 'countenance'	249	1.97	19
seio 'bosom'	183	1.45	14
sobrancelha 'eyebrow'	5	0.04	1
tarso 'tarsus'	1	0.01	1
têmpora 'têmpora'	4	0.03	1
testa 'forehead'	29	0.23	2
testículo 'testicle'	7	0.06	1
timo 'thyme'	6	0.05	1

Continued on next page

Table C.1 – *Continued from previous page*

Lemma	Count 1 (in the corpus)	%	Count 2 (selected)
tornozelo 'ankle'	13	0.10	1
traqueia 'trachea'	2	0.02	1
tronco 'trunk'	46	0.36	3
umbigo 'navel'	7	0.06	1
úmero 'humerus'	1	0.01	1
unha 'nail'	27	0.21	2
útero 'uterus'	22	0.17	1
vagina 'vagina'	6	0.05	1
veia 'vein'	24	0.19	1
ventre 'belly'	15	0.12	1
vesícula 'gallbladder'	2	0.02	1
Total:	**12,659**	**100**	**983**

Appendix D

Annotation Guidelines

Dear Annotator,

First of all, I would like to say that I appreciate very much your help in annotating this file.

The purpose of this annotation is to identify whole-part relations involving human nouns and body part nouns.

The file consists of **325** sentences from a newspaper corpus. Each sentence contains a human body part noun (*Nbp*), like *mão* 'hand', or a disease noun (*Nsick*), like *hepatite* 'hepatitis'.

There are different cases that require extraction of whole-part relation:

1. If there is a human noun in the sentence to whom the *Nbp* belongs, the whole-part relation should be established between the human noun and the *Nbp*:

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

`WHOLE-PART(João,braço)`

*O **Pedro** partiu o **braço*** 'Pedro broke [his] arm'

`WHOLE-PART(Pedro,braço)`

2. In some cases, instead of a noun, we find a pronoun; in that case, the whole-part relation should mention this pronoun:

*O **braço dele** está partido* (lit: The arm of him is broken) 'His arm is broken'

`WHOLE-PART(ele,braço)`

*O Pedro partiu-**lhe** o **braço*** 'Pedro broke him the arm'

`WHOLE-PART(lhe,braço)`

*O Pedro partiu o **teu braço*** 'Pedro broke your arm'

`WHOLE-PART(teu,braço)`

3. There may be a relation within the same sentence between different *Nbp*. In these cases, the whole-part relation should be established not only between the human noun and one of the *Nbp*, but also between the two *Nbp* in the sentence:

*A **Ana** pinta as **unhas** dos **pés*** (lit: Ana paints the nails of the feet) 'Ana paints her toes' nails'

`WHOLE-PART(Ana,unhas)`

`WHOLE-PART(pés,unhas)`

4. There may be a relation within the same sentence between an *Nbp* and a noun that designates a part of that same *Nbp*. In these cases, the whole-part relation should be established between the human noun and the *Nbp*, and a second whole-part relation should also be established between the determinative part of the *Nbp* and the *Nbp* itself.

Notice that in this case, the meaning of the sentence is **not** equivalent to *A Ana pinta os pés*. On the other hand, certain nouns that designate parts of *Nbp* allow this interpretation:

*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) - correct`

`WHOLE-PART(língua,ponta) - correct`

`WHOLE-PART(Pedro,ponta) - incorrect`

In this case, the sentence *O Pedro tocou com a **língua** no gelado da Ana* means approximately the same as *O Pedro tocou com a **ponta da língua** no gelado da Ana*, so the whole-part relations are different from the previous case.

5. In some cases, a whole-part relation is only implicit, and though *Nbp* are involved, they are not mentioned directly. For example, if someone has a *gastrite* 'gastritis' s/he has a disease in the stomach. In these cases, a whole-part relation between the human entity and the "hidden" *Nbp* should be established:

*O **Pedro** tem uma **gastrite*** 'Pedro has a gastritis'

`WHOLE-PART_HIDDEN(Pedro,estômago)`

6. Finally, there may be frozen sentences (or idioms) that involve *Nbp*. In these cases, a `FIXED` dependency is extracted:

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

`FIXED(perdeu,cabeça)`

If the `FIXED` dependency is extracted, there should not be a whole-part relation, as it can be considered to be irrelevant for the meaning of the sentence.

The goal of this work is to annotate whether a whole-part relation has been extracted **correctly**, or if it should be **removed**, **added** or **changed**:

- **correct** dependency - **do nothing**;
- **spurious** dependency (there should not be any whole-part dependency) - **remove** the dependency;
- **missing** dependency - **add** above the corresponding sentence the missing dependency:

  `WHOLE-PART(whole,part);`

- **partially correct** dependency - **change** the incorrect item in the dependency, either the whole or the part.

Thank you very much for your help.

Appendix E

Golden Standard

0>TOP{NP{As histórias} PP{de a poluição} PP{de o NOUN{rio Grande}} VF{correm} NP{toda aregião} , PP{desde o aparecimento} PP{de cadáveres} PP{de animais} PP{em a sua foz} PP{até a o boato} PP{de um surto} PP{de hepatite B} SC{que PP{em o NOUN{ano passado}} VF{afastou}} NP{centenas} PP{de veraneantes} .}

1>TOP{SC{Quando VF{alinhou}} PP{em o prólogo} PP{de Loulé} ADVP{já} VF{sabia} NP{o seu destino} : NP{o médico} VF{proibia} PP{a sua presença} PP{em a Volta} , PP{por suspeita} PP{de uma hepatite} .}

2>TOP{" NP{A prevenção} VCOP{é} AP{fundamental} , ADVP{principalmente} PP{junto de as mulheres} e PP{de os mais jovens} " , VF{diz} NP{NOUN{Jaime Branco}} , NP{reumatologista} e NP{NOUN{vice-presidente de a Sociedade}} NP{Portuguesa} PP{de as Doenças Ósseas} AP{Metabólicas} (NP{SPODOM}) , que , ADVP{em conjunto} PP{com a Associação Nacional} PP{contra a Osteoporose} (NP{APOROS}) , VF{promove} NP{a campanha} .}

3>TOP{NP{Os médicos} VMOD{têm , actualmente , NP{meios} de} VINF{diagnosticar} NP{as doenças} PP{de origem} AP{genética} AP{mais comuns} , como NP{a distrofia muscular} e NP{a mucoviscidose} , mas VCOP{é} AP{também possível} VINF{avaliar} NP{a predisposição} PP{de certos indivíduos} SC{para VINF{contrair}} NP{certos tipos} PP{de cancro} e VCOP{é} AP{provável} que , PP{em breve} , VCOP{seja} AP{possível} VINF{prever} NP{o risco} VINF{de sofrer} PP{de diabetes} , PP{de doenças cardiovasculares} ou PP{de artrite reumatóide} .}

WHOLE-PART(Abdel Rahman,pâncreas)
4>TOP{NP{NOUN{Abdel Rahman}} , NP{55 anos} , SC{que VCOP{é}} AP{cego} e VF{sofre} PP{de diabetes} , VF{sentia} NP{se} ADVP{" bastante bem "} , VF{disse} NP{Batchelder} , NP{o seu advogado} , AP{conhecido} SC{por VTEMP{ter}} VPP{representado} NP{alguns clientes} AP{ligados} PP{a o crime organizado} .}

5>TOP{NP{Os organizadores} VASP{estão a} VINF{estudar} , VF{desde há} NP{algum tempo} , NP{a prevalência} PP{de a osteoporose} PP{em a população} AP{portuguesa} , VGER{sabendo} NP{se} que ADVP{só} PP{em a NOUN{região Sul de o}} NP{país} NP{mais de 25 por cento} PP{de as mulheres} PP{entre os 20} e NP{os 80 anos} VF{sofrem} PP{de a doença} .}

6>TOP{PP{No caso de a rubéola} , NP{o risco} PP{de artrite} VF{depende} ADVP{muito} PP{de a idade} PP{de a pessoa} AP{vacinada} : PP{em as crianças} , NP{o risco} VCOP{é} AP{pequeno} , mas VCOP{é} AP{maior} PP{em os adultos} .}

7>TOP{NP{Os tratamentos} PP{de a sida} VCOP{são} AP{semelhantes} PP{a os aplicadas} PP{em doenças crónicas} como NP{a diabetes} e NP{a artrite} , VF{disse} NP{Merson} .}

8>TOP{PP{De acordo com as previsões} PP{de o relatório} PP{de a OMS} , NP{as pessoas} VF{com " ADVP{mais de} NP{65 anos} " VF{passarão}} , PP{em 25 anos} , PP{de 380} PP{para 690 milhões} NP{o} SC{que VF{provocará}} NP{um crescimento} PP{de as artroses} e PP{de a osteoporose} .}

9>TOP{ADVP{Mais recentemente} , VF{individualizaram} NP{se} NP{as consultas} PP{de gravidez} e NP{diabetes} e PP{de toxicodependentes} e VF{criaram} NP{se} VF{consultas} PP{de referência} (PP{em articulação}

PP{com centros de saúde}) , NP{a consulta} PP{de senologia} , a PP{de andrologia} , a PP{de ginecologia} AP{pediátrica} , a PP{de menopausa} , a PP{de diagnóstico pré-natal} e a PP{de aconselhamento} AP{genético} .}

WHOLE-PART_HIDDEN(filho,brônquio)
10>TOP{NP{O filho} , PP{com dez meses} e SC{que VCOP{estava}} VCPART{entregue} PP{a os cuidados} PP{de uma ama} , VF{sofre} PP{de bronquite} NP{asmática} .}

11>TOP{" NP{A má qualidade da água} AP{canalizada} " , VF{garante} , " VCOP{é} AP{responsável} PP{por o aparecimento} PP{de doenças infecciosas} como NP{a disenteria} , NP{a hepatite A} e NP{infecções} VF{intestinais} NP{agudas} " .}

12>TOP{NP{A diabetes} VCOP{é} NP{uma doença} VCPART{envolta} PP{em algum mistério} .}

13>TOP{NP{As mulheres} SC{que VF{adquirem}} NP{diabetes} PP{durante a gravidez} VMOD{podem} VINF{desenvolver} NP{hipertensão} e NP{diversos problemas neurológicos} , além de que NP{um bebé} SC{que VF{nasce}} PP{de uma gravidez} AP{complicada} PP{por diabetes} VMOD{pode} VCOP{ser} AP{muito grande} e VMOD{pode} VINF{causar} NP{um grande trauma} PP{durante o parto} .}

14>TOP{NP{O} SC{que NP{se} VF{pensa}} SC{que VF{acontece}} PP{em a artrite reumatóide} VF{é} SC{que NP{a cartilagem} VCOP{é}} VCPART{atacada} PP{por as defesas} AP{imunitárias} PP{de o doente} , SC{como se NP{ela} VF{fosse}} NP{um autêntico NOUN{" corpo estranho "}} .}

15>TOP{NP{Uma substância} AP{inédita} PP{contra a artrite reumatóide} VCOP{foi} VCPART{experimentada} PP{em os Estados Unidos} PP{em 28 pessoas} ADVP{gravemente} NP{doentes} e VF{surtiu} NP{efeitos} AP{muito encorajadores} , VF{anunciou} NP{a revista} NP{americana NOUN{" Science "}} PP{em a sua última edição} .}

16>TOP{NP{A artrite reumatóide} VF{é} NP{uma doença crónica} SC{que NP{se} VF{caracteriza}} PP{por inflamações} e NP{dores} PP{em as articulações} e VF{dá} NP{lugar} PP{a a erosão} PP{de a cartilagem} SC{que VF{cobre}} NP{as extremidades} PP{de os ossos} , assim como NP{a lesões} PP{em os próprios ossos} .}

17>TOP{NP{Os dados} PP{sobre Portugal} VF{são} ADVP{muito} NP{vagos} : apesar de , ADVP{até} NP{o MEDOS} VINF{arrancar} , NP{o Ministério da Saúde} VINF{ter} ADVP{apenas} NP{estatísticas} AP{gerais} PP{sem distinção} PP{de sexos} , VF{há} PP{em Portugal} PP{entre 500} PP{a 750 mil pessoas} SC{que VF{têm}} NP{esta doença} , VGER{sabendo} NP{se} SC{que VF{ocorrem}} PP{entre quatro} PP{a cinco mil fracturas} PP{de a anca} PP{por ano} , SC{que VF{afectam}} NP{três mulheres} PP{por cada homem} e SC{que VF{custam}} , PP{por doente} , NP{900 contos} ADVP{apenas} PP{em tratamento hospitalar} .}

18>TOP{NP{A sua reputação} como NP{afrodisíaco} , AP{contestada} PP{por a medicina} , VF{provêm} de SC{o facto de NP{as cantáridas} VF{serem}} NP{um agente} AP{irritante} que , quando AP{tomadas} ADVP{internamente} , VF{inflamam} NP{as mucosas} PP{de o aparelho urinário} , VGER{provocando} ADVP{eventualmente} NP{uma erecção} AP{involuntária} , AP{geralmente dolorosa} .}

19>TOP{VF{Entre} NP{estas} , VF{contam} NP{se} NP{algumas} PP{de as principais artérias} AP{portuenses} , como as PP{de NOUN{Oliveira Monteiro}} , NP{NOUN{João de as Regras}} , NP{NOUN{Gonçalo Cristóvão}} , NP{NOUN{Santos Pousada}} , NP{NOUN{Sá de a Bandeira}} e NP{NOUN{Mousinho de a Silveira}} , PP{em a época} AP{apresentadas} como NP{" largas e magníficas ruas "} , assim como NP{o NOUN{cemitério de Agramonte}} , PP{para além de um grandioso projecto} - ADVP{infelizmente} AP{não realizado} , PP{de um extenso parque} PP{entre a NOUN{Rotunda de a Boavista}} e NP{a NOUN{Quinta de a Prelada}} , VF{projecto} NP{esse} SC{que VF{constituiu}} NP{uma antevisão} PP{de o futuro NOUN{Parque de a Cidade}} .}

20>TOP{VCPART{Percorrida} PP{por edifícios} PP{de importante significado} AP{patrimonial} e NP{outros} PP{de fraco índice arquitectónico} , NP{aquela artéria} VTEMP{tem} VPP{vindo} NP{a ser} AP{alvo} PP{de um processo} PP{de renovação} AP{construtiva} SC{que VTEMP{tem}} VPP{levado} PP{a os AP{mais apaixonantes} comentários} e NP{exaltadas} NP{defesas} .}

21>TOP{NP{O automóvel} VASP{acaba por} VINF{ser} NP{a única alternativa viável} , NP{o} que , ADVP{no entanto} , NP{se} VF{apresenta} como NP{uma faca} PP{de dois gumes} , SC{pois VF{agrava}} NP{o trânsito}

PP{em as artérias} PP{de Monsanto} , SC{quando NP{o desejável} VF{seria}} NP{o contrário} , NP{se} VF{considerarmos} SC{que VCOP{estamos}} PP{no interior de um parque florestal} .}

22>TOP{NP{O estacionamento} AP{automóvel} PP{em as principais artérias} PP{de a NOUN{cidade de Leiria}} VTEMP{vai} VASP{deixar de} VCOP{ser} AP{gratuito} , VF{anunciou} NP{a autarquia} , SC{que VF{prepara}} NP{a instalação} PP{de parquímetros} PP{em as NOUN{avenidas Marquês de Pombal}} , NP{Heróis} PP{de Angola} e NP{Combatentes} PP{de a Grande} NP{Guerra} .}

23>TOP{NP{A polícia} VF{montou} ADVP{imediatamente} NP{barreiras} PP{em as principais artérias} PP{de a cidade} , mas NP{a única coisa} SC{que VF{conseguiu}} VTEMP{foi} VINF{encontrar} NP{a furgoneta} em SC{que VF{seguiam}} NP{alguns} PP{de os comandos} .}

24>TOP{NP{Redondos} , PP{de metal} NP{prateado} , NP{baço} , NP{NOUN{Alain Mikli}} , NP{preço} NP{acons} .}

WHOLE-PART(Onésimo,barbas)
25>TOP{NP{Barbas} AP{compridas} e AP{esbranquiçadas} , VINF{olhar} AP{penetrante} , NP{tez} AP{clara} e NP{NOUN{" papillon "}} , NP{Onésimo} VCOP{é} VCPART{tido} , PP{em São Vicente} , como NP{o grande senhor} PP{de o barlavento} NP{cabo-verdiano} (VINF{ver} NP{caixa}) .}

WHOLE-PART(judeu,barba)
26>TOP{NP{A outra} VF{mostra} NP{um judeu} , NP{ultra-ortodoxo} , AP{identificado} como NP{tal} PP{por a farta barba} e NP{a NOUN{" kippa "}} , NP{a mitra} .}

27>TOP{NP{O NOUN{encenador NOUN{Eugenio Barba}}} , NP{NOUN{director de o NOUN{Odin Teatret}}} , e NP{o cineasta} NP{NOUN{Bernardo Bertolucci}} VTEMP{têm} NP{encontro} VPP{marcado} PP{com o público} PP{de Lisboa} e PP{de o Alentejo} , PP{em o âmbito} PP{de a terceira edição} PP{de o Festival} NP{Sete Sóis} , NP{Sete Luas} que , NP{este ano} , VF{decorre} PP{PP{entre 1} e NP{NP{24} PP{de Setembro}}} .}

28>TOP{VF{Como} SC{quem VF{faz}} NP{a barba} ! " VF{Deixei} VINF{crescer} PP{a barba} , SC{porque NP{as pessoas} ADVP{assim} VF{respeitam}} NP{me} ADVP{mais} . " .}

29>TOP{NP{Segundo} NP{NOUN{Fernando Barriga}} , NP{o outro responsável} PP{de a missão} , VF{há} NP{duas hipóteses} AP{relativas} PP{a a evolução} PP{de o novo campo} .}

30>TOP{ADVP{Tanto} NP{tempo} , que NP{a carne} , NP{o leite} , NP{a água} , NP{os legumes} , VMOD{podem} VASP{deixar de} VINF{alimentar} NP{a barriga} PP{de a cidade} .}

31>TOP{ADVP{Não} VF{há} NP{nenhuma relação} PP{entre o consumo} PP{de café} e NP{o cancro da bexiga} .}

32>TOP{NP{Segunda-feira} , NP{NP{12} PP{de Agosto}} * NP{Portugal} VF{acerta} PP{em Nova Iorque} NP{pormenores} PP{de a visita parlamentar} PP{a Timor-Leste} * NP{Comissão Política} PP{de o PCP} , NP{Lisboa} * NP{NOUN{VI Congresso de a Frelimo}} , PP{em Maputo} * NP{Natação} : NP{campeonatos} PP{de os Estados Unidos} , PP{em Boca} NP{Raton} .}

WHOLE-PART(Diabo,Boca)
33>TOP{NP{A experiência} VF{repetiria} NP{se} PP{em 1989} , PP{com idêntico sucesso} , em " NP{Boca} PP{de o Diabo} " NP{(1989)} , NP{outro magnífico romance} PP{em banda desenhada} .}

WHOLE-PART(dirigentes,boca)
34>TOP{SC{Para NP{o} VINF{conseguir}} , NP{os dirigentes} PP{de o PSD} VF{ouviram} PP{de a boca} PP{de o líder} PP{de o partido} PP{a argumentação} AP{necessária} SC{para VF{convencerem}} NP{o eleitorado} PP{até Dezembro} .}

WHOLE-PART(portistas,boca)
35>TOP{PP{NP{A noite} PP{de ontem}} PP{em o Porto} VTEMP{vai} , ADVP{aliás} , VCOP{ficar} ADVP{seguramente} PP{para a história} como NP{a noite} PP{de os palavrões} , NP{tantos} VF{eram} NP{os} SC{que NP{se} VF{ouviam}} PP{de a boca} PP{de os portistas} .}

FIXED(abrirmos,boca)
36>TOP{SC{Quando VF{abrirmos}} NP{a boca} , NP{o queijo} VF{cai} e NP{a raposa} VF{leva} NP{o} " .}

FIXED(apanhados,com,boca,em,botija)
37>TOP{Mas NP{o certo} VF{é} SC{que NP{elas} VF{assustam}} ADVP{particularmente} NP{os pequenos delinquentes} , SC{que VF{correm}} ADVP{sempre} NP{o risco} de VCOP{serem} VCPART{apanhados} PP{com a boca} PP{em a botija} .}

WHOLE-PART(sua,boca)
38>TOP{Mas , PP{em a sua boca} , NP{a palavra} NP{democratização} VF{tem} NP{o sentido inverso} PP{a o invocado} PP{por Smith} .}

WHOLE-PART(lhes,boca)
39>TOP{VF{Foi} NP{o encenador} e NP{único actor} PP{de a NOUN{peça " A os Crocodilos mete se lhes um pau em a boca "}} , SC{que VCOP{esteve}} PP{em cena} PP{em o Teatro Nacional} , PP{PP{em Dezembro} PP{de 1990}} .}

40>TOP{VF{Fugiu} NP{lhe} NP{a boca} PP{para a verdade} .}

41>TOP{PP{À saída de o hotel} SC{onde VF{lançou}} NP{o manifesto} , NP{Durão} ADVP{ainda} VF{deixou} VINF{cair} NP{outra NOUN{" boca "}} PP{a Santana} NP{Lopes} : " NP{O Sporting} ADVP{ontem} VF{ganhou} NP{4-0} .}

WHOLE-PART(seres,boca)
42>TOP{NP{NOUN{Marjorie Wallace}} , SC{quando NP{as} VF{viu}} PP{por a primeira vez} PP{em o julgamento} , VF{escreveu} SC{que VF{eram}} NP{dois seres} " AP{pequenos} e AP{vulneráveis} , e ADVP{não} VF{abriam} PP{a boca} VINF{a ADVP{não} VINF{ser}} SC{para VINF{emitir}} NP{uns murmúrios} SC{que NP{o tribunal} VF{interpretou}} como NP{sinais} AP{evidentes} PP{de culpabilidade} " .}

43>TOP{SC{Depois de NP{ele} VTEMP{ter}} VPP{feito} ADVP{pouco} PP{de ela} : " VF{foi} PP{por causa de umas bocas} AP{chatas} " , VF{adiantou} NP{o rapaz} .}

44>TOP{" VF{Há} ADVP{aí} NP{uma culpazinha} , NP{o} SC{que NP{lhe} VF{fazia}} ? " , VF{interrogou} NP{NOUN{Alexandra Lencastre}} , NP{um caso} PP{de vocação} SC{para VINF{fazer}} NP{perguntas} .}

45>TOP{VF{São} ADVP{só} NP{bocas} PP{de a reacção} .}

WHOLE-PART(balão,boca)
46>TOP{PP{De essa acção} AP{inaugural} , ADVP{todavia} , NP{o lance} AP{mais célebre} VF{foi} NP{a ideia} VF{de corrigirem} NP{a atrocidade} SC{que VF{era}} NP{a publicidade} PP{a uma câmara de vídeo} PP{com a famosa foto} PP{de uma criança} AP{vietnamita} VINF{a arder} PP{em napalm} - VF{pintaram} NP{lhe} NP{um balão} VINF{a sair} PP{de a boca} VGER{declarando} " NP{NOUN{Feliz Natal}} PP{da parte de os JAMS} " .}

47>TOP{NP{O príncipe} NP{marinheiro} e NP{a antiga directora editorial} NP{NOUN{Sarah Ferguson}} VF{tinham} SC{dado que VINF{falar}} PP{em os últimos meses} , PP{com as suas aparições públicas} e NP{gestos} PP{de afecto} , SC{que VF{incluíram}} NP{um beijo} PP{de despedida} PP{em a boca} .}

48>TOP{NP{O Braga} VF{marcou} ADVP{ADVP{logo} PP{a os 25 segundos}} , PP{em a primeira iniciativa} AP{atacante} , PP{com Andersson} , PP{à boca de a baliza} , VINF{a empurrar} NP{uma bola} AP{perdida} PP{por NOUN{Peter Rufai}} , SC{que VF{defendera}} ADVP{incompletamente} NP{um remate} AP{espectacular} PP{de Barroso} .}

49>TOP{VF{Parecia} NP{um sapato} PP{de banda desenhada} , mas VF{tinha} NP{dois olhinhos} AP{muito engraçados} e NP{boca} PP{de ratinho} .}

FIXED(abrir,boca)
50>TOP{Mas NP{a viúva} PP{de Rajiv} VASP{continua a} VINF{receber} NP{dissidentes} e NP{apoiantes} PP{de Rao} , SC{sem VINF{abrir}} NP{a boca} ou VINF{inclinar} NP{se} PP{para qualquer} PP{de os lados} .}

51>TOP{NP{Eu} VCOP{sou} PP{de boa boca} .}

52>TOP{VF{Faz} NP{uma pequeníssima pausa} e , PP{em um trejeito} PP{de boca} SC{que NP{lhe} VCOP{é}} AP{muito característico} , VF{diz} como SC{que VGER{procurando}} NP{cada uma} PP{de as palavras} .}

WHOLE-PART(santos,bocas)
53>TOP{PP{Em dois templos} VCOP{foram} VCPART{destruídos} NP{sacrários} e NP{as hóstias} AP{colocadas} PP{em as bocas} PP{de as imagens} PP{de santos} , mas NP{as caixas de esmolas} ADVP{não} VCOP{foram} VCPART{assaltadas} .}

54>TOP{PP{Em os loucos NOUN{anos 30}} , NP{Xangai} VF{abriu} NP{os braços} PP{a uma sociedade} AP{sedenta} PP{de aventuras} , AP{interrompidas} PP{durante a Segunda Grande Guerra} e AP{decepadas} PP{por a Revolução Cultural} .}

55>TOP{Apesar de NP{isso} , VASP{continua a} VINF{faltar} NP{dinheiro} PP{para projectos de investigação} e , PP{com a Rússia} PP{a braços} PP{com uma crise económica} PP{sem paralelo} , NP{a situação} VMOD{tende a} VINF{agravar} NP{se} ADVP{de dia para dia} .}

56>TOP{E PP{em a Assembleia} , NP{o PCP} VF{é} NP{o braço} NP{amigo} SC{que VTEMP{vai}} VINF{votar} NP{as propostas legislativas} SC{que NP{NOUN{Alberto Costa}} VF{prepara}} .}

57>TOP{PP{Por a primeira vez} PP{desde o início} PP{de a guerra comercial} , AP{iniciada} PP{em a passada terça-feira} , PP{a Petrofel} VF{baixou} NP{os braços} e ADVP{não} VF{respondeu} PP{a a líder} PP{de o mercado} , VGER{mantendo} NP{o desconto} PP{em 10 escudos} .}

58>TOP{VF{Há} NP{algum tempo} , NP{a Faculdade de Letras} VF{viu} NP{se} PP{a braços} PP{com uma queixa} PP{de um editor} NP{inglês} .}

59>TOP{NP{O braço de ferro} VF{mantém} NP{se} PP{desde o último Verão} SC{quando NP{o NOUN{padre NOUN{Benjamin Videira Pires}}} VF{recusou}} VCOP{ser} VCPART{substituído} PP{em o cargo} PP{de director} PP{por NOUN{Joseph Tai}} , NP{um jesuíta} PP{de Hong Kong} .}

60>TOP{NP{Ele} VF{substituiria} , ADVP{assim} , NP{NOUN{Ribeiro de a Costa}} , NP{outro braço-direito} PP{de o líder} PP{de os centristas} , SC{que VTEMP{tem}} VPP{assegurado} NP{a secretaria-geral} PP{de o partido} PP{em os últimos anos} e a SC{quem NP{este} ADVP{agora} VF{destina}} NP{outros voos} .}

61>TOP{Quer NP{os Daimler} quer NP{os congéneres} - NP{AEC} , NP{Leyland} , ADVP{etc.} - ADVP{não} VCOP{estavam} VCPART{equipados} PP{com direcção assistida} , NP{essa maravilha} AP{técnica} SC{que VF{permite}} PP{a os braços} AP{menos musculados} VINF{enfrentar} NP{as manobras} AP{mais exigentes} PP{de o ponto de vista} AP{físico} .}

62>TOP{P. - Mas NP{o sucesso} , ADVP{aqui} , VF{depende} ADVP{essencialmente} PP{de o NOUN{" visado "}} VINF{dar} NP{o braço} VINF{a torcer} , PP{de a disponibilidade} PP{de a Administração Pública} SC{para VINF{aceitar}} NP{a recomendação} PP{de a Provedoria} .}

63>TOP{NP{O Sinn Fein} , NP{o braço} AP{político} PP{de o IRA} , VF{suspeita} SC{que NP{o incidente} VF{decorreu}} PP{de uma} " NP{operação} NP{encoberta} " AP{executada} PP{por uma unidade} PP{de o comando} PP{de elite} PP{de o exército} AP{britânico} , NP{o SAS} , PP{em um momento} em SC{que NP{o NOUN{Governo de Londres}} VCOP{está}} ADVP{sob pressão} PP{de NOUN{líderes unionistas}} SC{para VINF{intensificar}} NP{o cerco} PP{contra os elementos} NP{suspeitos} PP{de a prática} PP{de terrorismo} PP{em a província} .}

64>TOP{SC{Quando VF{há}} NP{carros} AP{mais rápidos} NP{NUM{meio segundo}} , ou NP{um segundo} , VCOP{é} AP{difícil} VINF{tirar} NP{a diferença} PP{em o braço} .}

65>TOP{NP{Tudo isto} VCOP{é} AP{estranho} e VF{traduz} NP{um} VINF{agudizar} PP{de as tensões} PP{entre Belém} e NP{S. Bento} , PP{em um braço-de-ferro} SC{que VF{ameaça}} VINF{toldar} ADVP{ainda mais} NP{o já agitado} e NP{nebuloso clima político} .}

66>TOP{PP{A experiência} - NP{a primeira} AP{realizada} PP{em o Alentejo} SC{para VINF{contratar}} NP{médicos} AP{estrangeiros} - VMOD{parece} VINF{ter} NP{resultado} e VF{espera} NP{se} ADVP{agora} SC{que NP{novos concursos} VF{tenham}} NP{lugar} PP{para outros concelhos} PP{de o interior} AP{alentejano} , PP{a braços} PP{com a falta} PP{de médicos} .}

WHOLE-PART(mães,braços)
67>TOP{" NP{Crianças} AP{queimadas} AP{vivas} PP{em os braços} PP{de as mães} , SC{que VF{gritavam}} : ` NP{Jesus} , VF{recebe} NP{as nossas almas} !}

68>TOP{' " .}

69>TOP{E ADVP{não} VF{há} NP{grandes novidades} : NP{o ensino secundário} VF{sofre} PP{de uma considerável crise de identidade} , VCOP{vive} PP{a braços} PP{com o excesso} PP{de horas lectivas} , VF{padece} PP{de programas} AP{extensos} e NP{as provas globais} VF{precisam} VINF{de ser} ADVP{" legalmente "} NP{redifinidas} .}

70>TOP{VF{Ora} , NP{a paixão} PP{de as presidenciais} PP{de 1996} VF{será} , ADVP{antes de mais} , NP{o desenlace} PP{de este clímax} , NP{altura} como NP{nenhuma} NP{outra} AP{privilegiada} SC{para que NP{os dois grandes actores} NP{políticos nacionais} VF{meçam}} NP{forças} , PP{em um braço-de-ferro} AP{final} SC{que ADVP{não} NP{se} VF{avizinha}} AP{fácil} .}

71>TOP{NP{A ETA} e NP{o NOUN{Harri Batassuna}} , NP{partido} AP{normalmente apontado} como NP{braço} AP{político} PP{de a organização terrorista} , VF{encarregam} NP{se} PP{de isso} .}

72>TOP{E ADVP{talvez também} NP{uma} ou NP{outra dor} PP{de coluna} , PP{depois de a queda} SC{que VF{deu}} PP{de o telhado} PP{de a fábrica} , ou NP{as dores} PP{de os braços} , NP{fruto} PP{de os anos} em SC{que VF{carregou}} PP{com as pastas} PP{de as amostras} PP{em o estrangeiro} .}

73>TOP{NP{A modernização} AP{técnica} PP{de estas empresas} VF{faz} com SC{que NP{elas} VMOD{possam}} VINF{tornar} NP{se} AP{mais fortes} do que NP{certos estados} NP{fracos} PP{a braços} PP{com grandes problemas} " .}

WHOLE-PART(inimigo,braço)
74>TOP{VF{Acertou} PP{em o braço} PP{de o inimigo} .}

WHOLE-PART(pais,braços)
WHOLE-PART(mães,braços)
75>TOP{NP{Centenas de famílias} PP{de o bairro} NP{NOUN{residencial de Mycrorayan}} , AP{habitado} PP{por quadros} PP{de o antigo regime} , VF{aproveitaram} NP{a trégua} PP{de o meio-dia} SC{para VF{fugirem}} - NP{NOUN{pais e mães}} PP{com bebés} PP{em os braços} , NP{alguns} VGER{transportando} NP{apenas um pequeno saco de plástico} , PP{sem tempo} PP{para mais} .}

WHOLE-PART(operário,braços) 76>TOP{NP{O operário} , SC{que VF{trabalhava}} PP{com um poderoso pilão} , VF{caiu} VF{sobre o VF{malho}} , VGER{perdendo} ADVP{logo} NP{o braço} SC{para ADVP{depois} VINF{tombar}} PP{para o lado} , AP{já morto} .}

WHOLE-PART(comandante,braço) 77>TOP{VF{Confirmou} ADVP{assim} NP{a versão} PP{de o antigo comandante} PP{de o NOUN{posto de a GNR de Sacavém}} que , quando PP{de o início} PP{de o julgamento} , VF{explicou} PP{a o colectivo} NP{o movimento} SC{que VF{fez}} PP{com o braço} - - PP{em o sentido ascendente} - - e SC{que VF{provocou}} NP{o disparo} (VF{dito} AP{acidental}) .}

78>TOP{VF{Faziam} NP{se} AP{prisioneiras} , AP{rapidamente levadas} ADVP{em braços} PP{para as camionetas} NP{inimigas} .}

79>TOP{PP{De este braço de ferro} PP{entre os símbolos} PP{de o bem} e PP{de o mal} VF{ressalta} PP{a forte personalidade} PP{de o juíz} NP{Dredd} , NP{um homem} ADVP{incorruptível} e NP{justiceiro} , AP{armado} PP{com uma pistola} SC{que ADVP{apenas} NP{lhe} VF{obedece}} PP{a ele} e SC{que VF{desenvolve}} NP{uma}

ADVP{incansável} NP{luta} PP{contra todos os fora-da-lei} .}

80>TOP{NP{As formações} AP{germânicas} ADVP{nunca} VF{baixam} NP{os braços} .}

81>TOP{VF{Dizem} SC{que ADVP{não} VTEMP{vão}} VCOP{ficar} PP{de braços} AP{cruzados} e VASP{estão a} VINF{promover} NP{diligências} PP{junto de o Presidente da República} e PP{de o provedor de Justiça} .}

82>TOP{VGER{Pondo} NP{termo} PP{a um NOUN{" braço-de-ferro "}} SC{que NP{se} VF{arrasta}} ADVP{ADV{há cinco anos}} , NP{os advogados} PP{de os ex-futebolistas} e NP{os assessores jurídicos} PP{de a CBF} VF{prevêem} NP{um pagamento} PP{em a ordem} PP{de os dois milhões de reais} (NP{cerca de 315 mil contos}) .}

83>TOP{ADVP{Até} por que NP{a multidão} PP{de associações} SC{que VF{constam}} PP{de o CEC} , VGER{indo} PP{desde a agricultura} PP{a o pequeno comércio} , VGER{passando} PP{por a indústria} , AP{provenientes} PP{de zonas} PP{de rápido crescimento} como NP{Aveiro} ou NP{Leiria} ou PP{a braços} PP{com problemas} AP{graves} PP{de desertificação} como PP{a Guarda} e NP{Castelo Branco} , VF{impede} , PP{em boa parte} , NP{esta tarefa} .}

84>TOP{NP{A figura tutelar} PP{de NOUN{Peter Norton}} VASP{continuou a} VINF{aparecer} VCPART{relacionada} PP{com todos os produtos} AP{entretanto surgidos} , ADVP{mais} SC{que ADVP{não} VCOP{seja}} PP{em as páginas} PP{de publicidade} : ADVP{já} VF{é} NP{tradição} VINF{ver} NP{a imagem} PP{de Norton} , ADVP{em camisa} e PP{de braços} AP{cruzados} , PP{em todos os anúncios} PP{de a Symantec} SC{que NP{se} VF{relacionam}} PP{com a NOUN{Peter Norton Computing}} .}

85>TOP{NP{A situação} AP{actual} VF{caracteriza} NP{se} PP{por um braço-de-ferro} , AP{disfarçado} PP{em a linguagem} AP{aveludada} PP{de a diplomacia} .}

WHOLE-PART(lhe,braço) 86>TOP{PP{De acordo com os dados} AP{fornecidos} PP{por a PSP} , NP{os assaltantes} VF{rasgaram} NP{o bolso} PP{de a camisa} PP{de o funcionário} e , como NP{este tivese} AP{reagido} , VF{deram} NP{lhe} NP{duas navalhadas} PP{em o braço} AP{esquerdo} .}

87>TOP{VF{Entre} NP{outras coisas} , VF{fala} PP{de o amor} e PP{de a tatuagem} SC{que VF{tem}} PP{em o braço direito} : NP{NOUN{" NOUN{Winona Forever} "}} .}

88>TOP{NP{Os budistas} e NP{adeptos} PP{de o NOUN{" candomblé "}} VF{indicaram} SC{que VF{receberão}} NP{NOUN{João Paulo II}} PP{de braços} AP{abertos} .}

89>TOP{PP{Em os céus} PP{de a cidade} VF{erguem} NP{se} NP{tentáculos} AP{monstruosos} , NP{cabeças} PP{de dragões} , NP{torres} PP{de cidadelas} AP{antigas} , enquanto PP{sob elas} NP{um cavaleiro} AP{motorizado} , NP{uma princesa} NP{provocadora} e NP{um rei} AP{louco} VF{vivem} NP{sonhos} PP{ao som de o NOUN{" rap "}} e PP{de os NOUN{" flashes "}} PP{de o fogo de artifício} .}

WHOLE-PART(Donaciano,cabeça)
90>TOP{NP{Sorriso} AP{tímido} , NP{NOUN{Donaciano Gomes}} VF{sobe} NP{as escadas} PP{de acesso} PP{a o avião} , NP{ténis} , NP{gangas} e NP{mochila} , NP{um panamá} AP{preto} PP{de basebal} AP{enterrado} PP{em a cabeça} , NP{a foto} PP{de uma jovem} VF{VERB{" bibere "}} PP{a a lapela} .}

91>TOP{NP{Kamwango} , NP{uma aldeia} AP{perdida} PP{em a floresta} AP{africana} PP{de o Quénia} , VF{é} ADVP{também} NP{um nome} SC{que ADVP{não} VF{sai}} PP{de a cabeça} PP{de milhares} PP{de garimpeiros} SC{que ADVP{ADV{há cinco meses}} a VF{colocaram}} PP{em a geografia} PP{de o ouro} por VTEMP{terem} VCOP{sido} VCPART{descobertos} NP{novos filões} PP{de este precioso metal} .}

92>TOP{NP{O PÚBLICO} VF{tentou} , PP{sem sucesso} , VINF{ouvir} NP{as razões} PP{de o cabeça de lista} PP{de a candidatura} AP{contestada} .}

93>TOP{NP{Um espectador} AP{esclarecido} VMOD{deveria} VINF{ver} NP{este Strindberg} SC{antes de VINF{ver}} NP{o Genet} SC{que NP{a ordem alfabética} VF{colocou}} PP{à cabeça de esta lista} .}

94>TOP{PP{Em o NOUN{caminho de Pierce}} VF{está} ADVP{agora} NP{a bielorrussa NOUN{Natalia Zvereva}} ,

NP{cabeça-de-série} NP{NOUN{nº 8}} , SC{que VF{precisou}} PP{de três NOUN{" sets "}} SC{para VINF{eliminar}} NP{a japonesa NOUN{Kyoko Nagatsuka}} , NP{72ª} PP{de o NOUN{" ranking "}} .}

95>TOP{NP{Sofia} VF{defronta} NP{NOUN{Annouschka Poppe}} , NP{19 anos} e NP{NOUN{nº 193}} NP{WTA} , SC{que VF{eliminou}} ADVP{anteontem} NP{a oitava} NP{cabeça de série} , NP{a finladesa} NP{NOUN{Petra Thoren}} .}

96>TOP{VF{Discorda} ADVP{totalmente} PP{de os poderes} AP{agora dados} PP{a a Polícia Judiciária} , que , SC{se PP{de o ponto de vista} AP{funcional} VF{depende}} PP{de o NOUN{Ministério Público}} , " PP{de o ponto de vista} AP{orgânico} e AP{administrativo} VF{depende} PP{de a administração} e PP{de a cabeça} PP{de a administração} SC{que VF{é}} NP{o governo} " .}

WHOLE-PART(empresário,cabeça)
97>TOP{NP{O empresário} VCOP{foi} ADVP{gravemente} VCPART{atingido} PP{em a cabeça} e PP{em este momento} VF{encontra} NP{se} AP{ainda muito perturbado} PP{a nível} AP{psicológico} .}

98>TOP{NP{Puxei-o} NP{eu} ADVP{mesmo} PP{de a cabeça} .}

WHOLE-PART(Jorge Soares,cabeça)
WHOLE-PART(Gamarra,cabeça)
99>TOP{ADVP{Ainda} PP{em o mesmo jogo} , NP{destaque} PP{para o golo} PP{de NOUN{João Pinto}} , NP{outro tiro} PP{de fora de a área} , e NP{o primeiro} PP{de NOUN{Paulo Nunes}} , AP{acrobático} , PP{depois de dois toques} PP{de cabeça} PP{de NOUN{Jorge Soares}} e NP{Gamarra} .}