Thiago Alexandre Salgueiro Pardo
António Branco Aldebaro Klautau
Renata Vieira Vera Lúcia Strube de Lima (Eds.)

LNAI 6001

Computational Processing of the Portuguese Language

9th International Conference, PROPOR 2010
Porto Alegre, RS, Brazil, April 2010
Proceedings

Lecture Notes in Artificial Intelligence 6001

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Thiago Alexandre Salgueiro Pardo
António Branco Aldebaro Klautau
Renata Vieira Vera Lúcia Strube de Lima (Eds.)

Computational Processing of the Portuguese Language

9th International Conference, PROPOR 2010
Porto Alegre, RS, Brazil, April 27-30, 2010
Proceedings

Springer

Volume Editors

Thiago Alexandre Salgueiro Pardo
Universidade de São Paulo, São Carlos / SP, Brazil
E-mail: taspardo@icmc.usp.br

António Branco
Universidade de Lisboa, Portugal
E-mail: antonio.branco@di.fc.ul.pt

Aldebaro Klautau
Universidade Federal do Pará, Belém / PA, Brazil
E-mail: aldebaro@ufpa.br

Renata Vieira
PUCRS, Porto Alegre / RS, Brazil
E-mail: renata.vieira@pucrs.br

Vera Lúcia Strube de Lima
PUCRS, Porto Alegre / RS, Brazil
E-mail: vera.strube@pucrs.br

Library of Congress Control Number: 2010923533

CR Subject Classification (1998): I.2, H.3, H.4, I.4, I.5, H.2.8

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-12319-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-12319-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

The International Conference on Computational Processing of Portuguese—PROPOR—is the main event in the area of natural language processing that is focused on Portuguese and the theoretical and technological issues related to this language. It welcomes contributions for both written and spoken language processing.

The event is hosted in Brazil and in Portugal. The meetings have been held in Lisbon/Portugal (1993), Curitiba/Brazil (1996), Porto Alegre/Brazil (1998), Évora/Portugal (1999), Atibaia/Brazil (2000), Faro/Portugal (2003), Itatiaia/Brazil (2006) and Aveiro/Portugal (2008).

This meeting has been a highly productive forum for the progress of this area and to foster the cooperation among the researchers working on the automated processing of the Portuguese language. PROPOR brings together research groups, promoting the development of methodologies, resources and projects that can be shared among all researchers and practitioners in the field.

The ninth edition of this event was held in Porto Alegre, Brazil, at *Pontifícia Universidade Católica do Rio Grande do Sul* (PUCRS). It had two main tracks: one for language processing and another one for speech processing. This event hosted a special Demonstration Session and the first edition of the PhD and MSc Dissertation Contest, which aimed at recognizing the best academic work on processing of the Portuguese language in the last few years. This edition of the event featured tutorials on statistical machine translation and on speech recognition, as well as invited talks by renowned researchers of natural language processing.

A total of 48 submissions were received, 37 for the language track and 11 for the speech track, by authors from 10 countries: Brazil, China, Denmark, UK, Germany, Italy, Poland, Portugal, Spain and USA. Each submission was evaluated by at least three members from a multidisciplinary and international scientific committee.

This volume gathers a selection of the 21 best papers accepted to be presented at this meeting, of which 13 are full papers, corresponding to an acceptance rate of 27%. These papers cover the areas concerning applications for information handling and text processing, language processing, language resources, and speech recognition and synthesis.

We would like to express our thanks to everyone involved in the organization of the event, to the scientific committee members for their excellent work, to the researchers who kindly accepted to contribute to the event by delivering tutorials and invited talks, and to the institutions, organizations and funding agencies which allowed the realization of this event, namely, PUCRS, SBC (the Brazilian Computer Society), CEPLN (the SBC Special Interest Group on Natural Language Processing), CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico* – a Brazilian funding agency), NAACL (The North American Chapter of the Association for Computational Linguistics), ISCA (International Speech Communication Association), SIG-IL (the ISCA Special Interest Group on Iberian Languages) and CLARIN

(Common Language Resources and Technology Infrastructure – a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable).

April 2010

Thiago Alexandre Salgueiro Pardo
António Branco
Aldebaro Klautau
Renata Vieira
Vera Lúcia Strube de Lima

Organization

General Chair

Vera Lúcia Strube de Lima	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil

Program Chairs

António Branco	Universidade de Lisboa, Portugal - Language
Aldebaro Klautau	Universidade Federal do Pará, Brazil - Speech

Tutorial Chair

Maria das Graças Volpe Nunes	Universidade de São Paulo, Brazil

Editorial Chair

Thiago Alexandre Salgueiro Pardo	Universidade de São Paulo, Brazil

PhD and MSc Dissertation Contest Chair

David de Matos	Instituto de Engenharia de Sistemas e Computadores, Portugal

Demo Session Chair

António Teixeira	Universidade de Aveiro, Portugal

Local Organizing Chair

Renata Vieira	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil

Program Committee

Alexandre Agustini	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Aline Villavicencio	Universidade Federal do Rio Grande do Sul, Brazil
Amália Mendes	Universidade de Lisboa, Portugal
Andre Adami	Universidade de Caxias do Sul, Brazil
António Ribeiro	European Commission, Joint Research Centre, Italy
António Serralheiro	Instituto de Engenharia de Sistemas e Computadores / Instituto Superior Técnico, Portugal
António Teixeira	Universidade de Aveiro, Portugal
Antonio Bonafonte	Universitat Politècnia de Catalunya, Spain
Augusto Silva	Universidade Católica Portuguesa, Portugal
Bento Dias da Silva	Universidade Estadual Paulista, Brazil
Berthold Crysmann	Universität Bonn / Universität des Saarlandes, Germany
Carlos Prolo	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Caroline Hagège	Xerox Research Centre, France
Celso Kaestner	Universidade Tecnológica Federal do Paraná, Brazil
Climent Nadeu	Universitat Politècnia de Catalunya, Spain
Cristina Martins	Universidade de Coimbra, Portugal
Daniela Braga	Microsoft Language Development Center, Portugal
Dante Barone	Universidade Federal do Rio Grande do Sul, Brazil
David de Matos	Instituto de Engenharia de Sistemas e Computadores, Portugal
Diamantino Freitas	Universidade do Porto, Portugal
Edmilson Morais	Vocalize, Brazil
Eric Laporte	Université Paris-Est Marne-la-Vallée, France
Fábio Violaro	Universidade Estadual de Campinas, Brazil
Fernando Resende	Universidade Federal do Rio de Janeiro, Brazil
Florence Amardeilh	Mondeca, France
Gabriel Pereira Lopes	Universidade Nova de Lisboa, Portugal
Gael Harry Dias	Universidade da Beira Interior, Portugal
Horacio Saggion	University of Sheffield, UK
Irene Rodrigues	Universidade de Évora, Portugal
Isabel Trancoso	Instituto de Engenharia de Sistemas e Computadores / Instituto Superior Técnico, Portugal
Ivandré Paraboni	Universidade de São Paulo, Brazil
Ivandro Sanches	Centro Universitário da FEI, Brazil
Jason Baldridge	University of Texas, USA
Jean-Luc Minel	Université Paris Ouest - Nanterre La Défense, France
João Balsa	Universidade de Lisboa, Portugal
João Paulo Neto	Instituto de Engenharia de Sistemas e Computadores, Portugal
Jorge Baptista	Universidade do Algarve, Portugal
Julia Hirschberg	Columbia University, USA

Lúcia Rino	Universidade Federal de São Carlos, Brazil
Luísa Coheur	Instituto de Engenharia de Sistemas e Computadores, Portugal
Luiz Pizzato	University of Sydney, Australia
Marcelo Finger	Universidade de São Paulo, Brazil
Marco Gonzalez	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Maria Viana	Universidade de Lisboa, Portugal
Maria das Graças Volpe Nunes	Universidade de São Paulo, Brazil
Maria Francisca Xavier	Universidade Nova de Lisboa, Portugal
Maximiliano Saiz Noeda	Universidad de Alicante, Spain
Miguel Filgueiras	Universidade do Porto, Portugal
Nestor Yoma	Universidad de Chile, Chile
Nuno Mamede	Instituto de Engenharia de Sistemas e Computadores / Instituto Superior Técnico, Portugal
Nuno Cavalheiro Marques	Universidade Nova de Lisboa, Portugal
Pablo Gamallo	Universidad de Santiago de Compostela, Spain
Paulo Quaresma	Universidade de Évora, Portugal
Plínio Barbosa	Universidade Estadual de Campinas, Brazil
Ranniery Maia	Toshiba Research Europe Limited, UK
Renata Vieira	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Sandra Aluisio	Universidade de São Paulo, Brazil
Stanley Loh	Universidade Católica de Pelotas, Brazil
Steven Bird	University of Melbourne, Australia
Thiago Alexandre Salgueiro Pardo	Universidade de São Paulo, Brazil
Tracy King	Microsoft, USA
Violeta Quental	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Viviane Orengo	Universidade Federal do Rio Grande do Sul, Brazil

Additional Reviewers

Aldebaro Klautau	Universidade Federal do Pará, Brazil
Belinda Maia	Universidade do Porto, Portugal
Fábio Kepler	Universidade de São Paulo, Brazil
Guillaume Jacquet	Xerox Research Centre, France
Jussara Vieira	Vocalize, Brazil
Marta Costa-jussà	Barcelona Media, Spain
Milagros Fernández Gavilanes	Universidade da Coruña, Spain
Simon Zwarts	Macquarie University, Australia

Steering Committee

Maria das Graças Volpe Nunes	Universidade de São Paulo, Brazil
Vera Lucia Strube de Lima	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Isabel Trancoso	Instituto de Engenharia de Sistemas e Computadores / Instituto Superior Técnico, Portugal
Violeta Quental	Pontifícia Universidade Católica do Rio de Janeiro, Brazil
António Teixeira	Universidade de Aveiro, Portugal

Table of Contents

Applications: Information Handling

Improving IdSay: A Characterization of Strengths and Weaknesses in Question Answering Systems for Portuguese 1
Gracinda Carvalho, David Martins de Matos, and Vitor Rocio

Assessing the Impact of Stemming Accuracy on Information Retrieval 11
Felipe N. Flores, Viviane P. Moreira, and Carlos A. Heuser

Exploiting Multilingual Grammars and Machine Learning Techniques to Build an Event Extraction System for Portuguese 21
Vanni Zavarella, Hristo Tanev, Jens Linge, Jakub Piskorski, Martin Atkinson, and Ralf Steinberger

Formalizing CST-Based Content Selection Operations 25
Maria Lucía Castro Jorge and Thiago Alexandre Salgueiro Pardo

Applications: Text Processing

Translating from Complex to Simplified Sentences 30
Lucia Specia

Challenging Choices for Text Simplification 40
Caroline Gasperin, Erick Maziero, and Sandra M. Aluísio

Comparing Sentence-Level Features for Authorship Analysis in Portuguese 51
Rui Sousa-Silva, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia

Language Processing

A Machine Learning Approach to Portuguese Clause Identification 55
Eraldo R. Fernandes, Cícero N. dos Santos, and Ruy L. Milidiú

A Hybrid Approach for Multiword Expression Identification 65
Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto

Out-of-the-Box Robust Parsing of Portuguese 75
João Silva, António Branco, Sérgio Castro, and Ruben Reis

LXGram: A Deep Linguistic Processing Grammar for Portuguese 86
Francisco Costa and António Branco

Language Resources

InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language 90
Vladia Pinheiro, Tarcisio Pequeno, Vasco Furtado, and Wellington Franco

Comparing Verb Synonym Resources for Portuguese 100
Jorge Teixeira, Luís Sarmento, and Eugénio Oliveira

Auxiliary Verbs and Verbal Chains in European Portuguese 110
Jorge Baptista, Nuno Mamede, and Fernando Gomes

P-AWL: Academic Word List for Portuguese 120
Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral, and Nuno Mamede

Speech Recognition

Automatic Phone Clustering Based on Confusion Matrices 124
Carla Lopes, Arlindo Veiga, and Fernando Perdigão

An Open-Source Speech Recognizer for Brazilian Portuguese with a Windows Programming Interface 128
Patrick Silva, Pedro Batista, Nelson Neto, and Aldebaro Klautau

A Baseline System for Continuous Speech Recognition of Brazilian Portuguese Using the West Point Brazilian Portuguese Speech Corpus 132
Fabiano Weimar dos Santos, Dante Augusto Couto Barone, and André Gustavo Adami

Speech Synthesis

Voice Quality of European Portuguese Emotional Speech 142
Ana Nunes, Rosa Lídia Coimbra, and António Teixeira

Prosodic Prediction in Brazilian Portuguese: A Contribution to Speech Synthesis 152
Cirineu Cecote Stein

The Role of Morphology in Generating High-Quality Pronunciation Lexica for Regional Variants of Portuguese 162
Simone Ashby and José Pedro Ferreira

Author Index 167

Improving IdSay: A Characterization of Strengths and Weaknesses in Question Answering Systems for Portuguese

Gracinda Carvalho[1,2,3], David Martins de Matos[2,4], and Vitor Rocio[1,3]

[1] Universidade Aberta, Rua da Escola Politécnica, 147 1269-001 Lisboa, Portugal
gracindac@univ-ab.pt, vjr@univ-ab.pt
[2] L2F/INESC-ID Lisboa, Rua Alves Redol 9 1000-029 Lisboa, Portugal
david.matos@inesc-id.pt
[3] CITI - FCT/UNL
[4] Instituto Superior Técnico/UTL

Abstract. IdSay is a Question Answering system for Portuguese that participated at QA@CLEF 2008 with a baseline version (IdSayBL). Despite the encouraging results, there was still much room for improvement. The participation of six systems in the Portuguese task, with very good results either individually or in an hypothetical combination run, provided a valuable source of information. We made an analysis of all the answers submitted by all systems to identify their strengths and weaknesses. We used the conclusions of that analysis to guide our improvements, keeping in mind the two key characteristics we want for the system: efficiency in terms of response time and robustness to treat different types of data. As a result, an improved version of IdSay was developed, including as the most important enhancement the introduction of semantic information. We obtained significantly better results, from an accuracy in the first answer of 32.5% in IdSayBL to 50.5% in IdSay, without degradation of response time.

1 Introduction

Question Answering (QA) systems have the aim of providing a short and precise answer to an information request stated in Natural Language (NL). The knowledge base usually consists of a large amount of NL text which is searched using Information Retrieval (IR) techniques. The usefulness of these systems is to find the exact information in large volumes of text data.

In recent years there has been very active research in this field, that led to the creation of tracks dedicated to QA in several international evaluation initiatives that, although presenting several variations, generally focus on the task of extracting answers from large open domain text collections. Such is the case of the Text REtrieval Conference (TREC) that had a QA track from 1999 to 2007 and the Text Analysis Conference (TAC) in 2008, for the English language. Another case is the bi-annual NTCIR Workshop that has been running a QA track, Question Answering Challenge (QAC), since its third edition in 2002, and

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 1–10, 2010.

whose main focus is on Asian languages. The Cross Language Evaluation Forum CLEF, an initiative co-sponsored by the European Commission is running a QA track, QA@CLEF, since 2003, including the Portuguese language since 2004. Although this forum concentrates its attention on Cross Language issues, it also has available mono-lingual tasks.

1.1 QA Systems for Portuguese at QA@CLEF 2008

IdSay [1], presented at QA@CLEF 2008 [2], is a QA system for Portuguese. This was a baseline version (henceforth referred to as IdSayBL). It answered correctly to 65 out of 200 questions in the first answer, and to 85 answers considering the three answers that could be returned per question[1]. Despite having had very encouraging results, there was, however, room for improvement.

QA systems for Portuguese at QA@CLEF are worth careful analysis because they produced state of the art results [2], with the system from the Portuguese company Priberam [3] being the best system overall in the 2005 and 2008 campaigns (2nd in 2006 and 2007). At QA@CLEF 2008, out of the 21 systems participating, there were two systems for Portuguese among the three best systems: Priberam (1st) and Senso [4] (3rd). IdSayBL was 5th overall. In addition to Priberam, Senso, and IdSayBL, the remaining participants in the Portuguese monolingual task were Esfinge [5], QA@L2F [6], and Raposa [7]. In 2008, the score of a hypothetical combination run that chose the best answers among the six participating systems for Portuguese would have a score of 168 right answers for all 200 questions, giving an accuracy over the first answer of 84%.

Considering the success of this combination, we made a detailed analysis of the results of the other five participants in the Portuguese monolingual task of QA@CLEF 2008, to find the strengths and weaknesses of each system and make a comparison with IdSayBL. Since it was a baseline, we expected this analysis to help us decide which improvements and new functionality were more liable to enhance our results, keeping in mind the two key characteristics we want to keep in the system: efficiency in terms of response time (the 200 questions are answered in less than 1 minute) and robustness because we plan to be able to treat different types of data.

Table 1 summarizes the results of the systems participating at QA@CLEF 2008 for Portuguese, and Table 2 presents their main characteristics, as published in the QA@CLEF literature.

1.2 Structure of the Paper

In Section 2 we make a critical comparative analysis of IdSayBL results against the results of the other systems. In Section 3 we describe the improvements made to IdSayBL. In Section 4 we present the results of the new version of IdSay, and in Section 5 we present our conclusions and future work.

[1] Results for individual questions of IdSayBL to QA@CLEF 2008 question set for Portuguese can be found in `www.idsay.net`

Table 1. Results overview

System	First Answer		All Answers		MRR
	R	accuracy	R	accuracy	
Priberam	127	63.5%	139	69.5%	66.3%
Senso	93	46.5%	93	46.5%	46.5%
IdSayBL	65	32.5%	85	42.5%	37.1%
Esfinge	47	23.5%	61	30.5%	26.6%
QA@L2F	40	20.0%	41	20.5%	20.3%
Raposa	29	14.5%	29	14.5%	14.5%
Combination	168	84.0%	174	87.0%	85.4%

2 Analysis of Results for Portuguese at QA@CLEF

We analysed all answers submitted at QA@CLEF 2008 in the Portuguese exercise, which represented around 3600 answers. The analysis of IdSayBL results was done in more detail since we had information that allowed us to determine the exact causes of failure. However, the analysis of other systems' answers also allowed us to have a better insight into their behaviour.

2.1 Comparative Analysis of IdSayBL Results

To reflect the performance of IdSayBL relatively to other systems, we present a table that highlights the global results of all systems for Portuguese (Table 3).

We make a comparison using the results considering only the first answer (the 2 top rows), and considering all the three answers that could be returned per question (the 2 bottom rows). The first column shows the number of questions that IdSayBL got right (R) or otherwise (not R, which includes Unsupported, ineXact and Wrong answers), given the condition of the corresponding row (1st answer or all three answers). In the rest of the row, each column indicates the number of systems that got right answers.

For instance, taking the first row of the table, we are considering only the first answers, so we start with the number of questions IdSayBL got right in those circumstances which is 65, and for these 65 we check the number of systems that also got first right answers. So the value 3, next to 65, means that from the 65 questions, 3 were only answered correctly by IdSayBL (or 0 other systems got them right). On the other end of that row, the value 2 indicates that of the 65 questions, 2 have been answered correctly by all systems.

In the next row we consider the questions that were not answered correctly by IdSayBL (135), and we make a similar analysis of the number of systems that got them right. Of these questions, 32 were not answered correctly by any of the systems, while on the other end of the row we can see that the number of questions that IdSayBL was not able to answer correctly but that all the other systems were able to give right answers is 0.

In general, a higher number of questions on the left hand side of the table represents a better performance of IdSayBL relative to the other systems.

Table 2. Characteristics of systems for Portuguese at QA@CLEF2008

System	Collection Pre-Analysis, Storage and Retrieval	Question Treatment	Answers Extraction and Scoring	Knowledge used besides the collections	Additional tools and techniques used
Priberam	Proprietary IR system; Stop Words and Lemmatization; Syntactical Processing; Indexing includes lemmas, heads of derivation, synonyms, ontological domains, and question categories.	Question Classification (around 84 categories); Syntactic Processing.	Manually obtained patterns; Scoring through patterns; Redundancy.	Proprietary Ontology; Thesaurus for query expansion; Lexicon for Lemmatization.	Proprietary parser; Morphological disambiguation; Named Entity Recognition.
Senso	Information Retrieval: system Lucene; Stop Words and Proprietary Stemmer.	Multi Strategy: Semantic Representation; Extraction of query words for ad hoc search.	Multi Strategy: Top Documents from IR are treated in two different ways: Semantic Representation and inference mode PROLOG style; Search documents via manually obtained patterns. Answer Validation using web search.	Proprietary Ontology including Thesaurus for query expansion	Parser Palavras; Answer Validation using web search.
IdSayBL	Proprietary IR system - Boolean Model; Lemmatization; Indexes for words and entities.	Question Classification (7 types of question); Extraction of Reference Entity and other words and entities for retrieval.	Entity Recognition; Manually obtained patterns; Direct extraction taking advantage of encyclopaedic structure of Wikipedia; Redundancy.	Lexicon POLLUX for Lemmatization.	Exact algorithm to find entities (words co-occurring more than a given threshold in the collection).
Esfinge	Documents stored as sentences: IMS Workbench for News Articles; MySQL for Wikipedia.	Question transformed into patterns in two ways: String processing; And using Parsing.	Three strategies: 1st Manually constructed patterns; 2nd NER SIEMÊS; 3rd N-grams model. Score: Answer frequency; Words of question in support; Candidate answer length.	Database of Word Co-Occurrences BACO.	Parser Palavras; Named Entity Recognizer SIEMÊS; Morphological Analyser JSpell; N-gram filtering Ngram Statistics Package (NSP); Search answers in web using Yahoo search API; Filters to exclude undesirable answers.
QA@L2F	MySQL News Articles; Heavy linguistic processing to build two data bases from the collection: Relation-Concept database; Named Entities database; and Raw text database. Wikipedia: Raw text data base.	Questions analysed by type and a script per type builds query for database.	Patterns; Named Entities Recognition.		L2F morphology processing chain (POS tagging, Tokenization, Disambiguation: Palavroso, Rudrico, Marv) XIP Parser (syntactic analysis). NER based in the above NLP chain.
Raposa	MySQL; stores the collection as sentences or paragraphs.	Rule based classification of question with special incidence in people questions.	NER; Context Rules or if there aren't any, Semantic Label/ Redundancy;	Database of Word Co-Occurrences BACO; Verb Thesaurus automatically built from an external large collection.	Morphological Analyser JSpell; Named Entity Recognizer SIEMÊS.

Table 3. Comparison of IdSayBL results and those of other systems

		IdSayBL	0 systems	1 system	2 systems	3 systems	4 systems	5 systems
First Answer	R	65	3	17	24	12	7	2
	Not R	135	32	40	38	19	6	0
All (3) answers	R	85	3	20	35	12	11	4
	Not R	115	26	32	36	15	6	0

In a row where we are considering right answers from IdSayBL, the questions that are included on the right of this row can be considered easy questions, and those included on the left side are questions in which IdSayBL proved to make a positive contribution when compared to other systems.

In a row where we consider the questions that IdSayBL was not able to answer correctly, the questions on the left are those generally not well covered by the systems, and can be considered difficult for the current state of the art. On the right hand side of this row are the questions which we take into special consideration, because they reveal weaknesses of our system, and that all or almost all the other systems were able to address successfully.

2.2 Characterization of QA Systems

For a characterization of the relative performance of a QA system, we created a condensed form of Table 3, first answers only. We call it a results quadrant and its meaning is explained in Table 4. In each quadrant is the percentage of questions that meet the criteria in the corresponding upper and left border.

We built a results quadrant for each of the six systems, but for reasons of space we present the summary of the most significant results: Priberam is the only system classified as "innovative" (34%), also having the highest value in the "good coverage of easier questions" (30%). Senso (32%) and IdSayBL (36%) are in the quadrant "needs to make an investment to cover more difficult questions" with the first being more innovative (19% versus 10%) and having a better coverage for easier questions than IdSayBL (28% versus 23%), while IdSayBL has more room to learn from the others (32% against 22%). Esfinge, QA@L2F, and Raposa all have higher percentages in the "liable to learn with the experience of others" quadrant, respectively 40%, 43%, and 48%.

Table 4. Meaning of Results Quadrant

Results Quadrant	0 or 1 systems	2, 3, 4 or 5 systems
R	innovative	good coverage of easier questions
Not R	needs to invest in new techniques to cover more difficult questions	liable to improve with the experience of others

Taking into account these results, we made a detailed question-based analysis, focusing on the questions in the quadrant "liable to improve with the experience of others", in order to identify the enhancements that could lead to better results. These improvements are described in the next section.

3 Improving IdSay

The results of the analysis validate the architecture of IdSay, which is presented in Fig. 1, and is similar to that of IdSayBL.

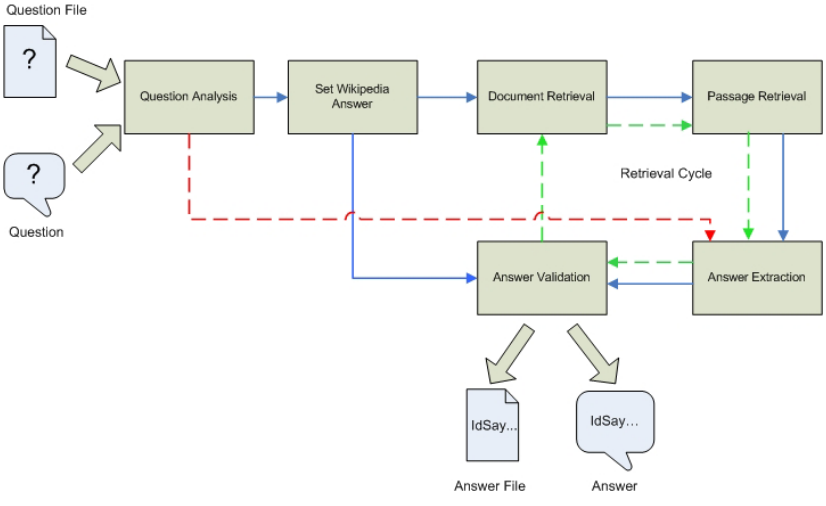

Fig. 1. IdSay system architecture

The results of the analysis seem to validate our option to use an IR module as the base component for search. This option is common among QA Systems, that rely on the ranking of documents both to treat a small number of documents and to score answers. Our approach in IdSay is different: we aim at processing all documents that have occurrences of the words and entities (two or more words co-occurring together in the same order) in the question. For that we developed a retrieval module that has two indexes: one for words and the other for entities. We tried to keep the entire process as simple and fast as possible, especially the modules in the retrieval cycle, so that it would be possible to process a large amount of information (many documents) a large number of times (many cycles). We focus our ranking efforts on the answers, taking redundancy as a key factor.

In our analysis we identified, as main sources of other systems' success, the use of semantic information resources that would influence the retrieval and validation of answers, and the use of linguistic analysis tools that accounted for better question treatment and answer extraction. Since one of the goals of IdSay is robustness, in the sense that it may be used with noisy data, as that produced

automatically by speech transcription or machine translation, we opted for the introduction of semantic information instead of deep linguistic processing.

3.1 Introduction of Semantic Equivalences

Given the design options described for IdSay, it is natural to give priority to improving the recall of our baseline by searching for words and entities that are semantically equivalent to those present in the question.

The introduction of semantic equivalences between words and entities used two different sources. Since we privilege the use of public free resources, for equivalences at a word level for Portuguese we used the TeP base [8,9], and for equivalences at an entity level, we built a resource based on Wikipedia, which we called Wikipedia Entity Synonyms (WES base) that we compiled using the Portuguese Wikipedia Version that is also part of the text collection of QA@CLEF[2]. The WES base is obtained from the Wikipedia using context pages names and redirect files. The format is compatible with TeP base: the synset[3] starts by the canonical name of the entity in the Wikipedia (the name of the file that has content) followed by the existing alternative names of the entity (names of redirecting files to the content file). The alternative names are separated by a comma surrounded by spaces, since the names may themselves contain commas, but never surrounded by spaces on both sides. The file has 46586 synsets, ordered alphabetically by canonical names, and the names are kept exactly as they appear in Wikipedia, without any processing. An Example is:

15713. [Wikipedia] {*Fernando Henrique Cardoso , Presidente Fernando Henrique Cardoso , Fernando Henrique , FHC*}

The WES base proved to be a valuable resource to the system and helped answering several questions. Using the example given, for (Question#23 "Quem é FHC"?) [Who is FHC?], the equivalence between FHC and "Fernando Henrique Cardoso" allowed the retrieval of the correct answer with information on the former Brazilian President.

3.2 Ontological Knowledge Based on Wikipedia

Both Priberam and Senso use a proprietary ontology. Priberam's ontology is described in [10], and mentions that it was obtained by translating 195,000 English entries. These entries are organised by senses, in an hierarchical four level structure that lead from 28 top nodes to 3387 end nodes. Senso's ontology [11] has 3500 concepts connected by relations such as *isA*, *usedFor*, *locatedAt*, *capableOf* and *madeOf*. Although the two systems make different uses of this information, our analysis suggested that it is an important component for the overall results, so we opted for incorporating ontological knowledge into IdSay, to be used during answer extraction, and to validate if a candidate answer has the type determined

[2] Available at the resources section of IdSay web page.

[3] Synset is an entry representing a semantic equivalence between entities, in this case.

by the Question Analysis module. The most natural way to do that was once more to use Wikipedia[4], which can be viewed as a simple ontology if one considers its category structure [12].

We implemented a generic *IsA* filter, for which the category is extracted from the question text directly, and filters for Person, Organization, Location. Since they do not correspond to exact categories of Wikipedia, we manually built authority lists for words that can be considered as part of that category, and we search for them in the category string. We used the structure of synonyms described in the previous section: for instance for the entity we were checking we used the synonyms between entities in order not to miss the correct Wikipedia page if it existed, and for lists of words to search in the category string.

Although we were limited to the fact that only entities in the Wikipedia could be checked, and had to rely on Wikipedia's sometimes incomplete or tangled category structure, we consider the filters implemented very helpful. The filters that yielded better results were generic *IsA* and *IsAPerson*. Organizations and Locations proved more difficult to filter, but nevertheless there was an improvement in answering this type of questions.

3.3 Other Improvements

The current version of IdSay also benefited from changes and enhancements in other parts of the system, namely: the scoring mechanism of the answers, that in IdSayBL relied almost exclusively in the number of occurrences of the answer in the collection, was substantially changed. It now takes into account information that starts at passage level, reflecting the percentage of words that were presented in the question versus the ones that were obtained through equivalencies, and reflects the method that was used to obtain the answer, for instance valuing information that was obtained directly from the Wikipedia web page, or the number of retrieval cycle in which the answer was obtained, among other factors. We improved the treatment of acronyms and abbreviations, normalizing them, and this change also had benefits in terms of dates with centuries with AC[BC] qualification. The treatment of numeric values was also taken into account with normalization taking place in terms of milliard separator and decimal separator. The passage retrieval module was improved in a way that they become closer to meaningful sentences, and list questions were considered for the first time. Some minor adjustments were also made in the Question Analysis module, especially regarding co-reference resolution in cluster questions.

4 Results of Improved Version

The results of IdSay after the improvements described in the last section, are presented in Table 5.

[4] In accordance with our preference for using Community Knowledge Bases that require minimum manual maintenance from single researchers/groups.

Table 5. Comparison of IdSay results and those of other systems

		IdSay	0 systems	1 system	2 systems	3 systems	4 systems	5 systems
First Answer	R	101	5	23	37	26	8	2
	Not R	99	30	34	25	5	5	0
All (3) answers	R	124	9	27	48	23	13	4
	Not R	76	20	25	23	4	4	0

Although IdSay still has room for improvement, our approach seems to be validated, keeping efficiency (the 200 questions are answered in less than 2 minutes). IdSay produced original results in some cases, as denoted by the 3 questions that it was the only system that was able to answer in the BL version, that increased to 5 and 9 in the current version, considering the first answer only or all the three possible answers, respectively.

Since the changes are made only to IdSay, an improvement means that a question moves from the row of "not R" to the above "R" row.

In Table 6, we present the changes according to the quadrant of Table 4.

Table 6. Final Results Quadrant for IdSay

Results Quadrants IdSay (IdSayBL)	0 or 1 systems	2, 3, 4 or 5 systems
R	14% (10%)	37% (23%)
Not R	32% (36%)	18% (32%)

These results indicate that we were able to explore some of the potential that lies in the experience of veteran systems, while being able to also innovate.

5 Conclusions and Future Improvements

We described a new version of IdSay, specifying the enhancements that were made to IdSayBL. We obtained significant improvement, from an accuracy in the first answer of 32.5% in IdSayBL to 50.5% in IdSay, without degradation of response time.

There are still many improvements that can be added to IdSay. We need to implement a more thorough treatment of time, with the normalization of time values and the resolution of relative time values, possibly with the use of metadata, which we do not currently use. Co-reference resolution is still a process that can be further pursued and that is especially relevant for cluster questions, but is also important in terms of the data collection, especially references to names of Wikipedia pages in their body.

We plan to improve on the techniques we already have implemented, for example the treatment of ambiguity at the semantic equivalences level, or extend the classification of questions using machine learning.

Among all possibilities, we must make choices, and as our next step we intend to use IdSay with other types of data, namely data obtained from Automatic Speech Recognition (ASR). We also intend to extend the system to other languages besides Portuguese.

References

1. Carvalho, G., de Matos, D.M., Rocio, V.: IdSay: Question Answering for Portuguese, pp. 345–352. Springer, Heidelberg (2009)
2. Forner, P., Peñas, A., Agirre, E., Alegria, I., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P.: Overview of the Clef 2008 Multilingual Question Answering Track, pp. 262–295. Springer, Heidelberg (2009)
3. Amaral, C., Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pina, J., Pinto, C.: Priberamas Question Answering System in QA@CLEF 2008, pp. 337–344. Springer, Heidelberg (2009)
4. Saias, J., Quaresma, P.: The senso question answering system at qa@clef 2008. In: Working Notes of the CLEF 2008 Workshop (2008)
5. Costa, L.F.: Using Answer Retrieval Patterns to Answer Portuguese Questions, pp. 361–368. Springer, Heidelberg (2009)
6. Coheur, L., Mendes, A., Guimarães, J., Mamede, N.J., Ribeiro, R.: Question Interpretation in QA@L2F, pp. 377–384. Springer, Heidelberg (2009)
7. Sarmento, L., Teixeira, J., Oliveira, E.: Assessing the Impact of Thesaurus-Based Expansion Techniques in QA-Centric IR, pp. 325–332. Springer, Heidelberg (2009)
8. Dias-da-Silva, B.C., Oliveira, M.F., Moraes, H.R., Hasegawa, R., Amorim, D., Paschoalino, C., Nascimento, A.C.A.: Construção de um thesaurus eletrônico para o português do brasil. In: V Encontro para o Processamento computacional da Língua Portuguesa Escrita e Falada, Atibaia, Brazil, vol. 4, pp. 1–10 (2000)
9. Maziero, E.G., Pardo, T.A., Felippo, A.D., Dias-da-Silva, B.C.: A base de dados lexical e a interface web do tep 2.0 - thesaurus electrônico para o português do brasil. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), pp. 390–392 (2008)
10. Amaral, C., Laurent, D., Martins, A., Mendes, A., Pinto, C.: Design and implementation of a semantic search engine for portuguese. In: ELRA (ed.) Proc. of the 4th Language Resources and Evaluation Conference – LREC 2004, Lisboa, Portugal, pp. 247–250 (2004)
11. Saias, J., Quaresma, P.: The University of Évora's Participation in QA@CLEF-2007, pp. 316–323. Springer, Heidelberg (2008)
12. Yu, J., Thom, J.A., Tam, A.: Ontology evaluation using wikipedia categories for browsing. In: CIKM 2007: Proc. of the 16th ACM Conf. on Information and Knowledge Management, pp. 223–232. ACM, New York (2007)

Assessing the Impact of Stemming Accuracy on Information Retrieval

Felipe N. Flores, Viviane P. Moreira, and Carlos A. Heuser

Instituto de Informática – UFRGS
Av. Bento Gonçalves, 9500, 91501-970 Porto Alegre, Brazil
{fnflores,viviane,heuser}@inf.ufrgs.br

Abstract. The quality of stemming algorithms is typically measured in two different ways: (i) how accurately they map the variant forms of a word to the same stem; or (ii) how much improvement they bring to Information Retrieval. In this paper, we evaluate different Portuguese stemming algorithms in terms of accuracy and in terms of their aid to Information Retrieval. The aim is to assess whether the most accurate stemmers are also the ones that bring the biggest gain in Information Retrieval. Our results show that some kind of correlation does exist, but it is not as strong as one might have expected.

Keywords: Stemming, information retrieval, evaluation.

1 Introduction

Stemming is the conflation of the variant forms of a word into a single representation, i.e. the stem. For example, the terms *presentation, presenting*, and *presented* could all be stemmed to *present*. The stem does not have to be a valid word, but it needs to capture the meaning of the words. Stemming is usually done by algorithms that strip word suffixes.

The quality of stemming algorithms is typically assessed in one of these manners: (*i*) how correctly the stemmer maps semantically and morphologically related words to the same stem; or (*ii*) how much improvement the stemmer brings to Information Retrieval (IR).

Stemming is widely used in IR with the aim of increasing recall (i.e. the number of relevant documents retrieved in response to a user query). A number of studies report on the effectiveness of using stemming in an IR system, especially for the English language [5, 6, 8].

There are also studies on the quality of stemming algorithms [2, 7, 11], which use an evaluation method proposed by Paice [13]. But a study on the impact of stemming accuracy on IR performance has not yet been done. Thus, the purpose of this paper is to answer the question: Is the most accurate stemmer the best for IR?

We carried out experiments with all Portuguese stemming algorithms found in the literature to measure their accuracy and also to assess the gain they bring to IR. Thus, as a byproduct, this paper identifies the most accurate Portuguese stemmer and the one that yields the biggest IR improvement.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 11–20, 2010.

The remainder of this paper is organized as follows: Section 2 discusses related work and introduces some background concepts for the evaluation of stemming algorithms; Section 3 presents the experiments that measure the accuracy of the Portuguese stemmers; Section 4 describes the IR experiments done to compare the impact of the stemmers over retrieval effectiveness; Section 5 investigates the correlation between both quality indicators, and Section 6 presents our conclusions.

2 Background and Related Work

Paice [13] proposed a method to evaluate the quality of stemmers using four metrics:

- **Overstemming Index** (OI), which calculates the number of times a stemmer mistakenly removes part of the stem as if it were part of the suffix. This type of error will typically cause unrelated words to be combined, e.g. *adoção* (adoption) and *adoçante* (sweetener) are both stemmed to *adoç*. OI is zero when there are no overstemming errors and one when all words are stemmed to the same stem. In IR, a high OI will potentially lead to a decrease in precision (the fraction of the retrieved documents which is indeed relevant).
- **Understemming Index** (UI), which calculates the number of times a stemmer fails to remove a suffix. This type of error will typically prevent related words from being conflated, e.g. if *jornalista* (journalist) is stemmed to *jornalist* and *jornalismo* (journalism) is stemmed to *jornalism*. UI is zero when there are no understemming errors and one when no words are correctly combined by the stemmer. In IR, understemming errors will potentially lead to a decrease in recall (the fraction of the relevant documents which is actually retrieved).
- **Stemming Weight** (SW), which is the ratio OI / UI. A 'heavy' stemmer will apply lots of affix stripping rules and thus have a high OI and a low UI, while a 'light' stemmer will apply fewer affix stripping rules and thus have a low OI and a high UI. Having a high or low SW does not indicate that we are dealing with a better stemmer, since it is only used to assess a stemmer's 'weight'.
- **Error Rate Relative to Truncation** (ERRT). The idea is that the values of (UI,OI) for a series of truncation lengths (e.g. trunc-3 to trunc-8) determines a line against which the quality of the stemmer could be assessed. The coordinates (UI,OI) for an adequate stemmer should be on the lower side of this truncation line (see Figure 1). The farther away from the truncation line, the better the stemmer. ERRT is obtained by extending a line from the origin through the (UI, OI) point *P* until it intersects the truncation line at *T*. The ERRT is then calculated by the following formula: ERRT = length(OP) / length(OT). The lower the ERRT, the better the stemmer.

Paice [13] uses his proposed method to evaluate three English stemmers: Lovins [9], Paice/Husk [12], and Porter [14]. He concludes that the most accurate stemmer is Paice/Husk, followed by Porter. The author also observed that Porter is lighter than Paice/Husk. The study tested word samples with different sizes (from 2654 to 9757) and concluded that the pattern of values were similar for all sample sizes. This study did not compare the effect of the stemmers on retrieval performance.

Kraaij & Pohlmann [7] employed Paice's evaluation method to assess the quality of a Dutch version of the Porter stemmer. Again this study does not test the effect of the stemmer on retrieval accuracy.

For the Portuguese language, Orengo & Huyck [11] use Paice's evaluation method to compare their proposed RSLP stemmer to the Portuguese version of the Porter stemmer. They conclude that RSLP is more accurate than Portuguese Porter. Also in Portuguese, Alvares et al.[2] compare STEMBR, RLSP and Portuguese Porter stemmers. The study concludes that STEMBR is heavier than Porter and RSLP and thus makes slightly fewer understemming errors than RSLP and Porter. Both papers, however, have not used the ERRT metric in their evaluation. Also none of them assesses the effects of their stemmers over retrieval performance.

There are some papers that study the impact of stemming in an IR system, such as Harman [5], Krovetz [8], and Hull [6]. The general conclusion is that looking at the average result over a number of queries, the improvement brought by stemming is small. However, for some queries, stemming brings a large improvement. For the Portuguese language, the work by Orengo et al. [10] evaluates the effects of three different stemmers for IR and concludes that the light stemmer is the best alternative. In turn, this study did not evaluate the accuracy of the stemmers.

To evaluate the impact of stemming over retrieval effectiveness, the standard approach is to compare a baseline run with no stemming to a run in which stemming is used as a preprocessing step over documents and queries. The most widely used measure to assess the quality of retrieval results for a given query is *average precision* (AvP). AvP emphasizes returning more relevant documents earlier in the ranking. When more than one query is used, the average of the AvPs, which is called Mean Average Precision (MAP), must be calculated. To compute the retrieval effectiveness measures, it is necessary to use a test collection composed by documents, queries, and relevance judgments which list the relevant documents for each query.

To the best of our knowledge, the comparison between the quality of a stemming algorithm and its effectiveness in an IR system has not yet been done. Thus the main goal of our present study is to fill this gap and investigate the relationship between these two quality indicators.

3 Quality of Portuguese Stemming Algorithms

This section describes the procedures and the results obtained with our experiments of the accuracy of stemmers.

3.1 Experimental Setup

The following stemmers were used in the experiments:

- **Porter:** Portuguese version of the Porter stemmer, originally proposed for English in [14] and adapted Portuguese in [17]. It works with a set of suffix stripping rules.
- **RSLP:** Proposed in [11], it is also a suffix stripping algorithm based on rules.
- **RSLP-S:** Light RSLP algorithm, using only its Plural Reduction Step, which removes, for example, the *–s* and *–ns* plural suffixes.
- **Savoy:** Presented in [15], this stemmer removes inflectional suffixes for nouns and adjectives. It has many more rules than light stemmers RSLP-S and Stemmer-S, but also fewer rules than Porter, RSLP and StemBR.

- **Stemmer-S:** A simple light stemming algorithm, implemented by us, which only removes the final *s* of words.
- **StemBR:** StemBR algorithm, proposed in [2], which removes prefixes and suffixes.
- **TruncN:** A simple truncation algorithm, which simply removes all but the first *n* letters of a word. We have tested with *n* varying from three to eight. They will be referenced as **Trunc3**, **Trunc4**, **Trunc5**, **Trunc6**, **Trunc7** and **Trunc8**.

To serve as a baseline, there is also a run **NoStem**, in which no stemming was used. Table 1 shows the output of each stemmer with the same text as input. We can see that some stemmers are more aggressive than others in removing affixes.

Table 1. Sample output from the different stemmers to the same input text

NoStem	O debate político, pelo menos o que vem a público, parece, de modo nada surpreendente, restrito a temas menores. Mas há, evidentemente, grandes questões em jogo nas eleições que se aproximam.
Porter	o debat polit, pel men o que vir a public, parec, de mod nad surpreendent, restrit a tem menor. mas hav, evident, grand questõ em jog nas eleiçõ que se aproxim.
RSLP	o debat politic, pel menos o que vem a public, parec, de mod nad surpreend, restrit a tem men. mas ha, evid, grand quest em jog na ele que se aproxim.
RSLP-S	o debate político, pelo menos o que vem a público, parece, de modo nada surpreendente, restrito a tema menor. mas há, evidentemente, grande questão em jogo na eleição que se aproximam.
Savoy	o debat politic, pelo meno o que vem a public, parec, de modo nada surpreendent, restrit a tema menor. mas ha, evident, grand questa em jogo nas eleica que se aproximam
Stemmer-S	o debate político, pelo meno o que vem a público, parece, de modo nada surpreendente, restrito a tema menore. ma há, evidentemente, grande questõe em jogo na eleição que se aproximam.
StemBR	o deba polit, pelo meno o que vem a public, parec, de modo nad surpreend, restrit a tem meno. ma ha, evident, grand quest em jogo na ele que se aproxim.
Trunc-4	o deba polí, pelo meno o que vem a públ, pare, de modo nada surp, rest a tema meno. mas há, evid, gran ques em jogo nas elei que se apro.

The computation of the quality metrics proposed by Paice requires groups of semantically and morphologically related words. In order to generate such groups, we obtained a list of 31986 words available in [16] and selected a sample of 2854 words. The guideline in word selection was to have words with diverse starting characters and also to have groups of different sizes. The sample was manually divided into 888 groups. Thus each group had on average 3.21 words.

We implemented a tool to calculate Paice's evaluation method. The tool calculates all four quality metrics including ERRT, which involves computing overstemming and understemming indices for different truncation algorithms and constructing a line (the truncation line) by connecting the coordinates (UI, OI) for each truncation length (see Figure 1). Then, the tool calculates the distance between the coordinate (UI, OI) obtained from the stemming algorithm and the truncation line. This implementation,

which is language-independent, i.e. it can be used to evaluate stemming algorithms of any Latin-character languages (English, Portuguese, Spanish, Italian, Finnish…), can be freely accessed at [1].

3.2 Results

In this section, we show the results obtained using Paice's evaluation method. Table 2 shows the figures for all four quality metrics. The table is divided in three parts, the top rows refer to Portuguese stemming algorithms, the rows in the middle refer to truncation algorithms, and the bottom row shows the figures obtained when no stemming is used. The best results for each metric are in bold.

Table 2. Result of Paice's evaluation

	UI	OI	SW	ERRT
Porter	0.3014530471	0.0001468286	0.0004870694	0.7159116889
RSLP	**0.1905226632**	0.0002680360	0.0014068458	**0.5691374097**
RSLP-S	0.9515289525	**0.0000004927**	0.0000005178	0.9959400023
Savoy	0.8863372767	0.0000123683	0.0000139544	1.1300856452
Stemmer-S	0.9580351334	**0.0000004927**	0.0000005143	1.0027066652
StemBR	0.2668618521	0.0003271616	0.0012259587	0.7650424144
Trunc3	**0.0304706137**	0.0157951554	0.5183733933	1.0000000000
Trunc4	0.1078941661	0.0029855960	0.0276715241	1.0000000000
Trunc5	0.2676209065	0.0006777271	0.0025324146	1.0000000000
Trunc6	0.4520711342	0.0001096287	0.0002425032	1.0000000000
Trunc7	0.6484493602	0.0000209403	0.0000322929	1.0000000000
Trunc8	0.8150075905	**0.0000083761**	0.0000102774	1.0000000000
NoStem	1.0000000000	0.0000000000	0.0000000000	1.0000000000

In terms of UI, the best result is obtained by the Trunc-3 as it rarely fails to conflate related forms of words. On the other hand, Trunc-3 has the highest OI. The best UI of a stemmer is achieved by RSLP. In terms of OI, the best result is when no stemming is applied, however that has the highest UI. The best OI of a stemming algorithm is obtained by RSLP-S. UI and OI alone do not enable the identification of the best algorithms.

According to ERRT, the best stemmers are, in order, RSLP, Porter, StemBR and RSLP-S, with the fourth well behind the first three. We can also notice that RSLP and StemBR are heavier than Porter and Savoy, according to SW. This fact was also found in [2] and [11]. Besides, even though RSLP has a similar SW to StemBR, the first outperforms the last in both UI and OI. However, RSLP outperforms Porter and Savoy only in UI, since Porter and Savoy have better OIs. Finally, we can also see that the light stemmers, RSLP-S and Stemmer-S, had very bad results in ERRT, being very close to 1. This happens because their UI is very high, also close to 1, since they have very few affix stripping rules and therefore only a small number of related words end up with the same stem. Besides, ERRT is exactly 1 for all truncate stemmers

(since they are obviously on the truncation line) and for NoStem, since it can be seen as a TruncN algorithm, with *n* being the length of the longest word of the sample.

Figure 1 shows a graphic interpretation of ERRT. Recall that a good stemmer should be on the lower side of the truncation line, and as far as possible from it. Thus RSLP can be considered the best stemmer, followed by Porter and StemBR. The light stemmers RSLP-S and Stemmer-S, as well as the medium weight stemmer Savoy, are close to or above the truncation line, and therefore can be considered worse.

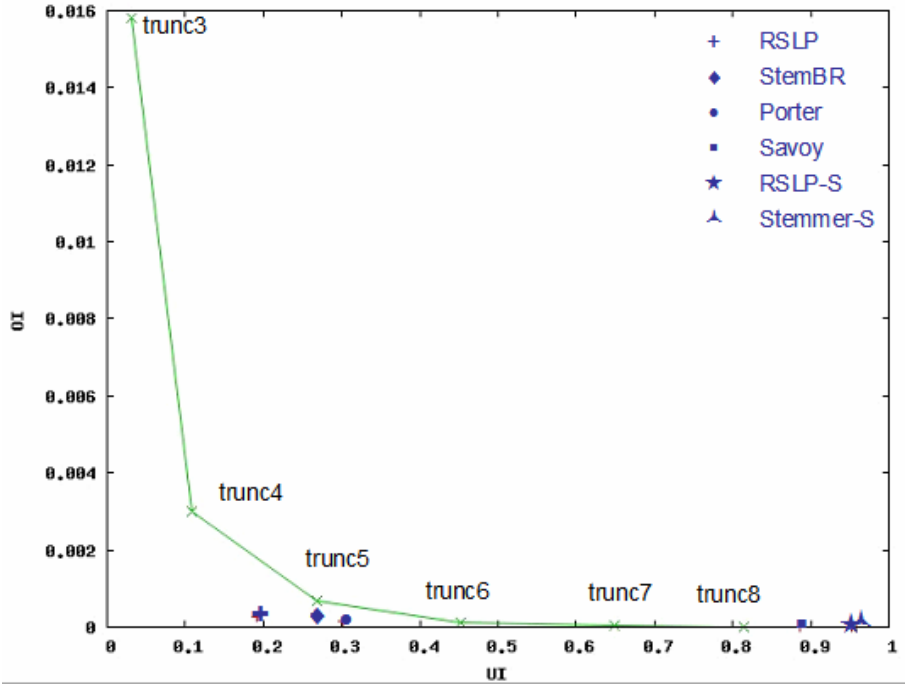

Fig. 1. ERRT Plot for the Portuguese Stemmers

4 Portuguese Stemmers and Information Retrieval

This section presents the experiments we carried out in order to assess how the different stemmers impact retrieval effectiveness.

4.1 Experimental Setup

In this experiment, we used the Portuguese data collection from the CLEF campaigns [3] of 2005 and 2006 (obtained from ELDA [4]), which consists of articles from *Folha de São Paulo* newspaper, from the years 1994 and 1995. A total of 98 query topics were used.

To index the data collection we used the Zettair Search Engine [18]. Zettair has a stemming option that we can use when indexing documents. However, we chose not

to use it, since it was designed for English and we wanted to test our own Portuguese stemmers. Thus, we created a simple algorithm to process each word of the documents and the topics file with the stemmers, and then used these stemmed files as input to Zettair.

4.2 Results

In this section, we show the results of mean average precision (MAP), number of relevant documents retrieved, number of distinct terms indexed and speed performance for each stemmer. The results are summarized on Table 3.

Table 3. Results of the Retrieval Experiment

	Distinct Index Terms	**Relevant Retrieved**	**Speed Performance**	**MAP**
Porter	250864 (-26.51%)	1680	00:57:35	0.2981
RSLP	237490 (-30.43%)	1684	00:06:43	0.2898
RSLP-S	313737 (-8.09%)	1695	00:06:34	0.2938
Savoy	306836 (-10.11%)	**1725**	00:01:15	**0.2984**
Stemmer-S	316753 (-7.21%)	1664	**00:01:08**	0.2736
StemBR	233536 (-31.59%)	1668	04:24:37	0.2829
Trunc3	**72750 (-77.36%)**	1355	00:01:14	0.1784
Trunc4	109093 (-68.04%)	1685	00:01:14	0.2706
Trunc5	163068 (-52.23%)	1696	00:01:14	0.2957
Trunc6	212723 (-37.68%)	1540	00:01:14	0.2751
Trunc7	250782 (-26.53%)	1512	00:01:14	0.2812
Trunc8	280803 (-17.74%)	1465	00:01:14	0.2718
NoStem	341362	1367	-	0.2587

The second column of Table 3 shows the number of distinct index terms after Zettair indexed the collection. Trunc3 had the best score here, since one of the goals of stemming is to reduce the size of the index. The largest reduction rate among the stemmers was obtained by StemBR.

The third column of Table 3 shows the total number of relevant documents retrieved. With the exception of Trunc-3, all stemming strategies enabled the retrieval of more relevant documents. The stemmer which enabled the retrieval of the largest number of relevant documents was Savoy, followed by RSLP-S.

The fourth column of Table 3 shows the speed performance, i.e. how long the stemmers spent in processing all files from *Folha de São Paulo*. The results show that StemBR was by far the slowest algorithm, spending more than four hours to complete the stemming process, while Porter took almost one hour and RSLP took only six minutes. StemmerS and TruncN, which are very simple, only took about one minute. The fastest "real" stemming algorithm was Savoy. We can also see that RSLP and RSLP-S had no significant difference in speed, meaning that the algorithm does not perform much faster if it ignores all but the plural reduction rules.

The fifth column of Table 3 shows the results for MAP. We can see that Savoy stemmer achieved the best result in improving IR effectiveness, while Trunc3 was the worst. We also performed paired *T-tests*, in order to see if the differences in MAP

were statistically significant. We used 0.05 as threshold for statistical significance. By analyzing the results of the *T-tests*, we can see that even though all stemmers (except Trunc3) had a better MAP than NoStem, only Savoy, Porter, RSLP, RSLP-S, StemmerS, Trunc5 and Trunc7 have shown a statistically significant improvement. Trunc3 was significantly worse than all other stemmers, including NoStem. We also concluded that StemBR was significantly worse than Savoy, Porter, and RSLP. Besides, RSLP-S was significantly better than StemmerS and Trunc8.

5 Correlation between Stemming Accuracy and IR performance

In this section, we analyze correlations between stemming accuracy (as calculated by Paice's evaluation method) and IR's retrieval performance measured by MAP and the number of relevant documents retrieved.

Table 4. Correlation between Stemming Accuracy (OI, UI, SW, ERRT) and IR Performance (MAP and Relevant Retrieved)

	MAP	Relevant Retrieved
Overstemming Index	-0.92	-0.50
Understemming Index	0.34	-0.08
Stemming Weight	-0.92	-0.26
ERRT	-0.17	-0.53

Intuitively, one would expect a strong negative correlation between the errors made by the stemmer and the measures for IR performance such as MAP and number of relevant documents retrieved. However, that assumption is not always correct. The correlation between ERRT and MAP is weak, only -0.17. Analyzing the results of the experiments from Sections 3 and 4, we observed that a stemmer that has lower ERRT is not necessarily better for IR than another stemmer with a higher ERRT. This is especially true in cases of light stemming, like RSLP-S, which had a very good performance in IR, but an ERRT close to one. Remarkably, Savoy and StemmerS, which had the highest ERRTs, have shown statistically significant improvements in relation to NoStem.

ERRT computes the distance from the truncation line, assuming that a truncation stemming algorithm is bad and should be avoided. However, that turned out to be false, as Trunc5 and Trunc7 had good results. Nevertheless, ERRT was able to distinguish the quality of stemmers of similar weight. For the light stemmers, RSLP-S was significantly better in MAP than StemmerS and had lower ERRT as well. For the heavy stemmers, StemBR had higher ERRT than Porter and RSLP, and also performed significantly worse in terms of MAP. RSLP, however, was much better than Porter in ERRT, but they did not have a significant difference in terms of MAP.

The strong negative correlations were between MAP and OI, and MAP and SW. This fact indicates that, for IR purposes, overstemming errors are more serious than understemming errors. As shown in Table 4, UI is positively correlated with MAP, even though this correlation is weak. This conclusion corroborates previous studies which advocate the use of lighter stemming alternatives for IR.

It was also expected that the number of relevant documents retrieved would decrease in the presence of understemming errors. However, we could not find a correlation between these two measurements.

6 Conclusions

This paper evaluates all Portuguese stemmers found in the literature in terms of their accuracy and in terms of their impact on retrieval effectiveness. Our aim was to assess if the quality of a stemming algorithm, particularly the quality measured by Paice's evaluation method, would translate into IR improvement. Our conclusion was that the most accurate stemmer was not the one to achieve the greatest improvement in IR.

We have implemented Paice's evaluation method and made the tool available on the Internet. With our implementation and a sample of words, it becomes very simple and fast to analyze stemmers for any language according to Paice's evaluation method. The tool calculates UI, OI, SW, and ERRT. If one does not have any previous information about the stemmers to be tested, Paice's evaluation method correctly shows which ones are light and which ones are heavy stemmers. It also correctly spotted the better stemmers when they had similar weights but significant difference in quality.

Further studies can be done to see if the same correlations found in this paper will apply to other languages besides Portuguese. Also, further studies can use our implementation of Paice's evaluation method and the results we obtained here to improve existing stemming algorithms.

Acknowledgments

This work was partially supported by CNPq (Brazil).

References

1. Flores, F.N.: http://www.inf.ufrgs.br/~fnflores/paice_tool
2. Alvares, R.V., Garcia, A.C.B., Ferraz, I.: STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 693–701. Springer, Heidelberg (2005)
3. CLEF. Cross-Language Evaluation Forum, http://www.clef-campaign.org (July 11, 2007)
4. ELDA. ELDA, http://www.elra.info/ (October 15, 2008)
5. Harman, D.: How Effective is Suffixing? Journal of the American Society for Information Science 42(1), 7–15 (1991)
6. Hull, D.A.: Stemming Algorithms: A Case Study for Detailed Evaluation. Journal of the American Society for Information Science 47(1), 70–84 (1996)
7. Kraaij, W., Pohlmann, R.: Evaluation of a Dutch Stemming Algorithm. In: T.G. Publishing (ed.) The New Review of Document & Text Management, London, UK, pp. 25–43 (1995)
8. Krovetz, R.: Viewing Morphology as an Inference Process. In: 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191–202 (1993)

9. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics 11, 22–31 (1968)
10. Orengo, V.M., Buriol, L.S., Coelho, A.R.: A Study on the Use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 91–98. Springer, Heidelberg (2007)
11. Orengo, V.M., Huyck, C.R.: A Stemming Algorithm for the Portuguese Language. In: 8th International Symposium on String Processing and Information Retrieval (SPIRE), Laguna de San Raphael, Chile, pp. 183–193 (2001)
12. Paice, C.D.: Another Stemmer. SIGIR Forum 24, 56–61 (1990)
13. Paice, C.D.: An Evaluation Method for Stemming Algorithms. In: Croft, W.B., van Rijsbergen, C.J. (eds.) 17th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–50. ACM, Dublin (1994)
14. Porter, M.F.: An Algorithm for Suffix Stripping. Program 14(3), 130–137 (1980)
15. Savoy, J.: IR Multilingual Resources at UniNE (Portuguese Stemmer) (October 6, 2009)
16. Snowball, `http://snowball.tartarus.org/portuguese/voc.txt` (July 29, 2003)
17. Snowball_Stemmers. Snowball Stemmers, `http://snowball.tartarus.org/texts/stemmersoverview.html` (October 28, 2008)
18. Zettair, `http://www.seg.rmit.edu.au/zettair/` (cited 2007, June 11, 2007)

Exploiting Multilingual Grammars and Machine Learning Techniques to Build an Event Extraction System for Portuguese

Vanni Zavarella[1], Hristo Tanev[1], Jens Linge[1], Jakub Piskorski[2], Martin Atkinson[1], and Ralf Steinberger[1]

[1] JRC - European Commission
[2] Polish Academy of Sciences
vanni.zavarella@ext.jrc.ec.europa.eu

Abstract. We describe a methodology for building event extraction systems. The approach is based on multilingual domain-specific grammars and exploits weakly supervised machine learning algorithms for lexical acquisition. We report on the process of adapting an already existing event extraction system for the domain of conflicts and crises to the Portuguese language.

1 Introduction

The multilingual event extraction system NEXUS, which is part of the Europe Media Monitor family of applications (EMM) [1], aims at identifying violent events, man made and natural disasters and humanitarian crises in news reports. Currently, Nexus can handle 4 languages - English, French, Italian, and Russian.

The system's overall architecture and algorithms are highly language independent, involving the use of language-specific dictionaries and extraction grammars which are plugged in as external resources; therefore, adding a new language requires no modification of the system itself.

NEXUS takes on its input the result of EMM processing, consisting of clusters of news articles grouped according to content similarity, and tries to extract event report summaries from each such a cluster. Each article in the cluster is first linguistically preprocessed by our in-house core linguistic engine (including tokenization, sentence splitting, dictionary look-up), then a cascade of extraction grammars is applied in order to identify phrases which report about victims and entities and their participation in the event.

For each cluster, NEXUS tries to detect and extract only the main event by analyzing all of of its articles, to produce a frame whose main slots are: *Date*, *Location*, *Dead (Number, Description)*, *Injured (Number, Description)*, *Kidnapped (Number, Description)*, *Arrested (Description)*, *Perpetrator(Description)* and *Weapons*.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 21–24, 2010.

2 Outline of the Extraction Grammars

The role of the grammars deployed in NEXUS is recognition of phrases which introduce event participants. The extraction process is performed by devising a multi-layer grammar cascade using EXPRESS, a finite state-based grammar formalism which processes regular expressions over flat feature structures [2].

Grammars detect certain named entities (e.g., person names), quantity expressions, simple chunks referring to person groups (e.g., *cinco policiais, milhares de portugueses*), with their appositive and coordinated combinations.

Person and entity recognition grammar is abstracting from surface forms and relies mainly on: a) person name and partial noun phrase syntactic structure; b) lexical resources for the target language. Since a) has little cross-language variation the bulk of system development consists mostly of providing suitable lexical resources, namely domain-specific lexicons, listing a number of (possibly multiword) expressions subcategorized into semantic classes.

Prior to person recognition grammar application, 1 or 2-slot event triggering surface patterns like the following English and Portuguese samples are matched on text for extraction of partial information on event roles, such as actors, victims, etc.

```
<DEAD> was shot by <PERPETRATOR>
<DEAD> foram mortas
```

Role slots (with their assignments in brackets) can be filled by phrases referring to persons or person groups. Linear patterns can be then combined with detected person and person group entities via extraction rules detecting event information.

3 Semi-supervised Resource Acquisition

The core of our approach is the usage of weakly supervised machine learning tools to acquire the language-specific resources which the system needs in order to process new languages. First, in order to acquire the linear patterns for extraction of victims, perpetrators, etc., we implemented an iterative news cluster-based pattern acquisition algorithm, which was originally suggested by [4]. Secondly, a new weakly supervised approach (based on [5]) is deployed for learning of semantic categories referring to people, weapons and others.

For each language our event extraction system should have lists of phrases belonging to certain semantic categories, namely weapons and persons (used intensively by the extraction grammars) and others.

Given our specific settings - only unannotated corpus of Portuguese news available, only few semantic classes to be learnt, we found quite relevant the weakly supervised term classification approach described in [5]: based on it we created our own term extraction and classification approach - Ontopopulis, which takes on its input a set of seed terms for each semantic category under consideration and an unannotated corpus of news articles.

The system performs two learning stages - Feature Extraction and Term Extraction.

For each category, we consider as a context feature each uni-gram, bi-gram or tri-gram n which co-occurs at least 3 times in the corpus as adjacent with any of the seed terms from this category. For each such a context feature n and a semantic category cat we calculate the score: $score(n, cat) = \sum_{st \in seeds(cat)} PMI(n, st)$ where $seeds(cat)$ are the seed terms of the category cat and $PMI(n, st)$ is the pointwise mutual information which shows the co-occurrence between the feature n and the seed term st.

Then the user performs manual feature selection from a list of 250 best scored features suggested by the system, so as to guarantee high quality features (e.g. *tiro de W*, *golpes de W*, for the class *weapons*).

The term extraction and learning stage takes the features and extracts as candidate terms uni-grams and bi-grams, which tend to co-occur with these features and do not contain stop words, numbers or capitalized letters.

We weight the term candidates, using algorithm similar to the one introduced in [5].

Eventually, the system orders the term candidates for each category by decreasing weight and filters out terms with a weight under a certain threshold. Then, the term list is given to the user for manual cleaning.

4 Experiments and Evaluation

For each language we performed the following resource-creation steps: running Ontopopulis to learn a dictionary of persons, weapons and other categories and manually validate and clean them; running the pattern learning algorithm for each of the following semantic roles: dead, wounded, kidnapped, perpetrator, and arrested and manually clean the output of the pattern learning algorithm.

Evaluation of Ontopopulis. Table 1 shows for each semantic category the number of seed terms, the number of the new terms learned by the Ontopopulis, the number of correct learned terms, the overall precision and the precision in the top 20 ranked terms, in an experiment using an unannotated corpus of 3,4 million titles of news articles as training data.

Table 1. Evaluation of Ontopopulis

	person	weapon	politician	vehicle	watercraft	edged weapon	crime	building
seed terms	48	26	46	135	28	20	33	73
learned	930	122	990	315	173	45	911	1035
correct	473	44	226	123	39	4	397	360
precision	51%	36%	22%	39%	22%	8.8%	43%	34%
prec.top 20	90%	60%	75%	85%	70%	20%	85%	75%

The fact that the accuracy is relatively high in the top 20 shows that the system properly orders the learned terms by decreasing reliability.

Evaluation of NEXUS. On a sample of 100 clusters on security and disaster-related topics collected by EMM during 30 consecutive days (April 2009) we ran a baseline version of the system (BL), based on seed linear patterns and seed dictionaries of persons and weapons, and a target version (TG) in which we added the cleaned output of Ontopopulis for the classes *person* and *weapons* and the output of the pattern learning algorithm.

Table 2. Evaluation of extraction of different roles in terms of F1-measure

	DEAD	WOUNDED	KIDNAPPED	ARRESTED
BL	0.62	0.53	0.54	0.29
TG	0.69	0.51	0.67	0.47

Table 2 compares F-measure figures for a subset of the event roles (due to data sparseness, some of them were not instantiated).

The average recall increase was of 12%, with the best one (for KIDNAPPED role) up to 20%. Moreover, the improvement in the recall was not at the cost of reduced precision.

In our experimental settings, we only performed one learning stage, with no fine-tuning. Therefore, whereas system performance in absolute terms was not excellent, we believe that figures on the improved performance of the learned systems are encouraging. Moreover, the approach seems to be portable in the same way over semantic domains. One possible research direction would be then to test the methodology on adapting the event extaction system to new application domains. The live event extraction system for Portuguese is publicly accessible at `http://press.jrc.it/geo?type=event&format=html&language=pt`

References

1. Steinberger, R., Pouliquen, B., van der Goot, E.: An Introduction to the Europe Media Monitor Family of Applications. In: Gey, F., Kando, N., Karlgren, J. (eds.) Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (2009)
2. Piskorski, J.: ExPRESS - Extraction Pattern Recognition Engine and Specification Suite. In: Proceedings of the International Workshop Finite-State Methods and Natural language Processing (FSMNLP 2007), Potsdam, Germany (2007)
3. Eleuterio, S., Ranchhod, E., Freire, H., Baptista, J.: A System of Electronic Dictionaries of Portuguese Lingvisticae Investigationes, vol. XIX, p. 2. Jonh Benjamins, Amsterdam (1995)
4. Piskorski, J., Tanev, H., Wennerberg, P.O.: Wennerberg: Extracting Violent Events From On-Line News for Ontology Population. In: 10th International Conference on Business Information Systems (2007)
5. Tanev, H., Magnini, B.: Weakly Supervised Approaches for Ontology Population. In: Proceedings of the European Chapter of the Association of Computational Linguistics (2006)

Formalizing CST-Based Content Selection Operations*

Maria Lucía Castro Jorge and Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
{mluciacj,taspardo}@icmc.usp.br

Abstract. This paper presents the definition and formalization of content selection operations based on CST (Cross-document Structure Theory) for multidocument summarization purposes.

Keywords: Multidocument summarization, CST, user preferences.

1 Introduction

Multidocument summarization (MDS) is a research area that intends to automatically produce a summary from a group of texts on the same topic [3]. A good MDS system must be able to produce a single summary telling the main points of the story or some parts of it for an interested reader, as background information or the evolution in time of some event. Therefore, content selection is an important task for MDS.

In this paper, we define and formalize content selection operations based on Cross-document Structure Theory (CST) [5], a multidocument theory/model that intends to represent the relatedness of a group of texts on the same topic. Our operations are represented as template-like operators composed of rules that use information processing primitive functions and correlate CST relations with user summarization preferences. This paper builds on previous work in the area [1][2][4][5].

2 Content Selection Operators: Definition and Formalization

The general idea of this work is that, given a group of texts (annotated with CST) and some summarization preference (from the user), the content selection operators are applied in order to indicate the best information units to be part of the final summary. The existence of more than one operator mirrors the diverse interests that users may have: while some of them may be interested only in the main information, some may want to read general background information about the subject, while others might want to visualize contradictions among the information sources.

Formally, by content selection operators we mean computational artifacts that are able to process a content representation and to produce a condensed version of it. The operators we define take as input a "raw" initial rank of information units derived from the CST-annotated texts (which take the form of a graph – the CST graph) and

* This work was supported by FAPESP and CNPq.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 25–29, 2010.

produces a refined rank, whose best ranked units (the "preferred ones" according to the specified user preference) should be selected for the final summary (respecting a specified compression rate). The CST relations from the CST graph are also included in the ranks, so that we do not need to keep consulting the graph for retrieving them.

The initial rank is simply built by taking all units from the CST graph and ordering them by the number of relations that they present. This process incorporates the underlying MDS assumption that the most important units are often repeated and elaborated in the several sources under consideration. Such units are usually highly connected with other ones by CST relations. Figure 1 shows a small hypothetical CST graph and the initial rank.

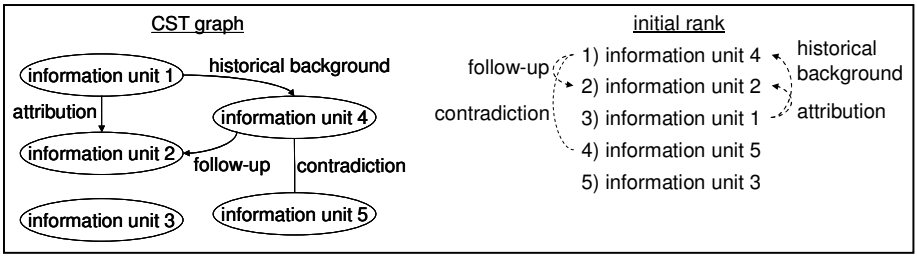

Fig. 1. Example of CST graph and initial rank

For the moment, when some units present the same number of relations, they are ranked in the order they are read from the graph. In this work, we consider sentences as our information units.

Our CST-based content selection operators are defined in a template-like format containing rules. The rules are specified as conditions and restrictions that must be satisfied for triggering actions, which are defined in terms of information processing primitive functions. Each rule is defined as follows:

$$\text{CONDITIONS, RESTRICTIONS} \Rightarrow \text{ACTIONS}$$

Each condition assumes the form CONDITION(S_i, S_j, Directionality, Relation) and is satisfied if there is the specified relation with the corresponding directionality among two sentences S_i and S_j (from S_i to S_j: $\rightarrow$; the opposite way: $\leftarrow$; or no directionality: —). Restrictions are optional and may represent extra necessary requirements for the operator to be applied. If all the conditions and restrictions are satisfied, the actions are applied to the initial rank, producing its refined version. Actions are defined in terms of at least one of the three primitive functions below:

- MOVE_UP(S_i,S_j): re-ranks sentence j, putting it at a higher position in the rank, immediately after sentence i;
- SWITCH(S_i,S_j): switches the position of the sentences i and j in the rank;
- ELIMINATE(S_i): eliminates sentence i from the rank.

For this work, we defined and formalized 5 operators that represent possible user summarization preferences, namely, contextual information presentation, overview of evolving events, contradiction exhibition, authorship identification, and redundancy

treatment. We also see the process of building the initial rank as an operator, which gives preference for the main information. We refer to it by generic (or main information) operator. Each operator is defined by three fields: a reference name, its description, and the rules, which consist in its most important part. The first operator, the one for contextual information presentation, is shown in Figure 2. It looks for pairs of sentences (in any positions in the rank) that present the elaboration or the historical background CST relations (since these are the relations that give such contextual information) and move up the appropriate sentence in the rank. Therefore, the contextual information gets higher preference for being in the summary.

Name	Contextual information presentation
Description	Preference for historical and complementary information
Rules	CONDITION(S_i, S_j, ←, Elaboration) ⇒ MOVE_UP (S_i,S_j) CONDITION(S_i, S_j, ←, Historical background) ⇒ MOVE_UP(S_i,S_j)

Fig. 2. Contextual information presentation operator

The application of this operator to the initial rank of Figure 1 would produce the refined rank in Figure 3 (the initial rank is also shown for comparison). One may see that the historical information (information unit 1) move up in the rank, going right below the sentence to which it refers (information unit 4).

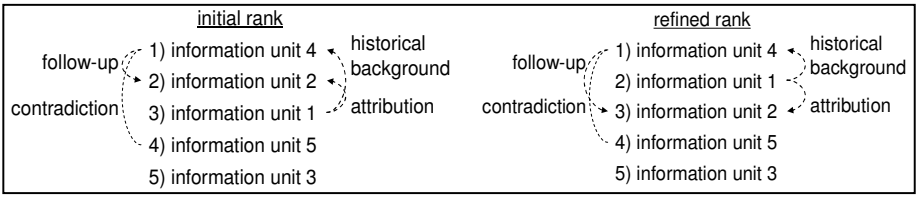

Fig. 3. Refined rank after application of the contextual information presentation operation

The next operator (Figure 4) is the one for giving preference for viewing the evolution of events in time. Such evolution is modeled in CST by the historical background and follow-up relations. Notice that, as the directionality does not matter, each rule appears twice, changing only the direction of the relation.

Name	Evolving events overview
Description	Preference for information about the evolution of events in time
Rules	CONDITION(S_i, S_j, ←, Historical background) ⇒ MOVE_UP(S_i,S_j) CONDITION(S_i, S_j, →, Historical background) ⇒ MOVE_UP(S_i,S_j) CONDITION(S_i, S_j, ←, Follow-up) ⇒ MOVE_UP(S_i,S_j) CONDITION(S_i, S_j, →, Follow-up) ⇒ MOVE_UP(S_i,S_j)

Fig. 4. Evolving events overview operator

Figure 5 shows the operator for exhibiting contradictions (which are expressed by the contradiction relation), while Figure 6 shows the operator for authorship identification (which is expressed by the attribution and citation relations).

Name	Contradiction exhibition
Description	Preference for contradictory information
Rules	CONDITION(S_i, S_j, —, Contradiction) ⇒ MOVE_UP (S_i,S_j)

Fig. 5. Contradiction exhibition operator

Note that the rules in the authorship identification operator have more than one condition. All of them must be satisfied in order to the corresponding rule to be applied. This happens because the attribution and citation relations usually come with some other relation.

Name	Authorship identification
Description	Preference for authorship information
Rules	CONDITION(S_i, S_j, ←, Attribution), CONDITION(S_i, S_j, ←, Subsumption) ⇒ SWITCH(S_i,S_j), ELIMINATE(S_i) CONDITION(S_i, S_j, ←, Citation), CONDITION(S_i, S_j, ←, Subsumption) ⇒ SWITCH(S_i,S_j), ELIMINATE(S_i)

Fig. 6. Authorship identification operator

Finally, the redundancy treatment operator is shown in Figure 7. Besides the conditions, it also presents restrictions on the length of the sentences.

Name	Redundancy treatment
Description	Preference for non-redundant information
Rules	CONDITION(S_i, S_j, —, Identity) ⇒ ELIMINATE(S_j) CONDITION(S_i, S_j, —, Equivalence), $\|S_i\| \leq \|S_j\|$ ⇒ ELIMINATE(S_j) CONDITION(S_i, S_j, —, Equivalence), $\|S_i\| > \|S_j\|$ ⇒ SWITCH(S_i,S_j), ELIMINATE(S_i) CONDITION(S_i, S_j, ←, Subsumption) ⇒ SWITCH(S_i,S_j), ELIMINATE(S_i) CONDITION(S_i, S_j, →, Subsumption) ⇒ ELIMINATE(S_j)

Fig. 7. Redundancy treatment operator

It is important to say that there is another important CST relation that we are not dealing with at this moment: the overlap relation, which specifies that the sentences share some content, but also have unique parts. The overlap relation should be considered in the redundancy treatment and in the authorship identification operators at the cost of having some available automatic sentence fusion module. For instance, in the redundancy treatment operator, if there is an overlap relation among the sentences S_i and S_j, another sentence S_k should be built by fusing S_i and S_j.

An MDS system should be able to read the user preferences, select the appropriate operators, and produce the corresponding summary from a group of texts.

References

1. Aleixo, P., Pardo, T.A.S.: Finding Related Sentences in Multiple Documents for Multi-document Discourse Parsing of Brazilian Portuguese Texts. In: Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL, pp. 298–303 (2008)
2. Jorge, M.L.C., Pardo, T.A.S.: Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory. In: The Proceedings of the VII Brazilian Symposium in Information and Human Language Technology (2009)
3. Mani, I.: Automatic Summarization. John Benjamins Publishing Co., Amsterdam (2001)
4. Radev, D.R., McKeown, K.: Generating natural language summaries from multiple on-line sources. Computational Linguistics 24(3), 469–500 (1998)
5. Radev, D.R.: A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: The Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue (2000)

Translating from Complex to Simplified Sentences

Lucia Specia

Research Group in Computation Linguistics,
University of Wolverhampton,
Wolverhmapton, UK
L.Specia@wlv.ac.uk

Abstract. We address the problem of simplifying Portuguese texts at the sentence level by treating it as a "translation task". We use the Statistical Machine Translation (SMT) framework to learn how to translate from complex to simplified sentences. Given a parallel corpus of original and simplified texts, aligned at the sentence level, we train a standard SMT system and evaluate the "translations" produced using both standard SMT metrics like BLEU and manual inspection. Results are promising according to both evaluations, showing that while the model is usually overcautious in producing simplifications, the overall quality of the sentences is not degraded and certain types of simplification operations, mainly lexical, are appropriately captured.

Keywords: Text Simplification, Statistical Machine Translation, Portuguese.

1 Introduction and Background

The ability to use written language to obtain and record information, express themselves, plan and learn continuously is called *literacy*, or "letramento", in Brazil. According to the index used to measure the literacy level of the population in Brazil (*INAF - National Indicator of Functional Literacy*), a vast number of people belong to the so called *rudimentary* and *basic* literacy levels [1]. These people are able to find explicit information in short texts (rudimentary level) and also process slightly longer texts and make simple inferences (basic level).

The PorSimples project[1] aims at producing tools to simplify complex texts and therefore promote digital inclusion and accessibility for people with such levels of literacy, and possibly other kinds of reading disabilities. Amongst the tools envisaged is a browser plugin that automatically simplifies texts on the web for the end-user. The focus is on texts published in government sites or by relevant news agencies, both expected to be of importance to a large audience with various literacy levels. The language of the texts is Brazilian Portuguese, for which there are no text simplification systems, to the best of our knowledge.

Text Simplification (TS) has been exploited in other languages for helping poor literacy readers [2], [3] and [4] and people with disabilities such as aphasia [5]. It has also been used for improving the accuracy of other natural language processing tasks

[1] http://caravelas.icmc.usp.br/wiki/index.php/English

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 30–39, 2010.

[6, 7]. Most approaches focus on developing rule-based systems based on parser structures and limited to certain simplification operations.

Corpus-based systems, which can learn the simplification operations, have also been exploited through the use of parallel aligned corpora of original and simplified texts. [8] investigates the induction of syntactic rules from a corpus by extracting syntactic correspondences and generalizing them. [9] uses corpora of TV program transcripts and subtitles to generate subtitles for hearing-impaired people, focusing on summarization and lexical substitution. [10] uses machine learning techniques to aid second language teachers by proposing when to split or drop sentences.

In the scope of the PorSimples project, one of the main text processing strategies to produce simplified texts is to identify and apply lexical and syntactic simplification operations. A rule-based system based on a pre-defined set of operations has been implemented [11] and a machine learning algorithm has been used to support the decision of whether or not to simplify a sentence [12]. Both systems rely on the set of pre-defined simplification operations. These operations, described in Section 2, were carefully designed based on a study of the Portuguese grammar. Their general definition cover all phenomena believed to cause sentences to be difficult to read. However, the specific implementation of such operations as rules is not a trivial task. The rules are defined in terms of syntactic representations and lexical clues, and it is impossible to foresee all variations of the same phenomena with different realizations. Generalizing the rules while making sure they do not cover illicit phenomena is therefore a complex issue. Moreover, the dependency on a syntactic parser, which is not error-free, might incur inadequate uses of the rules. Finally, a very important limitation is the fact that this system can only cover the phenomena that fall into the pre-defined set of operations.

All the existing corpus-based approaches aim at the simplification of English texts, mostly to learn when to apply operations or to induce the actual simplification rules based on syntactic representations. In this paper we proposed an alternative strategy to exploit parallel text with no additional annotation. We learn how to translate from complex to simplified Portuguese texts using the Statistical Machine Translation (SMT) framework, or more specifically, the noisy-channel model as implemented in a standard phrase-based SMT system. The SMT framework has been used for other tasks than multilingual translation, including recasing (http://www.statmt.org/moses/) and other post-edition for problems [13], but not for Text Simplification.

In what follows, we first describe the experimental setting, including the corpus of original and simplified texts and the SMT framework (Section 2) to then present our experiments and results (Section 3).

2 Experimental Setting

2.1 The Parallel Corpus of Original and Simplified Texts

Once specified the set of syntactic and lexical simplification operations to be addressed by the PorSimples project, one of the main tasks was the manual creation, by an expert in TS, native speaker of Portuguese, of a parallel corpus of original and simplified texts [14]. The pre-defined specifications aim to simplify texts as much as

possible, targeting mainly the rudimentary level readers. However, this sometimes leads to oversimplification of the text, making it unnatural, tiresome or even disturbing for readers with basic (or higher) literacy level. Therefore, the Project also considers the case of performing simplification only if the resulting text still looks "natural". These two levels of simplification are denominated "strong" and "natural", respectively, and were taken into account during the creation of the parallel corpus. In this paper, we use the corpus of natural simplifications to train the "translation" system. This type of simplification is more difficult to be captured by rules, as it does not follow rigid constraints like those imposed for strong simplifications.

The pre-defined simplification operations are the following: (1) non-simplification; (2) lexical substitution to replace unusual or complex words; (3) simple rewriting (to replace discourse markers or sets of words, like idioms or collocations) (4) putting the sentence in its subject-verb-object order; (5) putting the sentence in the active voice; (6) inverting the clause ordering; (7) splitting or (8) joining sentences; (9) dropping the sentence or (10) dropping parts of the sentence. Besides the 10 regular operations, during the annotation process, the annotator could choose (11): strong rewriting, for any sort of free rewriting of sentence that was not covered by the other operations, as defined in [10].

The corpus was extracted from two of the main Brazilian newspapers: first page news articles from *Zero Hora* and science articles from *Folha de Sao Paulo*. The human annotator used a tool to perform the simplifications, which recorded, among other information, the alignment between original and simplified sentences and the type of operation used [14]. Table 1 shows examples of original sentences in (A) and (C), and their natural simplifications in (B) and (D). The operations applied to (A) to make (B) were regular ones (splitting the sentence and replacing unusual words). (D), on the other hand, was produced from (C) using "strong rewriting" for natural modifications not covered by the pre-defined set.

Table 1. An example of an original text (A) and its (natural) simplified version (B)

A	*A cerração, que ocultou o amanhecer dos porto-alegrenses, voltou a comprometer as operações do aeroporto por mais de oito horas, entre 1h50 min e 9h56 min.*
B	*A neblina comprometeu outra vez as operações do aeroporto por mais de oito horas. A neblina comprometeu as operações do aeroporto entre 1h50 min e 9h56 min. A neblina ocultou o amanhecer dos porto-alegrenses.*
C	*Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea.*
D	*Cientistas britânicos detectaram em adultos que células-tronco da medula óssea produziram células do fígado.*

The resulting parallel corpus is composed of 104 news articles from *Zero Hora* and 37 from *Caderno da Ciencia*, amounting to 4,483 original sentences and their corresponding simplifications. Multiple sentences produced from a single complex sentence were put together and aligned to the original sentence. We call them *segments*. The distribution of the operations used in the manual simplification process of the original sentences is depicted in Table 2. A detailed corpus study is presented in [14].

Table 2. Statistics on the types of simplification in the parallel corpus

Syntactic and Lexical Operations	Number of (original) sentences
Non-simplification	595
Lexical Substitution	1487
Simple rewriting	732
Subject-verb-object ordering	41
Transformation to active voice	100
Inversion of clause ordering	256
Splitting sentences	823
Joining sentences	12
Dropping one sentence	0
Dropping sentence parts	424
Strong rewriting	13

2.2 SMT Framework and Moses

The standard framework for Statistical Machine Translation (SMT) is based on the noisy-channel model from information theory. The translation of a string *f* in a foreign language into a string *e* in a native language is based on estimations of the probability *p(e|f)*. The Bayes Theorem is applied to decompose this problem into two others:

$$p(e \mid f) = \frac{p(f \mid e)\, p(e)}{p(f)} \tag{1}$$

where *p(f|e)*, called the *translation model*, is the probability that the foreign string is the translation of the native string, and *p(e)*, called the *language model* is the probability of seeing that native string. *p(f)* is a normalization constant and can be disregarded. Therefore finding the best translation ê becomes the problem of finding the translation with the highest probability:

$$\hat{e} = \arg\max_{e \in e^*} p(e \mid f) = \arg\max_{e \in e^*} p(f \mid e)\, p(e) \tag{2}$$

The basic assumption of this paper is that if SMT works well for translating between two different languages, it could also succeed in translating from complex to simplified texts in the same language. The goal is to learn a probabilistic dictionary of phrases and their corresponding simplified versions (*translation model*) along with a model that indicates how likely it is that the segment is a good simplified segment in Portuguese (*language model*). Besides, it is necessary to estimate weights for these and additional models governing, for example, how to distribute such phrases in a simplified segment (*reordering model*), the penalty for source and translation with different lengths (*word penalty*), etc. Each candidate translation *t* for a given source segment is scored according to a linear combination of all these models (*feature functions* f_j) and their weights λ_j learned during a tuning phase:

$$score(t) = \sum_j \lambda_j f_j(t) \tag{3}$$

We use Moses, a standard SMT system [15] considered by many as the state of the art in phrase-based SMT. Moses is a generalization of the noisy channel, where the following feature functions are considered to cover the language and translation models[2]:

- phrase translation model: 5 features covering conditional forward and inverse phrase and lexical translation probabilities, and phrase penalty;
- target language model;
- distortion limit: a limit on the amount of reordering allowed;
- word penalty: to penalize translations that differ in length as compared to the source segment;
- lexicalized reordering model: 6 features to cover reordering of lexical items.

The weights of these features are estimated using *Minimum Error Rate Training* [16] based on some tuning data (source and reference translation). We use configuration of the system suggested in the series of shared translation tasks where Moses is used as the *baseline system*[3]. This is the best setting for most language-pairs and domains.

3 Experiments and Results

For the set of experiments presented here we put together all the segments in Table 2. With that, we are targeting a very robust scenario in which given a certain sentence and no further information about it, the model should decide whether it needs to be simplified and, if so, perform the necessary simplifications. In that scenario, sentences that do not need simplification should simply be copied to the output.

From our corpus of 4,483 segment pairs, we randomly selected:

- 3,383 segments for *training* (i.e., for generating the dictionary of original-simplified phrases);
- 500 segments for *tuning* the parameters in Moses; and
- 500 segments for *test* (evaluation).

3.1 Training

The phrase-table produced from training contains 229,000 phrases. The default configuration in Moses builds phrases with up to 7 tokens. While many of them simply have the same source and target phrases with high probability scores, others cover adequately many types of lexical simplification and simple rewriting. Some interesting examples are shown in Table 3, together with their main phrase probability feature *p(fle)*, which is given the highest weight in the tuning phase (Section 3.2).

It is important to notice that many spurious phrases are also produced, such as the three last cases in Table 3. However, they either have low probabilities and are therefore ruled-out from the candidate translation set by the translation model, or are ruled-out by the other models (mainly word-penalty and language model).

[2] http://www.statmt.org/moses/?n=Moses.FeatureFunctions
[3] http://www.statmt.org/wmt09/baseline.html

Table 3. Examples of phrased in our phrase table and their probabilities

Source phrase	Target phrase	*p(fle)*
para	com o objetivo de	0.8
para	com o objetivo	0.709677
para a melhor compreensão da	com o objetivo de compreender melhor a	1
, quando	, momento em que	0.625
, quando	, no momento em que	0.5
, quando	, no momento	1
-- não gostamos quando	-- não gostamos se	1
segundo	de acordo com	0.266667
dias , segundo	dias de acordo com	1
“ (editora maranta , são paulo ,	rambelli	0.004
negocio ainda precisa ser	mas	0.0023
de acordo com	a imprensa reproduzisse	0.5

3.2 Tuning

By inspecting the weights estimated for each feature from the tuning data, we found that the features with highest individual weight were word penalty (negative weight) and language model (positive weight). Absolute values for the weights of phrase probability and some lexicalized reordering features are also significantly higher than the remaining features. This is an indication that our model prioritizes simplifications which are not very different in length and order as compared to the original segment, and which are also common segments in the target (simplified) language.

3.3 Test

After building the phrase-table and tuning Moses, we decoded the 500 test segments and computed the two most common MT evaluation metrics: BLEU [17] (1 to 4-grams) and NIST [18] (1 to 5-grams). Both BLEU and NIST check the overlapping of n-grams between two sets of segments, in this case, human and machine simplification. There are no standard automatic metrics like these for Text Simplification.

Let *source* be a given original sentence, *target* the system simplification and *reference* the human simplification for that source sentence. The scores obtained for the target against the reference are **BLEU = 0.6075** and **NIST = 9.6244**. The upper bound for BLEU and NIST with this corpus would be **1** and **~13**, respectively. Although it is impossible to compare such scores with those reported in standard translation tasks, a BLEU of ~0.60 in a translation task usually means good translation quality, particularly for News texts (the shared evaluation task reports scores of around 0.25 for translation from English into several languages) [19].

Using these evaluation metrics and also checking string matching (ignoring case, spaces and punctuations), we performed three other tests:

1) **Amount of simplification performed:** we check whether the target segments differ significantly from the source sentences. The higher the BLEU/NIST scores between these two versions, the closer the target segments are to the source. The scores obtained are **BLEU = 0.7664** and **NIST = 11.509**. This shows that the simplifications produced by the system are

closer to the source than to the reference segments, which is an indication that the system is overcautious in performing simplifications. In other words, in terms of standard metrics recall and precision, the system favors precision. By measuring string matching, we found that **374 (75%)** segments in the target differ from the source, the remaining were not modified.

2) **Contribution of the simplifications performed:** we check whether the reference segments match better the target or source segments. The scores comparing source and reference segments are **BLEU = 0.5726** and **NIST = 9.5101.** This shows that target segments are closer to the reference (**BLEU = 0.6075** and **NIST = 9.6244**), which proves that the simplifications being performed are likely to be correct, since they approximate the segments to the gold-standard ones. However, exact matching is rare: **452 (90%)** segments in the target differ from the reference.
3) **Preservation of originally simple sentences:** we compare the string matching between source and reference segments, finding that that **443 (89%)** source sentences were modified somehow by the human annotator, and should, in principle, have also been simplified by our model. Out of these 443 segments, the SMT model only modified **350 (79%)**, which shows again that the system was overcautious. If we recall that in total the model modified 374 segments (item (1)), we see that only **24 segments** that were not simplified by the annotator were modified by our model.

Although this inspection gives some intuition on the performance of the model, automatic metrics like BLEU/NIST and string matching are known to be very limited even for standard translation evaluation tasks. We therefore also performed manual inspection of a subset of segments. We randomly selected 20 segments out of the 500-test set which suffered some modification and checked whether the simplified versions produced were fluent and adequate, and whether the expected simplifications were performed. These aspects were judged in a scale of **1 (worse)** to **3 (best)**. For simplification quality, in particular, the 1-3 scores were given according to the following criteria: 1 = none of the expected simplifications was performed or at least one inadequate simplification was performed; 2 = some of the expected simplifications were performed, but not all of them; and 3 = all of the expected and possibly additional adequate simplifications were performed. The average scores obtained for the 20 segments are: **Fluency = 2.5**, **Adequacy = 2.5**; **Simplification = 2.35**. Some examples of original segments (source), their simplified version according to our model (target) and the human annotator (reference) are shown in Table 4.

According to the manual inspection, as we can see by the examples in Table 4, the lexical operations (complex words and discourse markers) are in general well captured. Some short-distance reordering is also correctly performed, mostly for adverbs. Simple cases of sentence splitting are performed correctly, such as that of clauses separated by conjunctions. Long distance operations like canonical (S-V-O) ordering and clause inversion in general are not performed. The most problematic cases are very long sentences, particularly those containing brackets: the model usually tries to remove the brackets, but cannot always deal with their content adequately. Some of the deficiencies are due to the nature of the approach, since only information about words and short phrases is used.

Table 4. Examples of source, simplified and reference segments

Source	Target	Reference
no dia da liberdade de impostos, criado pela associação da classe média (aclame) *para marcar a* data em que os brasileiros, simbolicamente, deixam de trabalhar para pagar tributos, *foram vendidos* em novo hamburgo gasolina, computadores e um carro.	no dia da liberdade de impostos, criado a associação da classe média (aclame), *com o objetivo de marcar* a data em que os brasileiros, simbolicamente, deixam de trabalhar para pagar tributos. *em novo hamburgo, nesse dia, venderam-se* gasolina, computadores e um carro.	a associação da classe média (aclame) criou o dia da liberdade de impostos. *o objetivo é marcar* a data em que os brasileiros, simbolicamente, deixam de trabalhar para pagar tributos. *em novo hamburgo, nesse dia, venderam-se* gasolina, computadores e um carro.
segundo o presidente da comissão de finanças da assembléia, ricardo sanguino, o processo ...	*de acordo com* o presidente da comissão de finanças da assembléia, ricardo sanguino, o processo ...	ricardo sanguino é o presidente da comissão de finanças da assembléia. *ele disse que* o processo ...
o anúncio do resultado ocorre em um megaevento *em lisboa, portugal, amanhã à noite*.	o anúncio do resultado ocorre em um megaevento *em lisboa, amanhã à noite em portugal* .	o anúncio do resultado ocorre *amanhã à noite em um grande evento em lisboa. lisboa fica em portugal* .
segundo bush, as mudanças em anbar são um exemplo ...	*na opinião de* bush, as mudanças em anbar são um exemplo ...	*na opinião de bush*, as mudanças em anbar são um exemplo ...
o animal deu alguns passos e começou a *verter sangue pela boca.*	o animal deu alguns passos e começou a *verter sangue pelo estabelecimento.*	o animal deu alguns passos e começou a *despejar sangue pela boca.*
-- eu sei que é *imprudência, mas* só dirijo na praia, *só para* aprender.	-- eu sei que é *descuido. mas* só dirijo na praia*, para* aprender.	-- eu sei que é *descuido. mas* só dirijo na praia*, para* aprender.
o relatório *produzido pela* comissão criada para revisar processos administrativos ...	o relatório *da comissão* criada para revisar processos administrativos ...	o relatório *da comissão* criada para examinar os processos administrativos ...
ontem, a moeda norte-americana fechou cotada a r 2,026*, queda* de 3,20%	*ontem*, a moeda americana fechou cotada a r 2,026*. a queda foi* de 3,20%	a moeda norte-americano fechou a r 2,026 *ontem. a queda foi* de 3,20%
tombou primeiro sob as duas patas da frente *e então* caiu de vez .	tombou primeiro sob as duas patas da frente. *em seguida* caiu de vez.	ele tombou primeiro sob as duas patas da frente*. em seguida,* caiu de vez .
mesmo que a incerteza sobre o tamanho da crise *dificulte* projeções de prazo mais longo*, seus* efeitos abrem janelas para ...	*a incerteza* sobre o tamanho da crise *dificulte* projeções de prazo mais longo*. mas seus* efeitos abrem janelas ...	*a incerteza* sobre o tamanho da crise *torna difíceis* as projeções de prazo mais longo*. mas seus* efeitos abrem janelas para ...

4 Conclusions and Future Work

We have presented a new approach for Text Simplification, based on the framework of Statistical Machine Translation. The experiments have shown that the framework

can appropriately simplify certain phenomena, particularly those related to lexical operations: lexical simplification and simple rewriting. As for the other operations, the system is usually overcautious and does not perform simplifications. While this yields low recall, it guarantees precision, by preventing complex operations to be applied when they would most likely degrade the results.

These experimental results in general are very promising, particularly considering the size of the parallel corpus and the fact that a basic SMT system was used as such, without any adaptation to the simplification task. It is well-known that SMT requires considerably more data than the amount used in this paper. For training standard systems to translate news texts like ours, it is recommended a few hundred thousands sentence pairs, but we only used ~4,000 sentence pairs. While translating between variations of the same language alleviates the need for such large corpora, the fact that several types of simplification were considered here makes the task considerably harder. Finally, complex operations like clause reordering depend on deeper information than words and phrases. Certain additional features, like part-of-speech tags, can already be used as *factors* in standard phrase-based systems like Moses, although more training data would be necessary. The use of other information sources, particularly syntax, is also possible with frameworks for hierarchical SMT [20].

We believe that having a larger corpus of appropriately annotated original-simplified texts, covering enough examples of the several types of simplifications, is very likely to yield a larger number of "simple" simplifications without degrading the quality of the outcome segments. With a larger parallel corpus, one could build one "translation" model for each type of simplification and then apply all models to a given complex input text, either in sequence or in combination.

As future work, besides building larger corpora, we plan to investigate how the standard SMT framework could be adapted to further improve the results. We also consider adding syntactic information to the framework in order to better address long-distance operations, like clause inversion. Finally, we would like to evaluate whether the simplifications produced here are useful for a given target user by testing them within a certain task, such as answering a questionnaire or solving some problem.

References

1. Ribeiro, V.M.: Analfabetismo e alfabetismo funcional no Brasil. In: Boletim INAF. Instituto Paulo Montenegro, São Paulo (2006)
2. Max, A.: Writing for Language-impaired Readers. In: Proceedings of 7th Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, pp. 567–570 (2006)
3. Petersen, S.E.: Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education. PhD thesis, University of Washington (2007)
4. Siddharthan, A.: Syntactic Simplification and Text Cohesion. PhD thesis, University of Cambridge (2003)
5. Devlin, S., Unthank, G.: Helping aphasic people process online information. In: Proceedings of the ACM Conference on Computers and Accessibility, Portland, Oregon, pp. 225–226 (2006)

6. Klebanov, B., Knight, K., Marcu, D.: Text Simplification for Information-Seeking Applications. In: Meersman, R., Tari, Z. (eds.) OTM 2004. LNCS, vol. 3290, pp. 735–747. Springer, Heidelberg (2004)
7. Vickrey, D., Koller, D.: Sentence Simplification for Semantic Role Labeling. In: Proceedings of the ACL-HLT, pp. 344–352 (2008)
8. Chandrasekar, R., Srinivas, B.: Automatic Induction of Rules for Text Simplification. Knowledge-Based Systems 10, 183–190 (1997)
9. Daelemans, W., Hothker, A., Sang, E.T.K.: Automatic Sentence Simplification for Subtitling in Dutch and English. In: Proceedings of the 4th Conference on Language Resources and Evaluation, Lisbon, Portugal, pp. 1045–1048 (2004)
10. Petersen, S.E., Ostendorf, M.: Text Simplification for Language Learners: A Corpus Analysis. In: Proceedings of the Speech and Language Technology for Education Workshop, Pennsylvania, USA, pp. 69–72 (2007)
11. Candido Jr., A., Maziero, E., Gasperin, C., Pardo, T.A.S., Specia, L., Aluisio, S.M.: Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In: Proceedings of the NAACL/HLT Workshop on Innovative Use of NLP for Building Educational Applications, Boulder, Colorado, pp. 34–42 (2009)
12. Gasperin, C., Specia, L., Pereira, T., Aluisio, S.M.: Learning When to Simplify Sentences for Natural Text Simplification. In: Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA), Bento Gonçalves, Brazil, pp. 809–818 (2009)
13. Simard, W., Goutte, C., Isabelle, P.: Statistical Phrase-based Post-editing. In: Proceedings of NAACL HLT, Rochester, USA, pp. 508–515 (2007)
14. Caseli, H.M., Pereira, T.F., Specia, L., Pardo, T.A.S., Gasperin, C., Aluísio, S.M.: Building a Brazilian Portuguese parallel corpus of original and simplified texts. In: 10th Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, pp. 59–70 (2009)
15. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, C., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th ACL, demonstration session, Prague, Czech Republic (2007)
16. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st ACL, Sapporo, Japan, pp. 160–167 (2003)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th ACL, Morristown, pp. 311–318 (2002)
18. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the 2nd Conference on Human Language Technology Research, San Diego, pp. 138–145 (2002)
19. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Proceedings of the 4th Workshop on Statistical Machine Translation, Athens, Greece, pp. 1–28 (2009)
20. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd ACL, Ann Arbor, USA, pp. 263–270 (2005)

Challenging Choices for Text Simplification

Caroline Gasperin, Erick Maziero, and Sandra M. Aluísio

NILC - Núcleo Interinstitucional de Linguística Computacional
ICMC, Universidade de São Paulo, São Carlos, SP, P.O. Box 668, 13560-970, Brazil
{cgasperin,sandra}@icmc.usp.br, erickgm@grad.icmc.usp.br

Abstract. In this paper we discuss particular choices we made during the development of a rule-based syntactic text simplification system. Such choices concern 1) how to deal with adverbial phrases in order to simplify sentences, and 2) the order in which to apply our set of simplification rules. Adverbial phrases have not been considered by previous work on text simplification, but have a considerable impact on the complexity of a sentence. Considering our whole set of simplification rules, we discuss and compare two different orders in which to apply them: empirical and hierarchical.

1 Introduction

In Brazil, according to the index used to measure the literacy level of the population (INAF - National Indicator of Functional Literacy), a vast number of people belong to the so called rudimentary and basic literacy levels. These people are only able to find explicit information in short texts (rudimentary level) or process slightly longer texts and make simple inferences (basic level). INAF [1] reports that 7% of the individuals were classified as illiterate; 25% as literate at the rudimentary level; 40% as literate at the basic level; and only 28% as literate at the advanced level.

The PorSimples project (*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*[1]) aims at producing Text Simplification (TS) tools for promoting digital inclusion and accessibility for people with such levels of literacy, and possibly other kinds of reading disabilities. More specifically, the goal is to help these readers to process documents available on the web. The focus is on texts published in government sites or by relevant news agencies, both expected to be of importance to a large audience with various literacy levels. The language of the texts is Brazilian Portuguese, for which there are no text simplification systems, to the best of our knowledge.

TS aims to maximize the comprehension of written texts through the simplification of their linguistic structure. This may involve simplifying lexical and syntactic phenomena, by substituting words that are only understood by a few people with words that are more usual, and by breaking down and changing the syntactic structure of the sentence.

[1] http://caravelas.icmc.usp.br/wiki/index.php/Principal

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 40–50, 2010.

TS has been exploited in other languages for helping poor literacy readers [2], bilingual readers [3] and special kinds of readers such as aphasics [4] and deaf people [5]. It has also been used for improving the accuracy of other natural language processing tasks [6,7], like parsing and information extraction.

Within the PorSimples project we are developing a rule-based TS system [8], which aims to simplify complex syntactic constructions such as apposition, relative clauses, coordination and subordination. These have been the syntactic phenomena addressed by previous work on text simplification [9,2]. Moreover, we have built rules to transform to active voice a sentence in passive voice and to transform to Subject-Verb-Object (SVO) order a sentence in non-SVO order, mainly those stating with a verb phrase.

In this paper we address two issues that complement the work we have already done. Firstly, we present our study on developing a rule for simplification of adverbial phrases, which have a considerable impact on the complexity of sentences and which have not been dealt with in the past. Secondly, we discuss the order in which the simplification rules should be applied: we present two alternative orders, empirical and hierarchical, and compare the results of the system for each of them.

In Section 2 we describe briefly some related work and present our text simplification system. In Section 3 we discuss our strategy for simplifying adverbial phrases. In Section 4 we discuss the issue of defining an order of application of simplification rules. In Section 5 we describe our reference corpus and present the results of the evaluation of our system.

2 Rule-Based Syntactic Text Simplification

2.1 Related Work

A few rule-based systems have been developed for text simplification [9,2], focusing on different readers (poor literate, aphasic, etc). These systems contain a set of manually created simplification rules that are applied to each sentence. These are usually based on parser structures and limited to certain simplification operations. Siddharthan's approach uses a three-stage pipelined architecture for syntactic text simplification: analysis, transformation and regeneration. In his approach, POS-tagging and noun chunking are used in the analysis stage, followed by pattern-matching for known templates which can be simplified. The transformation stage applies seven handcrafted rules for simplifying conjoined clauses, relative clauses and appositives. The regeneration stage fixes mistakes introduced by the previous phase by generating referring expressions, selecting determiners, and generally preserving discourse structure with the goal of improving cohesion of the resulting text.

Inui et al. [5] proposes a rule-based system for text simplification aimed at deaf people. The authors create readability assessments based on questionnaires answered by teachers about the deaf. With approximately one thousand manually created rules, the authors generate several paraphrases for each sentence and train a classifier to select the simpler ones. Promising results are obtained,

although different types of errors on the paraphrase generation are encountered, such as problems with verb conjugation and regency.

2.2 Our Work

We have followed Siddharthan's approach, in the sense that we organize our system in phases: 1) identification of syntactic phenomena and 2) application of simplification rules. We are not working on the regeneration phase yet.

For phase 1 we have appointed a set of syntactic phenomena that are considered complex and require simplification. For phase 2 we have created a set of simplification rules which assign simplification operations to each phenomenon. Table 1 summarises the syntactic phenomena that we tackle and the simplification rules for each phenomenon considering phases 1 and 2. We use surface information, part-of-speech and syntactic clues provided by a parser for Portuguese [10] to detect the phenomena (these sources of information also assist in the process of simplification). When no phenomenon is detected, the sentence is not simplified. We describe in more detail the phenomena we treat and the operations assigned to them in [8].

In this paper we give particular attention to the last item in Table 1, Adverbial Phrases. Previous studies have not dealt with adverbial phrases, but we consider it is important to take them into account when simplifying sentences.

The order in which to apply the rules described in Table 1 is also a focus of this paper. It can minimise parsing errors and thus contribute to the quality of the resulting simplified sentences.

3 Simplification of Adverbial Phrases

A sentence that does not contain any of the syntactic phenomena from 1 to 21 (Table 1) but which contains long adverbial phrases can be complex to read, mainly if these phrases are in the beginning or in the middle of the sentence. Consider the following example sentence:

Pela decisão do STF, nos casos de mudança de partido depois de 27 de março, as legendas terão de encaminhar à corte eleitoral um pedido de investigação.

In this sentence there are two adverbial sentences before the subject of the sentence. The reader has to go through them before he/she reaches the main part of the sentence. That can severely disturb the comprehension process for a non-fluent reader. Because of that we decided to include adverbial phrases into the simplification process. According to Temperley [11], constructions in which the distance between the head of the adverbial phrase and the verb on which it is dependent is too long should be avoided to minimize the dependency length. Temperley finds statistical evidence for Gibson's Dependency Locality Theory, which proposes that the complexity of a sentence is related to the length of its syntactic dependencies: longer dependencies are more difficult to process.

Table 1. Phenomena identification and simplification steps

#	Cases	Ph	Rules
1	Passive voice clauses	1	Identify passive voice by syntactic tags
		2	Transform passive into active voice
2	Apposition	1	Identify apposition by syntactic tags; Identify the anchor of the appositive clause
		2	Remove apposition from original sentence; Create new sentence containing anchor as subject, "é" (*is*) as verb, and the appositive clause as predicate
3	Asyndetic coord. clauses	1	Identify pairing of clauses by syntactic tags
		2	Split original sentence into one sentence per clause
4	Additives coord. clauses	1	Identify pairing of clauses by syntactic tags; Identify additive conjunctions by PoS tags
		2	Split original sentence into one sentence per clause
5	Adversative coord. clauses	1	Identify adversative conjunctions by token matching
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by "Mas" (*But*)
6	Correlated coord. clauses	1	Identify correlation conjunctions by token matching
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by "Também" (*Also*)
7	Result coord. clauses	1	Identify result conjunctions by token matching and PoS tags
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by "Como resultado" (*As a result*)
8	Explanatory coord. clauses	1	Identify explanation conjunctions by token matching and PoS tags
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by "Isto ocorre porque" (*This happens because*)
9	Causal sub. clauses	1	Identify causal conjunctions by token matching
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by "Com isso," (*Given this*)
10	Comparative sub. clauses	1	Identify comparison conjunctions by token matching
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the subordinate clause starting by "Também" (*Also*)
11	Concessive sub. clauses	1	Identify concession conjunctions by token matching
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by "Mas" (*But*)
12	Conditional sub. clauses	1	Identify condition conjunctions by token matching and PoS tags
		2	Rearrange clauses into subordinate/main order
13	Consecutive sub. clauses	1	Identify consecutive conjunctions by token matching
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the subordinate clause starting by "Assim," (*Thus*)
14	Purpose sub. clauses	1	Identify purpose conjunctions by token matching and PoS tags
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the subordinate clause starting by "O objetivo é" (*The purpose is*)
15	Conformative sub. clauses	1	Identify conformative conjunctions by token matching
		2	Rearrange clauses into subordinate/main order, linked by "confirma que" (*confirms that*), or "que" (*that*) when the subordinate clause contains a verb
16	Temporal sub. clauses	1	Identify temporal markers by token matching
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by "Então," (*Then*)
17	Proportional sub. clauses	1	Identify proportional conjunctions by token matching
		2	Do not simplify.
18	Non-finite sub. clauses	1	Identify non-finite clauses by syntactic tags
		2	Do not simplify.
19	Restrictive relative clauses	1	Identify relative clause by syntactic tags and relative pronoun; Identify the anchor of the appositive clause
		2	Remove relative clause from original sentence; Create new sentence containing anchor as subject and the relative clause as predicate
20	Non-restr. relative clauses	1	Identify relative clause by syntactic tags, relative pronoun and punctuation; Identify the anchor of the appositive clause
		2	Remove relative clause from original sentence; Create new sentence containing anchor as subject and the relative clause as predicate
21	No Subj-Verb-Obj order	1	Identify non-SVO order by syntactic tags: for now, we only look for sentences starting with a VP
		2	Transform into SVO order
22	Adverbial phrases	1	Identify adverbial phrases by syntactic tags; Identify if adverbial phrases are long (carry post-modification)
		2	Remove long adverbial phrases (one or more, if consecutive) from original sentence; Create a sentence for the adverbial phrases starting by "Isso ocorre" (*This happens*)

Simply moving the adverbial phrases to the end of the sentence (canonical position) would not entirely solve the problem, since sentences like the one in the above example would continue to be long. We have then adopted the following strategy: in the case of long adverbial phrases, we remove them from the original sentence and insert them in a new sentence started by "Isso ocorre" (the verb "ocorrer" is inflected to match the tense of the original sentence). We consider as long adverbial phrases those which contain prepositional or clausal post-modification attached to its head. If there are more than one consecutive adverbial phrase, they are inserted in the same new sentence (instead of creating a new sentence for each of them). According to this strategy, the simplified version of the above sentence is:

As legendas terão de encaminhar à corte eleitoral um pedido de investigação. Isso ocorrerá pela decisão do STF, nos casos de mudança de partido depois de 27 de março.

We have decided not to alter adverbial phrases when they are short (no post-modification), because in this case they do not disturb dependency lengths considerably. Examples like the following would not be simplified.

Ontem , Antunes disse que a quantidade de contratos sob gerenciamento de um mesmo departamento pode ser revista .

The evaluation of the adequacy of the new rule and of the performance of the system on executing it is presented in Section 5. The removal of the adverbial phrase from the original sentence can also help further simplifications, given that the quality of the output of the syntactic parser is likely to improve once the original sentences have been "cleaned up" of adverbial phrases. We discuss that in Section 5 as well.

4 Order of Application of Rules

Given a set of simplification rules, one has to decide in which order to apply them. If a sentence contains more that one complex syntactic phenomena, it is necessary to define which should be simplified first and so on. The order of application of the rules can minimise parser errors and so influences the quality of the final simplified sentence.

We present two alternative orders in which to apply the rules. The first order, which we call "empirical", was defined empirically based on the quality of the parser output for the single phenomenon involved: the more reliable the parser output, the further up in the sequence of simplification operations a rule is. The empirical order is:

1. Passive voice; 2. Apposition; 3. Subordinate clauses; 4. Non-restrictive relative clauses; 5. Restrictive relative clauses; 6. Coordinate clauses; 7. Non-SVO

The second order, which we call "hierarchical", follows the order in which the syntactic phenomena appear in the tree resulting from the syntactic analysis of the sentence. The closer to the root of the tree the phenomena is, the earlier it is

going to be simplified. This is the order followed by Siddharthan [2]. The sentence below, which contains passive voice and coordination, shows a particular case in which the results from empirical and hierarchical orders differ.

Trechos do vídeo foram exibidos pela rede britânica BBC e mostram uma fileira de corpos com ferimentos claramente provocados por tiros.

According to the hierarchical order, the sentence splitting due to coordination would happen before the transformation to active voice.

H1, Coordination: *Trechos do vídeo foram exibidos pela rede britânica BBC. Os trechos mostram uma fileira de corpos com ferimentos claramente provocados por tiros.*
H2, Passive: *A rede britânica BBC exibiu trechos do vídeo. Os trechos mostram uma fileira de corpos com ferimentos claramente provocados por tiros.*

The opposite would happen if the hierarchical rule is followed:

E1, Passive: *A rede britânica BBC exibiu trechos do vídeo e* ***mostram*** *uma fileira de corpos com ferimentos claramente provocados por tiros.*
E2, Coordination: *A rede britânica BBC exibiu trechos do vídeo.* ***A rede britânica BBC*** *mostram uma fileira de corpos com ferimentos claramente provocados por tiros.*

In E1, the first clause was transformed into active voice, but that caused the second clause to turn awkward, since its subject is now "*A rede britânica BBC*". On the other hand, in H1 the coordination is resolved first and both clauses become independent. In H2, then, the transformation to active voice happens in the first clause and has no interference in the second.

The purpose in establishing different orders for application of the rules is to investigate which order can cause less trouble for the parser, increasing the chance that its resulting analyses will be correct, mainly in relation to the segmentation of phrases and clauses. The expected simplification is the same independently of the order of application of the rules, although the order of the output sentences may differ.

Our hypothesis is that the hierarchical order facilitates the parsing of the subsequent segments, improving the quality of the simplification. In the next section we present the results of our experiments, which compare the performance of the system using empirical and hierarchical order. We also evaluate whether the rule for moving adverbial phrases has an impact on the quality of the subsequent simplifications when applied before any other rule, as a preprocessing step.

5 Evaluation

5.1 Evaluation Corpus

We have built a reference corpus[2] composed by 143 sentences, which is used to evaluate the system. The corpus was build by selecting sample sentences containing the particular syntactic phenomena of interest from a larger corpus of

[2] The annotated corpus and the system output are available at http://caravelas.icmc.usp.br/SS

154 automatically-parsed news articles. Some sentences contain more than one phenomena, so some more common phenomena have higher frequency in the corpus than others. Table 2 presents the frequency of each syntactic phenomenon in the corpus. The sentences in the corpus have been simplified manually following the simplification rules that we have defined for each phenomenon. For each sentence, there are annotations containing: the syntactic phenomena present in the sentence, a simplified version of the sentence considering each phenomenon at a time, and a simplified version of the sentence containing all required simplifications at once. We can observe that adverbial phrases (those requiring simplification) are the second most frequent phenomenon in the reference corpus, only behind apposition.

Table 2. Frequency, identification and simplification performance p/ phenomena

Phenomenon	Frequency	Precision	Recall	F-measure	Edit Dist.
Passive voice clauses	5	83.3	100	90.8	0.221
Apposition	64	86.8	51.6	64.7	0.237
All coordinate clauses	40	62.1	90.0	73.4	0.108
Asyndetic coord. clauses	5	20.0	80.0	32.0	0.174
Additive coord. clauses	10	45.0	90.0	60.0	0.111
Adversative coord. clauses	9	75.0	100	85.7	0.082
Correlated coord. clauses	4	75.0	75.0	75.0	0.206
Explanatory coord. clauses	8	70.0	87.5	77.7	0.052
Result coord. clauses	4	100	100	100	0.098
All subordinate clauses	59	49.1	84.7	62.16	0.221
Causal sub. clauses	5	27.8	100	43.5	0.223
Comparative sub. clauses	5	45.5	100	62.5	0.453
Concessive sub. clauses	9	77.8	77.8	77.8	0.179
Conditional sub. clauses	5	55.6	100	71.46	0.279
Consecutive sub. clauses	3	100	100	100	0.052
Conformative sub. clauses	5	21.7	100	35.66	0.066
Proportional sub. clauses	4	00.0	00.0	00.0	-
Purpose sub. clauses	13	57.1	92.3	70.55	0.149
Temporal sub. clauses	5	36.4	80.0	50.0	0.613
Non-finite sub. clauses	5	44.4	80.0	57.1	0.057
Restr. relative clauses	12	34.8	66.7	45.73	0.224
Non-restr. relative clauses	14	33.3	92.9	49.0	0.167
No Subj-Verb-Obj order	20	00.0	00.0	00.0	-
Adverbial phrases	25	**36.7**	**52.4**	**43.1**	**0.438**

5.2 Evaluating the Adverbial Rule

In order to evaluate the adequacy of the rule for simplification of adverbial phrases, we asked two annotators to compare the manually simplified sentences in the reference corpus (simplified only according to the adverbial rule) with the original sentences and indicate if their meaning had been preserved and if they

were easier to read. They were asked to assign one of the following labels to the sentences: 0) Resulting sentence changed the meaning of the original sentence; 1) Resulting sentence kept the meaning of the original sentence but did not make it easier to read; 2) Resulting sentence kept the meaning of the original sentence and made it easier to read. Table 3 shows the results of this evaluation for the 25 cases in the reference corpus. For Annotator 2, the adverbial rule was effective in making the majority of cases easier to read. However, for Annotator 1, the same amount of cases were viewed as missing the original meaning and as keeping the original meaning and being easier to read. We calculated the agreement between the selections of each annotator and it reached a Kappa score of 0.295. This is a low score, showing that the annotators disagreed in several cases. However, if we merge labels 0 and 1 in one class, we reach a Kappa score of 0.577, which reflects moderate agreement. We understand that a more thorough evaluation is needed to validate our rule for adverbial phrases, with a higher number of cases.

Table 3. Manual evaluation of adverbial rule

Annotator 1			Annotator 2		
0	1	2	0	1	2
41.66%	16.66%	41.66%	16.66%	37.50%	45.83%

5.3 Evaluating the Orders of Rule Application

In order to evaluate the impact of the order of application of the simplification rules, we first evaluate the performance of the system for each rule individually – for phase 1, phenomenon identification, and phase 2, actual simplification – and then evaluated the different orders of application of rules.

Phenomenon Identification. We have used the reference corpus to evaluate our system's performance. Since the simplification process consists of two phases, (1) phenomena identification and (2) simplification of these phenomena, we evaluate each phase separately. In this section we show the performance of the system for phase 1. Table 2 shows precision, recall and f-measure scores for identification of the phenomena to be simplified.

We can observe that the system could not recover any of the non-SVO cases nor any cases of asyndetic coordination. The system is totally dependent on the parser output analysis in order to identify these cases, since no other feature is used. At the moment, the only non-SVO orders that we try to recover is VSO and VOS, although the reference corpus contains other cases. Unfortunately, the parser rarely gets these two cases right, usually mistagging the subject and/or object with other functions. Hence, our null performance on these cases. Concerning asyndetic clauses, we rely on the parser identifying that the clauses are coordinated, and that also happens very rarely.

The adverbial phrases, which we discussed in Section 3, were identified with very high recall but very low precision. Unfortunately the parser mistags as

adverbial phrases constructs which are not, such as the adnominal adjunct in bold in "*O prefeito prometeu apoio para a realização de obras de infra-estrutura* ***na área em que a arena será erguida***". The low precision in identifying adverbial phrases can cause oversimplification of the sentence and unnecessary creation of new sentences.

Simplification. In this section we evaluate phase 2 of the simplification process, that is, the output or the simplification rules. For each sentence, we assess the output of the simplification of one phenomenon at a time and the output of the overall simplification, considering all phenomena present in the sentence.

In order to automate the evaluation process we implemented a comparison strategy between the reference corpus and the output of the system using Levenshtein (Edit) Distance. This measure compares two strings (simplified sentences, in our case) and return a score of how distant they are in terms of missing and misplaced characters. The lowest the distance score, the most similar the strings are. We normalize the distance scores by the size of the manually simplified sentence from the reference corpus. The last column of Table 2 presents the average distance scores for each phenomena across the reference corpus. We can observe very good results for additive coordinate clauses, explanatory coordinate clauses and comformative subordinate clauses. The results for adverbial phrases are not as good, and we believe this is due to the parser's phrase segmentation problems.

Orders. Finally, we evaluate the impact of the order of application of rules on the quality of the final simplified sentences, the ones which had all present phenomena simplified in sequence, following each of the orders we proposed. We also use Levenshtein distance as the sentence comparison measure. We experimented with hierarchical order (H), empirical order starting with the rule for adverbial phrases (E-ADV-S), and empirical order ending with the rule for adverbial phrases (E-ADV-E). Our hypotheses were that (1) H order would perform better than the empirical orders, due to its adaptability to each sentence since it obeys the parse tree, and that (2) E-ADV-S would perform better than E-ADV-E, since in the start the adverbial phrase rule could work as a "clean up" step which would facilitate further simplification in the sentence. Table 4 shows the distance scores for each order option. Although E-ADV-E got the best result, the scores are too similar for us to draw any strong conclusion. We believe the difference between the three values can be solely due to noise related to the distance measure used.

However, when we analyse the edit distance scores for the simplification of each phenomena individually, we observe that those with the best (lowest) distance scores (coordination and subordination) are the ones that H order handles

Table 4. Average edit distance for recursive application of rules

E-ADV-S	E-ADV-E	H
0.363	0.353	0.375

first, since inter-clause phenomena come higher in the syntactic trees. This indicates that H order can be beneficial, although the overall evaluation was not conclusive.

6 Concluding Remarks

We have presented our text simplification system for Portuguese and introduced two important issues: the treatment of adverbial phrases, which have not been dealt with in previous studies, and the order of application of simplification rules.

Long adverbial phrases (which require simplification) are the second most frequent phenomenon in our reference corpus. We reached 36% precision and 52% recall on identifying them in the input sentences, and got a 0.438 average distance score for the simplified sentence. As future work we intend to refine the identification of these phrases in order to increase the precision of the system.

We have experimented with three orders of application of simplification rules: hierarchical (following the parse tree), empirical plus rule for adverbials in the beginning, and empirical plus rule for adverbials in the end. We expected the hierarchical order to work best of all because of the fact that it conforms to the parse tree of the sentence. We also expected that the rule for adverbial phrases, when applied before other rules, could improve the performance of these, given the fact that it excludes adverbial phrases from the sentence and that makes subsequent parsings easier. Unfortunately, the results of our evaluation were not strong enough to confirm these hypotheses. We believe this is due to the strong impact of parsing mistakes on the performance of hierarchical order.

We plan a fine-grained manual evaluation of the simplification output, both in terms of the rules proposed and the performance of the system. To evaluate the performance of the system we aim to analyse separately each step of the construction of the simplified sentences, such as the recovery of anchors and segmentation of textual segments.

Acknowledgements

We thank FAPESP and Microsoft Research for supporting the PorSimples project.

References

1. INAF: Indicador de alfabetismo funcional INAF/Brasil (2007), http://www.acaoeducativa.org.br/portal/images/stories/pdfs/inaf2007.pdf
2. Siddharthan, A.: Syntactic Simplification and Text Cohesion. PhD thesis, University of Cambridge (2003)
3. Petersen, S.E., Ostendorf, M.: Text simplification for language learners: A corpus analysis. In: Proceedings of the Speech and Language Technology for Education Workshop (SLaTE 2007), Pennsylvania, USA, pp. 69–72 (2007)

4. Devlin, S., Unthank, G.: Helping aphasic people process online information. In: Proceedings of the 8th international ACM SIGACCESS Conference on Computers and Accessibility, Portland, USA, pp. 225–226 (2006)
5. Inui, K., Fujita, A., Takahashi, T., Iida, R., Iwakura, T.: Text simplification for reading assistance. In: Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, pp. 9–16 (2003)
6. Chandrasekar, R., Srinivas, B.: Automatic induction of rules for text simplification. Knowledge-Based Systems 10(3), 183–190 (1997)
7. Klebanov, B.B., Knight, K., Marcu, D.: Text simplification for information-seeking applications. In: Meersman, R., Tari, Z. (eds.) OTM 2004. LNCS, vol. 3290, pp. 735–747. Springer, Heidelberg (2004)
8. Candido Jr., A., Maziero, E., Gasperin, C., Pardo, T.A.S., Specia, L., Aluisio, S.M.: Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In: Proceedings of Workshop of Innovative Use of NLP for Building Educational Applications at NAACL 2009, pp. 34–42 (2009)
9. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In: Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING 1996), pp. 1041–1044 (1996)
10. Bick, E.: The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University (2000)
11. Temperley, D.: Minimization of dependency length in written english. Cognition 105, 300–333 (2007)

Comparing Sentence-Level Features for Authorship Analysis in Portuguese

Rui Sousa-Silva[1,3], Luís Sarmento[2], Tim Grant[1],
Eugénio Oliveira[2], and Belinda Maia[3]

[1] Centre for Forensic Linguistics at Aston University
[2] Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
[3] CLUP - Centro de Linguística da Universidade do Porto

Abstract. In this paper we compare the robustness of several types of stylistic markers to help discriminate authorship at sentence level. We train a SVM-based classifier using each set of features separately and perform sentence-level authorship analysis over corpus of editorials published in a Portuguese quality newspaper. Results show that features based on POS information, punctuation and word / sentence length contribute to a more robust sentence-level authorship analysis.

1 Introduction

Authorship analysis consists in identifying an author from a limited number of candidates (see [2]), and is increasingly relevant in cases of plagiarism detection and information filtering. Previous research on forensic linguistics has shown that there are several stylistic markers that help determine the authorship of a text *independently* of the topic of those texts. For example, Eagleson [1] claims that context-independent features are related to grammatical and lexical choices, including *syntactic structure*, *morphological inflections*, *vocabulary*, *spelling* and *punctuation*. Grant [2] uses a sophisticated Discriminant Function Analysis (DFA) to determine which variables help discriminate between the texts of three different authors, i.e. which variables are the best predictors to attribute texts to authors. He concluded that DFA is able to tell which of three authors is most likely to have written a queried text, and obtain an indication of the weight of evidence for each attribution, using a further analysis of the probabilities. He further concluded that the system proposed, both strong and conservative, is a trade-off between a robust method against mis-attribution and the conservatism in terms of the number of texts not firmly attributed. Hirst and Feiguina [3] perform authorship discrimination based on syntactic analysis, in particular on the frequency of bigrams of syntactic labels, obtained from partial parsing of the text, that they treat as pseudo-words, and consider their relative frequencies. They concluded that bigrams of syntactic labels are more discriminating that other features such as frequencies of rewrite rules, even with fragments of little more than 200 words (in which case the accuracy was boosted by using features such as unigram POS frequencies). All these however require strings of text of considerable length.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 51–54, 2010.

In this exploratory study, we investigate authorship analysis in Portuguese texts at *sentence level*. Performing authorship analysis at this level raises the additional problem that style markers should be able to work with short text strings and intra-sentence information only (i.e. very low frequency counts). We compare the robustness of several types of stylistic markers extracted at sentence-level to help discriminate the authorship of sentences using an SVM (*Support Vector Machine*)-based classifier over a corpus of editorials published in a Portuguese quality newspaper.

2 Stylistic Features for Authorship Analysis

We focus on the following potential and observable markers of authorship, which are to be extracted at *sentence level* with minimum linguistic processing:

- *POS-based features*: Computation of the frequency of each POS label found in the sentence, including function words and tense information in the case of verbs. Words found "POS-ambiguous" and "unknown" (e.g. neologisms) are also included in the POS-based features since their discriminatory power is potentially relevant.
- *Punctuation*: Frequency information about the usage of commas, "strong" punctuation marks, quotes and brackets.
- *Length*: Quantitative features such as the number of characters per word, the number of words per sentence, the number of 1 to 20-letter words and the number of words of 20+ letters.
- *Suffixation - superlatives and diminutives*: Suffixes are found to vary greatly among authors (i.e., they are largely idiolectal) in that they act as optional modifiers. We consider two particular forms of affixation: superlatives and diminutive forms.
- *Pronouns*: Information about explicit usage of relative and personal pronouns, whose use is dependent on the individual choices of the authors.
- *Conjunctions*: Information about the use of seven types of conjunctions, as the use of dependent and independent clauses is also highly idiolectal.

It is important to emphasise that *all* feature sets listed are *content agnostic*, which is intended to isolate our experiments from the potentially significant impact that content could have on authorship attribution, especially if columnists tend to focus their posts on certain preferential topics (e.g. economics vs. politics).

3 Experimental Set-Up

We built a corpus of editorials and opinion articles posted by columnists of a Portuguese quality daily newspaper[1]. The corpus comprises 915 posts by 23

[1] Jornal de Notícias – http://www.jn.pt

commentators from November 2008 to September 2009. From these we selected the top three most productive columnists - denoted by C_1, C_2, C_3 - who write editorials more than once a week covering a wide variety of issues, and not specialised in any specific topic domain. Commentator C_1 writes many short editorials (176 posts with 5.1 sentences / post), while commentators C_2 and C_3 write less frequently (74 and 51 posts each), but often longer editorials (17.9 and 17.7 sentences / post respectively).

We then randomly selected a set of 750 sentences for each of these commentators and trained a binary classifier to discriminate sentences written by each commentator. The authorship of a given test sentence is thus determined by the binary classifier that produces the highest classification score. Since the scores produced by all the three binary classifiers may be quite low – which can reflect the fact that the classifier has a very low confidence level in its result or the sentence is somehow difficult to differentiate – we introduced a threshold on the minimum value of classification score to be considered valid, c_{min}. Only classification scores higher than c_{min} were considered, which means that if none of the three classifiers reaches that threshold the test sentence at stake remains *unclassified*, i.e. no authorship is attributed to it.

We opted for using SVM as the classification algorithm for their well-known robustness in several text classification settings. We used the *SVM-light* [4] implementation. In all our experiments we used the default *SVM-light* parameters (including the choice for a linear kernel). In order to obtain Precision vs. Recall curves, we attempted authorship attribution with different values on the threshold c_{min}. We performed 5-fold cross-validation in all our experiments, and we micro-averaged partial Precision and Recall results.

4 Results and Analysis

We ran the training and classification procedure using *only one* of the six subgroup of markers described in Section 2 at a time. Figure 1 presents the precision vs. recall curves for all the six runs plus an additional curve for the run made using *all* stylistic markers. As expected, the performance obtained using any of the subgroup of markers alone is lower than the performance obtained using all the stylistic features. Among all groups of markers, *POS-based* ones seem to carry more information, performing almost as well as all subgroups of markers together. Two other groups perform reasonably well alone: *Punctuation* and *Length*. Interestingly, these subgroups use practically *no lexical information*, but instead rather simple statistics related to punctuation and word / sentence length. The other groups of stylistic markers are not so *robust* (i.e., their performance drops sharply) since they tend to occur in only a limited number of sentences. However, it is important to emphasize that *all* groups of features seem to carry some relevant information for authorship analysis, as the results obtained using all groups of markers shows.

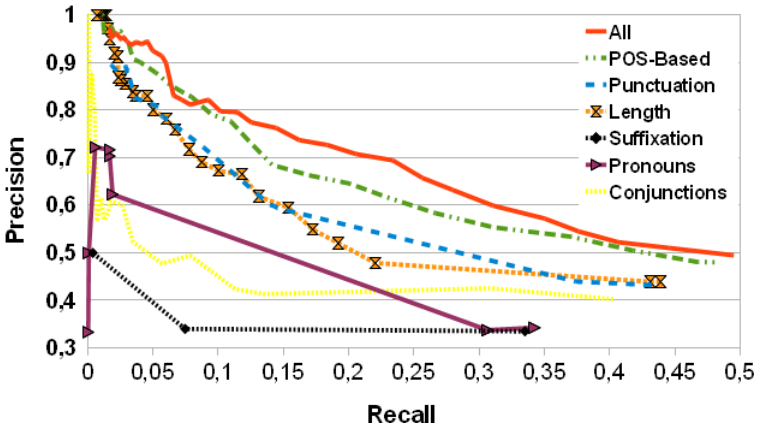

Fig. 1. Precision vs. Recall curves for each subgroup of stylistic markers

5 Conclusions and Future Work

This experiment confirms our initial assumptions that *content-agnostic* features can effectively be used for authorship analysis at sentence level. Among all the stylistic features, the excellent performance rate of punctuation stands out, which demonstrates that punctuation is one of the most robust stylistic features analysed. Affixes (superlatives and diminutives) and pronouns do not demonstrate enough robustness to perform *sentence-level* authorship attribution. Unsurprisingly, simple quantitative data (i.e. word and sentence length) perform well overall, with results that are similar to those obtained by punctuation.

Acknowledgments

This work was partially supported by grant SFRH/BD/23590/2005 FCT-Portugal, co-financed by POSI, and by grant SFRH/BD/47890/2008 FCT-Portugal, co-financed by POPH/FSE.

References

1. Eagleson, R.: Forensic analysis of personal written texts: a case study. In: Gibbons, J. (ed.) Forensic Linguistics: An Introduction to Language in the Justice System, pp. 362–373. Longman, Harlow (1994)
2. Grant, T.: Quantifying evidence in forensic authorship analysis. The International Journal of Speech, Language and the Law 14(1), 1–25 (2007)
3. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Lit Linguist Computing 22(4), 405–417 (2007)
4. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)

A Machine Learning Approach to Portuguese Clause Identification*

Eraldo R. Fernandes[1,2], Cícero N. dos Santos[3], and Ruy L. Milidiú[1]

[1] Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio
Rio de Janeiro, Brazil
[2] Laboratório de Automação
Instituto Federal de Educação, Ciência e Tecnologia de Goiás – IFG
Jataí, Brazil
[3] Mestrado em Informática Aplicada – MIA
Universidade de Fortaleza – UNIFOR
Fortaleza, Brazil
{efernandes,milidiu}@inf.puc-rio.br, cnogueira@unifor.br

Abstract. In this work, we apply and evaluate a machine-learning-based system to Portuguese clause identification. To the best of our knowledge, this is the first machine-learning-based approach to this task. The proposed system is based on *Entropy Guided Transformation Learning*. In order to train and evaluate the proposed system, we derive a clause annotated corpus from the *Bosque* corpus of the *Floresta Sintá(c)tica Project* – an European and Brazilian Portuguese treebank. We include part-of-speech (POS) tags to the derived corpus by using an automatic state-of-the-art tagger. Additionally, we use a simple heuristic to derive a phrase-chunk-like (PCL) feature from phrases in the Bosque corpus. We train an extractor to this sub-task and use it to automatically include the PCL feature in the derived clause corpus. We use POS and PCL tags as input features in the proposed clause identifier. This system achieves a $F_{\beta=1}$ of 73.90, when using the golden values of the PCL feature. When the automatic values are used, the system obtains $F_{\beta=1} = 69.31$. These are promising results for a first machine learning approach to Portuguese clause identification. Moreover, these results are achieved using a very simple PCL feature, which is generated by a PCL extractor developed with very little modeling effort.

1 Introduction

Clause identification [1] is a natural-language-processing task consisting of splitting a sentence into clauses. A clause is defined as a word sequence that contains a subject and a predicate. Clause identification is a special kind of shallow parsing, like phrase chunking [2]. Nevertheless, it is more difficult than phrase

* This work was partially funded by CNPq and FAPERJ grants 557.128/2009-9 and E-26/170028/2008. The first author was supported by a CNPq doctoral fellowship.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 55–64, 2010.

chunking, since some clauses also contain embedded clauses. Clause information is important for several more elaborated tasks such as full parsing and semantic role labeling.

The *PALAVRAS* parser [3] produces syntactic trees for Portuguese texts which include clause information. A manual-rule-based system to Portuguese clause identification is proposed in [4]. To the best of our knowledge, there is no machine-learning-based approach to Portuguese clause identification.

Conversely, for the English language, there are several such systems. The CoNLL'2001 shared task [1] is devoted to clause identification for English language texts. A corpus with clause annotations is provided and six systems have participated in the competition. The best system at CoNLL'2001 [5] shows a $F_{\beta=1} = 81.73$ and is based on boosted trees. After the competition, other systems were proposed and evaluated in the same corpus. A system based on *Entropy Guided Transformation Learning* (ETL) achieves a $F_{\beta=1} = 80.55$ with very little modeling effort [6]. The current state-of-the-art system [7] for this corpus achieves $F_{\beta=1} = 85.03$. This system is based on a modified perceptron algorithm specialized for phrase recognition.

In this work, we apply and evaluate an ETL system for Portuguese clause identification. ETL [8] is a machine learning strategy that generalizes *Transformation Based Learning* (TBL) [9] by automatically solving the TBL bottleneck: the construction of good template sets. ETL uses *entropy* in order to select the feature combinations that provide good template sets. First, ETL employs decision tree induction to perform an entropy guided template generation. Next, it applies the TBL algorithm to learn a set of transformation rules. ETL is an effective way to eliminate the need of a problem domain expert to build TBL templates.

Since our approach is based on a supervised machine learning method, we need a corpus annotated with clause boundaries in order to train our system. In this work, we derive the training corpus from the *Bosque* corpus of the *Floresta Sintá(c)tica Project* [10] – an European and Brazilian Portuguese treebank. We call this derived corpus the *clause corpus*. In our experiments, we randomly split it into three parts: train, development, and test.

The most effective systems to clause identification in English texts make use of part-of-speech tags and phrase chunks. We include POS tags in the clause corpus using a state-of-the-art tagger [11], which is also based on ETL. To the best of our knowledge, there is no phrase chunking definition for Portuguese language. Hence, using a simple heuristic, we derive a *phrase-chunk-like* (PCL) feature from phrases in the Bosque corpus. We train an ETL-based PCL extractor and use it to automatically generate this information in the clause corpus.

The proposed system achieves a $F_{\beta=1}$ of 73.90, when using the golden values of the PCL feature. When the automatic values are used, the system obtains $F_{\beta=1} = 69.31$. Using automatic values for the PCL feature yields more realistic estimates of the expected system performance for new texts. This sensitivity analysis indicates the potential impact of improvements on the PCL extractor.

The remainder of this paper is structured as follows. In Section 2, we describe the corpus derivation process. The general ETL method is briefly described in Section 3. In Section 4, we present the ETL modeling for Portuguese clause identification. Experimental results are reported and discussed in Section 5. Finally, in Section 6, we present our concluding remarks.

2 Corpus

Our approach is based on a supervised machine learning method. Therefore, we use a corpus annotated with clause boundaries in order to train our system. Here, we derive this training corpus from the *Bosque* corpus, which is a subset of the *Floresta Sintá(c)tica* [10] corpus. The *Floresta Sintá(c)tica Project*'s corpus consists in a treebank of European and Brazilian Portuguese texts. The syntactic trees has been automatically generated by the *PALAVRAS* parser [3]. However, the *Bosque* part of the corpus has been manually revised by linguists. Clause boundaries are one kind of syntactic information, among several others, that is available in the Bosque treebank. An example of a sentence from this corpus, broken into clauses by parentheses, is presented in Figure 1.

(Ninguém percebe (que ele quer (impor sua presença)) .)

Fig. 1. A clause annotated Bosque sentence

We call *clause corpus* the corpus annotated with clause boundaries that we derive from the Bosque. This corpus format is the same as the one provided in the CoNLL'2001 shared task [1]. Each line in the corpus contains a token along its corresponding features. As proposed in the CoNLL'2001 shared task, we tackle the clause identification task in three steps: (i) clause *start* identification; (ii) clause *end* identification; and (iii) complete *clause* identification. For each step, the clause corpus has one output feature. The format of the three output features is depicted in Table 1, based on the sentence in Figure 1. The *Start* column contains a binary feature that indicates the tokens where (at least) one clause *starts* (*S* tag). The *End* column contains a binary feature that indicates the tokens where (at least) one clause *ends* (*E* tag). Finally, the *Clause* feature codifies the complete clause set by using parentheses.

In the Bosque corpus, a clause is classified among three types: finite (*fcl*), non-finite (*icl*), and averbal (*acl*). In this work, we ignore the *averbal* clauses due to their unusual structure: they do not contain a verb. Additionally, we are not interested in classifying clauses according to their types. We just want to identify the clause boundaries. The corpus sizes are depicted in Table 2.

Table 1. Clause corpus format

Input			***Output***		
Word	*POS*	*PCL*	*Start*	*End*	*Clause*
Ninguém	pron-indp	B-NP	S	X	(S*
percebe	v-fin	B-VP	X	X	*
que	conj-s	B-PP	S	X	(S*
ele	pron-pers	B-NP	X	X	*
quer	v-fin	B-VP	X	X	*
impor	v-inf	B-VP	S	X	(S*
sua	pron-det	B-NP	X	X	*
presença	n	I-NP	X	E	*S)S)
.	.	O	X	E	*S)

Table 2. Clause corpus sizes

Part	*#Sentences*	*#Tokens*	*#Clauses*
Train	6,557	158,819	14,767
Development	1,405	34,596	3,180
Test	1,405	35,256	3,157

2.1 Input Features

The most effective systems to clause identification in English texts make use of POS tags and phrase chunks. We include POS tags in the clause corpus using a state-of-the-art tagger, also based on ETL. This tagger was proposed in [11] and its reported accuracy – evaluated on two Portuguese corpora – is over 96%.

The Bosque corpus includes phrase information, but *phrase chunking* information is not included. Although phrases and phrase chunks are closely related, there are important differences between them. For instance, phrases can contain another phrases, that is, phrases can be embedded. On the other hand, phrase chunks are flat and are never embedded. Therefore, phrase chunks are simpler than phrases and, consequently, more suitable for machine learning methods.

The idea of breaking a sentence into phrase chunks, for the English language, was firstly proposed by Abney [12]. Phrase chunking is a kind of shallow parsing, yet powerful. It is related to prosodic aspects of the sentence. Unfortunately, as far as we know, there is no equivalent proposal for Portuguese language. There are works related to nominal chunks (base noun phrases) [13]. However, prepositional and verbal chunks also provide valuable information.

In this work, we propose a simple heuristic in order to derive a *phrase-chunk-like* feature from the phrases in the Bosque corpus. We define as chunk all *consecutive* tokens within the same deepest-level phrase. We consider three types of phrase chunks: verbal, nominal, and prepositional. In order to codify this feature, we use the IOB2 tagging style, as in the English-language corpus provided in the CoNLL'2000 shared task [14]. In Figure 2, we show a sentence along with its phrase tree and the resulting PCL feature. Although this heuristic is very simple,

the resulting feature conveys relevant information for the clause identification task, as indicated by some experiments reported in Section 5.

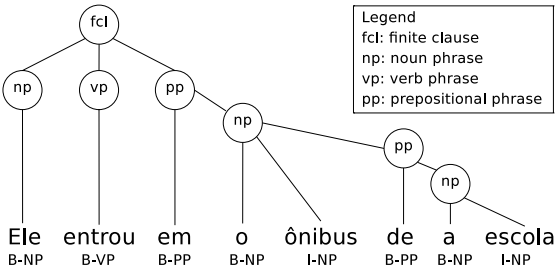

Fig. 2. Sentence along with its phrase tree and the corresponding PCL feature

We train an ETL-based PCL extractor and use it to automatically include this feature in the clause corpus. Therefore, we have two versions of the PCL feature: (i) the golden values derived from the Bosque treebank by the heuristic and (ii) the automatic values given by the trained PCL extractor. It is important to notice that this extractor has been developed with little modeling effort.

3 Entropy Guided Transformation Learning

Entropy Guided Transformation Learning [2] generalizes Transformation Based Learning (TBL) [9] by automatically generating rule templates. ETL employs an *entropy guided template generation* approach, which uses the *information gain* measure in order to select feature combinations that provide good template sets. ETL has been successfully applied to part-of-speech tagging [11], phrase chunking, named entity recognition [8,15], and dependency parsing [16] – producing results at least as good as the ones of TBL with handcrafted templates. In Figure 3, we present a concise description of the ETL algorithm. A detailed description of ETL can be found in [2,8]. Several ETL-based multi-language processors are freely available on the Web through the *F-EXT*[1] service [17].

4 ETL Modeling

In this section, we show our ETL modeling for the Portuguese clause identification task. This modeling is strongly based on the one proposed in [6] to English clause identification. We approach the clause identification problem in three steps: (i) clause start identification; (ii) clause end identification; and (iii) complete clause identification. We solve these three sub-tasks sequentially. Therefore, we use the information produced in previous steps as input to the next ones. First, we use *Start* tags as input for the end classifier. Next, we use both *Start* and *End* tags to identify the complete clauses.

[1] http://www.learn.inf.puc-rio.br/

1. Applies the baseline system to the training corpus.
2. Generates the rule templates by using an entropy-guided approach.
3. *Repeat*:
   (a) Generates, for each classification error in the current version of the training corpus, correcting rules by instantiating the templates.
   (b) Computes rule scores. The rule score is defined as the difference between the total number of repaired errors and the total number of generated errors.
   (c) *Stop*, if there is no rule with a score above a given threshold.
   (d) Applies the best-scoring rule to the training corpus.
   (e) Adds the best-scoring rule to the sequence of learned rules.
4. Returns the sequence of learned rules.

Fig. 3. Entropy Guided Transformation Learning

4.1 Baseline System

We adopt the simple baseline system proposed in the CoNLL'2001 shared task. This system just assigns one clause for the whole sentence. This baseline system is used in the three steps.

4.2 Clause Boundary Candidates

The first and second steps consist in identifying the clause boundary candidates, that is, start and end tokens. These steps identify the tokens that are good candidates to clause boundaries, without any concern to consistence among them. We model these two sub-tasks as token-classification problems. In Table 1, we illustrate the corpus format through an example. The *Start* and *End* columns in the table respectively indicate the *start* and *end* classifications. In the first step, if a token *starts* one or more clauses, it must be classified as *S*, otherwise, it must be classified as *X*. Similarly, in the second step, if a token *ends* one or more clauses, it must be classified as *E*, otherwise as *X*.

4.3 Complete Clause Identification

The last and most difficult step consists in splitting a given sentence into clauses. In the clause corpus, the complete clauses within a sentence are encoded through a unique token feature using the following tags: *(S** – indicating that the token starts a clause; **S)* – indicating that the token ends a clause; * – representing a token that neither starts nor ends a clause; and any combinations of the previous to represent tokens that start or end more than one clause. The *Clause* column in Table 1 contains the tags that encode the clauses within the sentence illustrated in Figure 1.

For this last sub-task, we present two modeling approaches: *ETL-Token* and *ETL-Pair*. The *ETL-Token* consists of a token classification approach. In this approach, we apply ETL in a straightforward manner. We train an ETL model to classify each token as *, *(S**, **S)*, or any tag combination appearing in the *Clause* column of the training corpus. This approach is very simple but also limited.

We observe that many clauses are tokenwise long. For instance, in the training corpus, the fraction of clauses with length longer than 14 tokens is greater than 40%. For such cases, even using a window of 27 tokens (the current token plus the thirteen tokens on each side), one clause boundary is not included when classifying the other one. We observe that this window size is computationally prohibitive for the ETL algorithm.

In order to capture a broader context, we try a second modeling approach to the third step: *ETL-Pair*. This approach uses the output of the *Start* and *End* classifiers to create a new corpus. For each start-end pair of tokens from a given sentence in the original corpus, we generate one example in the new corpus. We attach to this new example all the original input features of both start and end tokens. Next, we train a binary ETL model that learns to classify which examples (pairs of tokens) correspond to correct clause boundaries.

4.4 Derived Features

We use the three input features in the clause corpus – word, POS, and PCL – plus some derived features. We derive these additional features in the same fashion as in [18], although we use just a small subset of the features proposed by these authors.

The derived features inform about the occurrence of relevant elements within a specific sentence fragment. The following elements are the relevant ones: *pronouns*, *conjunctions*, *verbal chunks*, *start tokens*, and *end tokens*. We call *verbal chunks* the ones with chunk tag with value `verb`. We generate two features for each relevant element and sentence fragment: a flag indicating the occurrence within the fragment and the number of occurrences within the fragment.

For the token classifiers (*Start*, *End*, and *ETL-Token*) we use the same sentence fragmentation scheme. For each token we derive twenty features: ten for the sentence fragment before the token and ten for the sentence fragment after it. For the *ETL-Pair* classifier we use a different scheme. For each start-end pair of tokens we derive thirty features: ten for the sentence fragment before the start token; ten for the sentence fragment after the end token; and ten for the sentence fragment between the start and end tokens. Observe that a derived feature is only used when its required information is available.

5 Experiments

We use the development corpus in order to tune the ETL parameters. For the three token classifiers (*Start*, *End*, and *ETL-Token*), we set the context window size parameter to 7. Whereas for the *ETL-Pair* classifier we set the window size to 9. For all approaches, we set the rule score threshold to 2.

In order to evaluate the potential and real impact of the PCL feature in the proposed system performance, we train and evaluate three independent versions of the system: (i) using no information of the PCL feature; (ii) using the automatic values of the PCL feature; and (iii) using the golden values of the PCL

feature. Only in version (i), where no PCL information is used, we consider the verbal tokens (POS tag equal to verb) as relevant elements when generating the derived features. In (ii), the PCL values are provided by the automatic PCL extractor. In (iii), the PCL values are obtained directly from the Bosque treebank by the PCL derivation heuristic.

The resulting performances are presented in Table 3. One can observe that the $F_{\beta=1}$ for the version that uses the PCL golden values is almost seven points greater than the one that uses no PCL information. The sensitivity of the system performance to this feature clearly indicates its potential positive impact. Using automatic values for the PCL feature yields more realistic estimates of the expected system performance for new texts. The system performance using the automatic values of this feature also indicates the positive impact of improvements on the PCL extractor.

Table 3. PCL impact on *ETL-Pair* performance

PCL	*Precision*	*Recall*	$F_{\beta=1}$
No	75.18	60.34	66.95
Automatic	78.14	62.27	69.31
Golden	83.78	66.11	73.90

In Table 4, we present the performances for the four proposed classifiers – *Start*, *End*, *ETL-Token*, and *ETL-Pair* – on the test corpus. These results are divided into two groups: golden and automatic values of the PCL feature. The $F_{\beta=1}$ of the *ETL-Pair* system is over two points greater than the one of the *ETL-Token* system. We believe that this improvement is due to the stronger contextual information used by the *ETL-Pair* approach.

Table 4. Test corpus performances

Task/Strategy	*Golden*			*Automatic*		
	Precision	*Recall*	$F_{\beta=1}$	*Precision*	*Recall*	$F_{\beta=1}$
Start	93.37	87.25	90.20	90.63	84.25	87.32
End	85.64	79.89	82.67	84.29	74.78	79.25
ETL-Token	74.59	**68.45**	71.39	69.85	**64.65**	67.15
ETL-Pair	**83.78**	66.11	**73.90**	**78.14**	62.27	**69.31**
BLS	82.06	36.52	50.55	82.06	36.52	50.55

6 Conclusions

In this paper, we apply and evaluate a machine-learning-based system to Portuguese clause identification. The system is based on the machine learning technique called *Entropy Guided Transformation Learning*. In order to train and

evaluate our system, we derive a clause annotated corpus from the *Bosque* treebank of the *Floresta Sintá(c)tica Project*. We include POS tags in the clause corpus by using a state-of-the-art tagger, also based on ETL.

Phrase chunking is a very important feature for several Natural Language Processing tasks, including clause identification. However, to the best of our knowledge, there is no phrase chunking definition to Portuguese language. So, we propose a simple heuristic to derive a phrase-chunk-like feature from phrases in the Bosque treebank. We train an ETL extractor to this sub-task and use it to include this information in the derived clause corpus.

The clause identification modeling used in this work is based on the approach proposed in [6] to English language. The problem is divided into three steps: (i) clause start identification; (ii) clause end identification; and (iii) complete clause identification. We propose one system for the first step, another for the second step, and two systems for the third step.

We report the performance of the four systems on the derived clause corpus. The impact of the PCL feature and the automatic PCL extractor on the system performance is also evaluated. We report the system performance on three scenarios: using no PCL information, using the automatic values, and using the golden values of the PCL feature. These results indicate that the PCL feature is informative to the clause identification task and the ETL-based PCL extractor is effective to improve the performance of the proposed clause identifier.

We believe that using a better phrase chunking information we can improve our result in this task. We are working on a better heuristic to extract phrase chunks from Bosque. Additionally, the ETL-based PCL extractor can be substantially improved, since it has been developed with very little modeling effort.

Acknowledgments

We thank Bernardo A. Pires and Guilherme De Napoli for their efforts to code the scripts used to derive the clause corpus. We also thank Maria Cláudia de Freitas for the important clarifications about Bosque treebank.

References

1. Sang, E.F.T.K., Déjean, H.: Introduction to the CoNLL 2001 shared task: Clause identification. In: Proceedings of Fifth Conference on Computational Natural Language Learning, Toulouse, France (2001)
2. Milidiú, R.L., dos Santos, C.N., Duarte, J.C.: Phrase chunking using entropy guided transformation learning. In: Proceedings of ACL 2008: HLT, pp. 647–655. Association for Computational Linguistics, Columbus (2008)
3. Bick, E.: The Parsing System Palavras: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University, Aarhus, Denmark. Aarhus University Press (November 2000)
4. Leffa, V.J.: Clause processing in complex sentences. In: Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Espanha, vol. 2, pp. 937–943 (1998)

5. Carreras, X., Màrquez, L.: Boosting trees for clause splitting. In: Proceedings of Fifth Conference on Computational Natural Language Learning, Toulouse, France (2001)
6. Fernandes, E.R., Pires, B.A., dos Santos, C.N., Milidiú, R.L.: Clause identification using entropy guided transformation learning. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), São Carlos, Brazil (2009)
7. Carreras, X., Màrquez, L., Castro, J.: Filtering-ranking perceptron learning for partial parsing. Machine Learning 60(1–3), 41–71 (2005)
8. dos Santos, C.N., Milidiú, R.L.: Entropy Guided Transformation Learning. In: Foundations of Computational Intelligence, vol. 1 of Learning and Approximation. vol. 201 of Studies in Computational Intelligence, pp. 159–184. Springer, Heidelberg (2009)
9. Brill, E.: Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics 21(4), 543–565 (1995)
10. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, thicker and easier. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 216–219. Springer, Heidelberg (2008)
11. dos Santos, C.N., Milidiú, R.L., Renteria, R.P.: Portuguese part-of-speech tagging using entropy guided transformation learning. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 143–152. Springer, Heidelberg (2008)
12. Abney, S.: Parsing by Chunks. In: Principle-Based Parsing. Kluwer Academic Publishers, Dordrecht (1991)
13. Freitas, M.C., Garrao, M., Oliveira, C., Santos, C.N.d., Silveira, M.: A anotação de um corpus para o aprendizado supervisionado de um modelo de sn. In: Proceedings of the III TIL / XXV Congresso da SBC, São Leopoldo - RS - Brasil (2005)
14. Sang, E.F.T.K.: Text chunking by system combination. In: Proceedings of Conference on Computational Natural Language Learning, Lisbon, Portugal (2000)
15. Milidiú, R.L., dos Santos, C.N., Duarte, J.C.: Portuguese corpus-based learning using ETL. Journal of the Brazilian Computer Society 14(4) (2008)
16. Milidiú, R.L., dos Santos, C.N., Crestana, C.E.M.: A token classification approach to dependency parsing. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), São Carlos, Brazil (2009)
17. Fernandes, E.R., dos Santos, C.N., Milidiú, R.L.: Portuguese language processing service. In: Proceedings of the Web in Ibero-America Alternate Track of the 18th World Wide Web Conference, Madrid (2009)
18. Carreras, X., Màrquez, L., Punyakanok, V., Roth, D.: Learning and inference for clause identification. In: Proceedings of the Thirteenth European Conference on Machine Learning, pp. 35–47 (2002)

A Hybrid Approach for Multiword Expression Identification

Carlos Ramisch[1,2], Helena de Medeiros Caseli[3], Aline Villavicencio[2,4]
André Machado[2], and Maria José Finatto[5]

[1] GETALP/LIG, University of Grenoble (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Department of Computer Science, Federal University of São Carlos (Brazil)
[4] Department of Computer Sciences, Bath University (UK)
[5] Institute of Language and Linguistics, Federal University of Rio Grande do Sul (Brazil)
ceramisch@inf.ufrgs.br, helenacaseli@dc.ufscar.br,
{avillavicencio,ammachado}@inf.ufrgs.br, mfinatto@terra.com.br

Abstract. Considerable attention has been given to the problem of Multiword Expression (MWE) identification and treatment, for NLP tasks like parsing and generation, to improve the quality of results. Statistical methods have been often employed for MWE identification, as an inexpensive and language independent way of finding co-occurrence patterns. On the other hand, more linguistically motivated methods for identification, which employ information such as POS filters and lexical alignment between languages, can produce more targeted candidate lists. In this paper we propose a hybrid approach that combines the strenghts of different sources of information using a machine learning algorithm to produce more robust and precise results. Automatic evaluation on gold standards shows that the performance of our hybrid method is superior to the individual results of statistical and alignment-based MWE extraction approaches for Portuguese and for English. This method can be used to aid lexicographic work by providing a more targeted MWE candidate list.

1 Introduction

Recent research on Multiword Expressions (MWEs) has devoted considerable attention to their identification. One of the problems that these works address is that MWEs can be defined as combinations of words that have idyosincrasies in their lexical, syntactic, semantic, pragmatic or statistical properties [1], such as idioms (*make ends meet*), phrasal verbs (*find out*), light verbs (*give a speech*) and compounds (*mother nature*). However, MWEs are very numerous in languages accouting for between 30% and 45% of spoken English and 21% of academic prose [2], and having the same order of magnitude in a speaker's lexicon as the number of single words [3]. Moreover, if we consider that new MWEs are also constantly coined (*credit crunch*), and that for language from a specific domain

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 65–74, 2010.

the specialized vocabulary is going to consist largely of MWEs (*chromosomal mutation*), these estimates are likely to be conservative underestimates.

In this context, especially for NLP tasks that involve some kind of semantic processing, it is important to adequately identify and treat MWEs, as failing to do so may cause serious problems [1]. For example, in order to avoid the generation of unnatural sentences, a Machine Translation system must translate the idiom *to kick the bucket* differently in a sentence like *He was only 39 years old when he kicked the bucket* (meaning *to die*) than in *The janitor kicked the bucket with water*.

Therefore, there is a need for robust (semi-)automatic ways of acquiring lexical information for MWEs that can contribute to improving the quality of NLP systems. In this context, a number of methods for identifying MWEs from corpora have been proposed. For this task they employ information that ranges from purely statistical Association Measures (AMs), to more linguistically-based, such as e.g. Part-of-Speech (POS) patterns, with varying degrees of success [4,5].

While the former can retrieve a large list of multiword units, more linguistically motivated methods for MWE identification, such as those based on POS filtering or lexical alignment, on the other hand, may result in a more accurate list of candidates.

After evaluating statistical AM and alignment-based approaches separately in previous work [6,7], in this paper we investigate their weighted combination aiming at a more robust method that could output a more accurate set of MWE candidates than those of the individual methods. The proposed approach can be used to aid lexicographic work by providing a more targeted MWE candidate list to keep lexical resources up to date and also to improving the quality of NLP systems.

The remainder of this paper is structured as follows. In section 2 we briefly discuss MWEs and their identification. Section 3 presents the materials used in our experiments while section 4 describes the hybrid method proposed to extract MWEs. Section 5 presents the results and section 6 finishes this paper with some conclusions and proposals for future work.

2 Related Work

MWEs have been the focus of both linguistic and computational work, and they have proved to be a difficult problem to tackle from either field [1]. The different phenomena that are defined as MWEs form a very heterogeneous group, with phrasal verbs, idioms, compounds, among others, each with its particular characteristics. Moreover, even within a single MWE type there is considerable variation in their possible linguistic realizations. Verbal idioms, for example, vary in terms of morphosyntactic and semantic flexibility from more rigid combinations (*kick the bucket*) to more flexible ones (*touch a nerve*). As a consequence MWEs defy attempts to capture them uniformly.

Due to the tight connection between the elements of an MWE and their co-occurrence patterns, AMs have been often used to identify them [4,8], as they are sensitive to such patterns. Since we expect the component words of an MWE to occur frequently together, these measures can give an indication of whether a sequence of words is a MWE. The advantage of using them in the identification of MWEs is that AMs are an inexpensive language and type independent means of detecting recurrent patterns and can be democratically applied to any language and MWE type. The effectiveness of these methods seem to depend on the MWEs themselves (e.g. type, syntactic flexibility) [8], in characteristics of the corpus used (e.g. size and domain) [4], and on the gold standard used for evaluation [7].

A number of these works have also combined these measures with linguistic information such as syntactic and semantic properties of the MWEs [9,8] or automatic word alignment [10]. Fazly, Cook and Stevenson [8], for instance, use properties like lexical and syntactic flexibility in statistical measures for verb-noun idiom identification. Ramisch et al. [11] combine standard statistical measures with information about syntactic flexibility using a supervised machine learning approach for the identification of Verb-Particle Constructions (VPCs).

Some work has looked for evidence from other language for MWE identification. For instance, Melamed's proposal for the automatic detection of non-compositional compounds (NCC) [12] is based on the idea that their translation to another language does not usually correspond to their word-for-word literal translation. This method can successfully identify many NCCs, but it does not use monolingual information about possible NCCs within a language. The work of Villada Moirón and Tiedemann [10] seems to be the most similar to the approach proposed in this paper. Their method looks at the automatically generated translations of MWE candidates assuming that the translations of idiomatic expressions would be less predictable and less compositional than the non-idiomatic cases. However, while their method uses the alignment information just for ranking the MWE candidates, in this paper, the word alignment is the basis of MWE extraction process.

In this paper we investigate the combination of several sources of information for the identification of MWEs. In particular, we propose that the combination of statistical AMs and alignment-based information has a positive effect on the performance of this task, as they each capture different aspects of MWEs. We also evaluate thoroughly their contributions to the overall performance, looking at factors, such as language and size of the ngram, that influence these results.

3 The Corpus and Reference Lists

For our experiments, we used the Corpus of Pediatrics [13], a parallel corpus composed of 283 pairs of texts in Portuguese (785,448 words) and their translations to English (729,923 words) extracted from the *Jornal de Pediatria*[1]. We use a parallel corpus to evaluate the MWE identification for these two different

[1] www.jped.com.br

languages, Portuguese (pt) and English (en), and also to investigate whether the choice of language influences the results obtained.

Our automatic evaluation process uses the Pediatrics Glossary, a domain-specific resource built semi-automatically from the Corpus of Pediatrics for supporting translation studies.[2] The Portuguese Glossary was constructed by first extracting all ngrams (with n ranging from 2 to 4) from the texts which occurred at least 5 times in the corpus, then applying a POS filter to exclude candidates beginning with Article + Noun and beginning or finishing with Verbs and, finally, manually verifying the remaining entries. Subsequently an enrichment process was performed as described in [14] to include all the valid bigrams contained in the trigrams and removed during the construction of the Glossary. The English Glossary was built by a similar process with translations of the ngrams in the Portuguese Glossary. The final versions of the gold standards have 2,150 terms in Portuguese and 883 terms in English[3]. Due to the smaller number of entries in the English Glossary, we also considered as true positives the candidates contained in a general dictionary, such as the Cambridge International Dictionary of Idioms [15], and these two sources are marked in table 1 as specialized and generic respectively.

Table 1. Number of reference entries in each gold standard

	Specialized	Generic	Total
pt	2150	—	2150
en	883	1382	2190

4 MWE Extraction Methodology

In this paper we propose a hybrid method that combines two independent approaches for MWE identification using a Bayesian network classifier. The first approach applies well-know AMs to all the bigrams and trigrams generated from each corpus: Pointwise Mutual Information (PMI), Mutual Information (MI), t-score, χ^2, Dice coefficient, Fisher's exact test, Poisson-Stirling measure (PS) and Odds ratio, as implemented in the Ngram Statistics Package [16].

The second one, the alignment-based approach, is based on the automatic lexical alignment of Portuguese and English versions of the Corpus of Pediatrics generated by the statistical word aligner `GIZA++` [17]. The hypothesis is that when the lexical aligner encounters a sequence in the source language that cannot be resolved by aligning the target words individually, this sequence is taken to be a MWE candidate. Thus, the alignment-based approach considers as MWE candidates the sequences of two or more consecutive source words joined by the aligner regardless of whether they are aligned with one or more target words.

[2] www6.ufrgs.br/textquim/Dicionarios/DicPed

[3] www.inf.pucrs.br/~ontolp/downloads-ontolplista.php

Table 2. MWE candidates per method, language and ngram size

statistical						
	no filter			filters		
	$n = 2$	$n = 3$	Total	$n = 2$	$n = 3$	Total
pt	244420	513494	757914	11290	4553	15843
en	230130	492154	722284	10311	4526	14837
alignment-based						
	no filter			filters		
	$n = 2$	$n = 3$	Total	$n = 2$	$n = 3$	Total
pt	15333	7373	22706	12154	5518	17672
en	16345	7469	23814	12222	5154	17376
statistical ∩ alignment-based						(filters)
	$n = 2$		$n = 3$			Total
pt	1376		134			1510
en	1921		109			2030

The original corpora were POS tagged using the `Apertium`[4] tools [18] with augmented lexicon [19]. Morphological information was used to filter the candidate lists of both approaches. The filters were applied uniformly, removing:

- punctuation, numbers and special characters (dashes, slashes, brackets,...);
- candidates below a certain threshold (5 occurrences in corpus for AMs, and 5 occurrences as alignment for alignment-based method);
- candidates starting with function words (determiners, auxiliary verbs, pronouns, adverbs, conjunctions, forms of the verb *to be* and prepositions *from, to* and *of*. In this we follow Caseli et al. [20] who found that these patterns are effective for filtering out noise, without removing many false positives.

The corpora were then independently given as input to each of the approaches, and as a result, two lists of MWE candidates for each language were generated. Table 2 shows the number of original candidates extracted for each language before and after filtering.

Both languages have about the same number of candidates for each approach, and filtering considerably reduces the candidate lists, especially for the AMs.

The last section of table 2 shows the intersection between the two methods, which indicates that their candidates are essentially different: less than 15% of the candidates extracted by the alignment-based method are also captured by the statistical method and vice versa. One difference between the approaches is that the alignment-based approach is able to extract non-contiguous sequences. Therefore when it detects an ngram from the source aligned with an ngram in the target language (an n:m alignment), if there are intervening words between the two ngrams, the candidate will not include them.

[4] `Apertium` is an open-source machine translation engine and toolbox available at `http://www.apertium.org`.

Table 3. Sample of the English training set

ngram	align	statistical								Class
		Dice	Odds	PMI	PS	t-score	MI	χ^2	Fisher	
abnormal findings	Yes	.03	114.1	6.74	25.70	2.62	0	734.73	0	No
adrenal insufficiency	No	.46	10376	11.6	371.9	7.28	.0008	160784	0	Yes
óxido nítrico	Yes	.95	8553397	14.5	289.3	5.66	.0006	733177	0	Yes
academia americana	No	.52	74302	13.3	197.4	4.9	.0004	244244	0	No

For example, from *a mild pain* and *a characteristic pain* aligned with the Portuguese *uma discreta dor* and *uma dor característica* the aligner proposes *a pain* as a candidate even though these two words never occur adjacently in the English corpus, and consequently the statistical method does not propose them as a candidate ngram.

For combining the different methods, a classifier was constructed using the Weka package [21]. The input for each language was the set of filtered ngrams (15,843 for Portuguese and 14,837 for English) annotated with the values of the statistical measures and the judgement of the lexical aligner as to whether the ngram is a possible MWE candidate. Table 3 shows some examples of English and Portuguese entries from the training set. As discussed in the next section, the data sets are unbalanced, with a much larger proportion of non-MWEs than MWEs. Therefore, a Bayesian Network classifier is used to combine the different approaches, since it has been found to be robust and less sensitive to highly unbalanced classes.[5]

5 Experiments and Results

We evaluate the efficacy of the combined approach for MWE identification (from §4) in a domain-specific corpus using the gold standards for each language (§ 3). The results are reported in terms of precision (#correct vs #proposed candidates), recall (#correct vs #candidates in gold standard) and F-measure $((2 * precision * recall)/(precision + recall))$.

The baseline for comparison is obtained by evaluating the individual approaches independently. Table 4 shows the number of True Positives (TPs) in each candidate list considering both the MWEs in the specialized Portuguese and English Pediatrics Glossaries (pt_spec and en_spec) and those in the general English Dictionary (en_spec+gen), while table 5 shows the Mean Average Precision of each AM taken independently, which range from 11.23% for Fisher's exact test to 55.83% for Odds ratio.

In the results for Portuguese the statistical approach captures 86.14% of the MWEs in the text with a precision of 11.69%, while for English only 68.06% of the true instances are captured with a precision of 4%. The differences in the results for

[5] Using, e.g. decision trees on the English data generated a single class model guessing "No" for all candidates.

Table 4. True Positives (TP), precision, recall and F-measure of individual methods

	statistical				alignment-based			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
pt_spec	1852	11.69%	86.14%	20.59%	240	0.97%	11.16%	1.78%
en_spec	601	4.05%	68.06%	7.65%	84	0.35%	9.51%	0.68%
en_spec+gen	774	5.22%	35.34%	9.10%	224	0.94%	10.23%	1.72%

Table 5. Mean Average Precision of the statistical Association Measures (AMs) taken individually

	PMI	MI	PS	Dice	Odds	t-score	χ^2	Fisher
pt	28.18%	17.53%	23.99%	53.47%	55.83%	13.9%	54.38%	11.23%
en	13.17%	12.79%	7.07%	25.09%	25.7%	7.14%	25.82%	5.12%

these languages can be explained to a large extent due to the differences in coverage of the gold standards, with the Portuguese Glossary containing a much larger number of entries. Indeed, using the extended English gold standard with both specialized and generic MWEs improves the F-measure for both approaches, and this extended resource is adopted in the subsequent evaluations. The alignment-based method has a lower performance partly due to the larger number of candidates to consider (table 2). Although a higher alignment frequency threshold could considerably improve the precision of the aligner [7], more restrictive filters were not applied because we wanted to investigate how much the combination of these methods can filter out the noise in each of the candidates lists.

The results of the Bayesian network classifier using different feature sets and 10-fold cross validation are shown for each language in table 6. For both languages the hybrid model is able to generate much better candidates than the individual methods: e.g. for Portuguese, the Bayesian classifier yields an F-measure of around 50% against 20.59% and 1.78% for the statistical and alignment-based methods, respectively.

To evaluate the contribution of the individual methods to these results we consider four different feature sets: (a) a subset of the Association Measures (subAM), namely PMI, PS and MI, which do not involve the construction of contingency tables and can be straightforwardly applied to ngrams of arbitrary size, (b) the combination of this subset of AMs and the alignment-based approach (subAM + align), (c) all AMs (allAM) which includes subAMs for bigrams and trigrams, and the other AMs only for bigrams (since they rely on contingency tables) and (d) the combination of all AMs and the alignment feature (allAM + align). In terms of individual features, the aligner only improves the performance in subAM for English, where it provides enough extra information for an increase in performance from 0% (subAM) to 4.91% (subAM + align). This suggests that the alignment information helps to add robustness to the process. To further evaluate the contribution of this feature, we built a decision tree with the same

Table 6. Bayesian network classifier for different feature sets, in Portuguese and in English

	Portuguese				English			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
subAM	1102	48.29%	51.26%	49.73%	0	—	—	—
subAM + align	1103	47.98%	51.30%	49.58%	62	16.49%	2.88%	4.91%
allAM	1100	43.51%	51.16%	47.03%	464	19.74%	21.58%	20.62%
allAM + align	1084	43.41%	50.42%	46.65%	465	19.68%	21.63%	20.61%

Table 7. Classifier performance for different feature sets and languages, only bigrams

	Portuguese				English			
	TP	Precision	Recall	F-measure	TP	Precision	Recall	F-measure
subAM	1021	49.11%	71.90%	58.36%	228	32.66%	16.06%	21.53%
subAM + align	1026	45.72%	72.25%	56.00%	267	26.67%	18.80%	22.06%
allAM	1113	43.04%	78.38%	55.57%	459	19.94%	32.32%	24.66%
allAM + align	1113	42.68%	78.38%	55.26%	459	19.93%	32.32%	24.66%

training sets and in the resulting trees the alignment feature is only used after PMI, PS and Dice, which seem to be better predictors of MWEs. The addition of further AMs seems to also have the effect of providing more robustness to the task, as they result in considerably higher F-measures for English unlike for Portuguese. These different performances for the two languages are in line with Evert and Krenn's argument that statistical AMs are highly dependent on type and language [4].

Some of the statistical measures used as features are based upon contingency tables and are therefore not straightforwardly applicable to trigrams.[6] Therefore, in the training set, these measures are represented with a missing value ("?") for trigrams. In order to verify in more details whether these extra values affect the results obtained we performed a second evaluation where we only analyzed candidates which have non-null values for these features (i.e. bigrams).

The performance of the classifiers built on the bigram data set only are summarized in table 7[7]. The results further confirm that in some cases these extra features are adding enough information for the performance of the Bayesian Networks to improve. This is particularly clear in the case of English subAM, for which the F-measure improves by as much as 16% (for the case without alignment information). The difference in the results obtained by only considering the bigrams suggests that the methods propose a larger and more accurate set of bigram candidates, but they do not seem as effective for trigrams. Further investigation would have to be conducted to properly assess which factors play a role in the lower performance for trigrams.

[6] NSP, for example, does not implement these measures for trigrams.
[7] Recall considering only bigrams in the gold standards.

6 Conclusions and Future Work

In this paper we presented an inexpensive, language independent hybrid approach for the identification of MWEs, that combines the strenghts of statistical measures with alignment-based information. The results obtained with a Bayesian network classifier confirm the improved performance of the hybrid approach over the individual methods. The use of the alignment information, as well as a larger set of AMs, seem to add robustness to the task, providing enough additional confirmation for the classifier. In addition, the methods seem to perform better for bigrams than trigrams. Further investigation needs to be conducted to identify which factors determine the influence of alignment information on the final performance, since this feature can both introduce noise and improve the performance of the classifier according to the language and size of the ngram. In addition, we also plan on investigating the influence of domain in the performance of these methods, verifying whether domain-specific MWEs are easier to extract than general ones.

Acknowledgements

We want to thank the financial support of the Brazilian agencies FAPESP, CAPES and CNPq. This research has been partly funded by the FINEP/ SEBRAE project COMUNICA.

References

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002)
2. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Grammar of Spoken and Written English. Longman, Harlow (1999)
3. Jackendoff, R.: Twistin' the night away. Language 73, 534–559 (1997)
4. Evert, S., Krenn, B.: Using small random samples for the manual evaluation of statistical association measures. Computer Speech and Language, Special Issue on Multiword Expressions 19(4), 450–466 (2005)
5. Baldwin, T.: The deep lexical acquisition of English verb-particles. Computer Speech and Language, Special Issue on Multiword Expressions 19(4), 398–414 (2005)
6. Caseli, H.M., Villavicencio, A., Machado, A., Finatto, M.J.: Statistically-driven alignment-based multiword expression identification for technical domains. In: Proceedings of the ACL-IJCNLP 2009 Workshop on Multiword Expressions, pp. 1–8 (2009)
7. Villavicencio, A., Caseli, H.M., Machado, A.: Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology, São Carlos, SP (2009)
8. Fazly, A., Cook, P., Stevenson, S.: Unsupervised type and token identification of idiomatic expressions. Computational Linguistics 35(1), 61–103 (2009)

9. Van de Cruys, T., Villada Moirón, B.: Semantics-based Multiword Expression Extraction. In: Proceedings of the ACL 2007 Workshop on Multiword Expressions: A Broader Prespective, Prague, pp. 25–32 (2007)
10. Villada Moirón, B., Tiedemann, J.: Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of the EACL 2006 Workshop on Multiword expressions in a Multilingual Context, Trento, Italy, pp. 33–40 (2006)
11. Ramisch, C., Villavicencio, A., Moura, L., Idiart, M.: Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity. In: Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008), pp. 49–56 (2008)
12. Melamed, I.D.: Automatic Discovery of Non-Compositional Compounds in Parallel Data (1997) eprint arXiv:cmp-lg/9706027
13. Coulthard, R.J.: The application of corpus methodology to translation: the JPED parallel corpus and the Pediatrics comparable corpus. Master's thesis, Universidade Federal de Santa Catarina (2005)
14. Lopes, L., Vieira, R., Finatto, M.J., Martins, D., Zanette, A.: Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. RECIIS - Electronic Journal of communication information and innovation in healthq 3, 76–88 (2009)
15. Procter, P.: Cambridge International Dictionary of English. Cambridge University Press, Cambridge (1995)
16. Banerjee, S., Pedersen, T.: The Design, Implementation and Use of the Ngram Statistics Package. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pp. 370–381 (2003)
17. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the ACL, Hong Kong, China, pp. 440–447 (2000)
18. Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A.: Open-source Portuguese-Spanish machine translation. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 50–59. Springer, Heidelberg (2006)
19. Caseli, H.M., Nunes, M.G.V., Forcada, M.L.: On the automatic learning of bilingual resources: Some relevant factors for machine translation. In: Zaverucha, G., da Costa, A.L. (eds.) SBIA 2008. LNCS (LNAI), vol. 5249, pp. 258–267. Springer, Heidelberg (2008)
20. Caseli, H.M., Ramisch, C., Nunes, M.G.V., Villavicencio, A.: Alignment-based extraction of multiword expressions. Language Resources and Evaluation (2009) (to appear)
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2005)

Out-of-the-Box Robust Parsing of Portuguese

João Silva, António Branco, Sérgio Castro, and Ruben Reis

University of Lisbon
{jsilva,antonio.branco,sergio.castro,ruben.reis}@di.fc.ul.pt

Abstract. In this paper we assess to what extent the available Portuguese treebanks and available probabilistic parsers are suitable for out-of-the-box robust parsing of Portuguese. We also announce the release of the best parser coming out of this exercise, which is, to the best of our knowledge, the first robust parser widely available for Portuguese.

Keywords: Parsing, probabilistic, robust, out-of-the-box, Portuguese.

1 Introduction

The task of robust parsing seeks to address one of the well-known data acquisition bottlenecks in the field of natural language processing.

To its input, typically sentences, a parser associates the corresponding grammatical analysis or representation. Due to the inherent incompleteness of their lexica and grammar rules, hand-coded rule-based parsers are often brittle and eventually fail to deliver any outcome to input sentences whose lexical items are missing in its lexicon or with syntactic constructions not covered by its grammar rules. Robust parsers however are specifically designed to deliver an outcome to any input, even though at the cost of performing at suboptimal accuracy level.

In the last two decades, the field of robust parsing has undergone substantial research effort and progress. The basic technology of probabilistic context free grammars was researched and expanded to a point where the state-of-the-art is consistently scoring in the window 85–90% [1, p. 440] for the most successful solutions—obtained for the most studied natural language, English, with the largest and most widely used dataset, the Penn Treebank, and for its basic task, labeled syntactic constituency analysis.

This research effort has supported the development of a number of packages that implement language-independent approaches, have public releases, and allow to train top performing parsers. Concomitantly, there have been increased efforts to construct treebanks for languages other than English, many of them widely available with public distribution.

These circumstances have permitted to release out-of-the-box, top performing robust parsers for those languages for which such datasets were developed. And this observation brings us to the central issue motivating the present paper.

Though there have been treebanks developed for Portuguese and widely available, to the best of our knowledge, no encompassing and thorough research exercise has been performed to assess whether out-of-the-box, top performing robust

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 75–85, 2010.

parsers for Portuguese can be obtained, and a fortiori no such parser, if possible, has been widely released so far. The present paper is thus guided by the objective of gathering answers to the following leading questions:

Out-of-the-box? To what extent do the software packages currently available, that permit to train a robust parser out of a treebank, are language independent and support a smooth application to Portuguese data? Given their design features, to what extent do the available treebanks permit a smooth application of these packages and the development of an out-of-the-box robust parser?

Top-performing? What range of performing scores is attainable with such parsers? To what extent, if any, does the circumstance that they are out-of-the-box parsers, working in a stand-alone mode with no pre- or post-processing auxiliary tools, affect their chance to attain state-of-the-art performance?

On a par with these objectives, another important goal of this paper is to describe the characteristics and announce the public distribution of the best performing parser resulting from the present research exercise.

In Section 2, we indicate the datasets available for this exercise and their features, and assess their suitability to support the development of out-of-the-box parsers. The software packages available and their suitability are described and assessed in Section 3. In Section 4, we report on the impact on the performance of the parsers that results from the fact that they are used as they come out of their training over the dataset, without resorting to any pre- or post-processing auxiliary tool. Finally, Section 5 is devoted to concluding remarks.

2 Datasets

To the best of our knowledge, there are two treebanks of contemporary Portuguese available, Bosque and CINTIL Treebank. In this section, we ponder their suitability for training out-of-the-box parsers.

2.1 Bosque

Bosque is a treebank of newspaper articles (Brazilian and European Portuguese) that has been automatically annotated by the PALAVRAS parser and subsequently manually revised.[1]

Syntactic heads are explicitly marked, as well as arguments and other modifiers (with indication on whether they are pre- or post-modifiers). Trees are also annotated with tags for syntactic functions at the phrase level. Named entities and closed class multi-word expressions appear as one syntactic unit as their parts are concatenated (separated by underscore) into a single token. A sample, taken from [2], is shown in Figure 1.

In [2], the authors report on the use of Bikel's package [3] to train a parser over Bosque. Most of that paper actually consists in the description of the many difficulties that the authors need to cope with when adapting the tree format of Bosque to a format suited for training the parser.

[1] Distribution at `http://www.linguateca.pt`

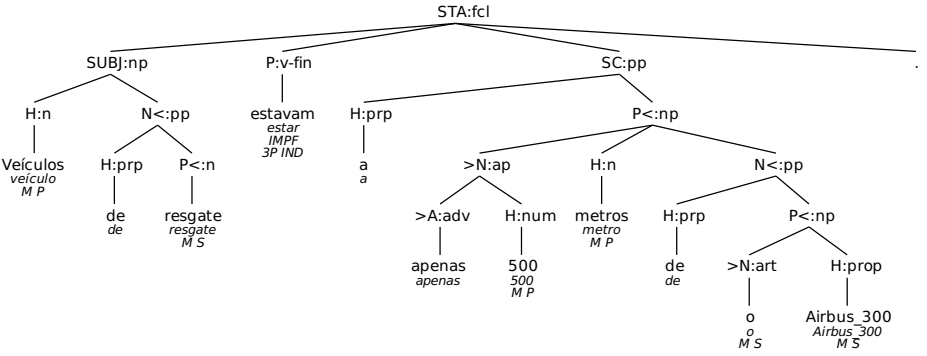

Fig. 1. A sample of Bosque

A dataset of 1,877 sentences was held-out for testing and the remaining 7,497 sentences were used for development. An out-of-the-box Bikel parser achieved a PARSEVAL f-score[2] of 36.3%, which can be taken as a baseline. Another, improved parser was then obtained by refining the training parameters while running 10-fold cross-validation tests over the development dataset. This parser was extended with Portuguese-specific head-finding rules, and with inflectional and derivational features for classifying unknown Portuguese words. Additionally, the annotation of Bosque was enriched to support a better performance of the parser. This improved parser achieved a PARSEVAL f-score of 63.2%.

2.2 CINTIL Treebank

The CINTIL treebank was produced from the output of LXGram, a deep linguistic processing grammar [4] by manually selecting the correct parse for a sentence from among all the possible parses that are delivered by the grammar [5].[3]

Like Bosque, constituents are marked with syntactic functions at the phrase level. Although heads are not explicitly indicated by a tag, they can be directly picked from the X-bar syntactic structure, which this treebank adheres to. Named entities and multi-word expressions have their internal constituency explicitly represented. For instance, `Hong Kong` is represented as `(N' (N Hong) (N Kong))`. In addition, it is possible to access a variety of morphological information, like the lemma of the words. A sample is shown in Figure 2.

As there were no published results on training a parser over CINTIL, we ran an experiment to assess its suitability for the exercise reported in this paper.

In its current version, CINTIL has 1,204 sentences, mostly from newspaper articles (European Portuguese). Since the sentences are already represented in the *de facto* standard Penn Treebank format, there was no need to adapt it for training with Bikel's parser, which is the one we chose to ease comparison with

[2] See Section 3.5 for more details on evaluation metrics.
[3] Distribution at `http://nlx.di.fc.ul.pt/lxgram/`

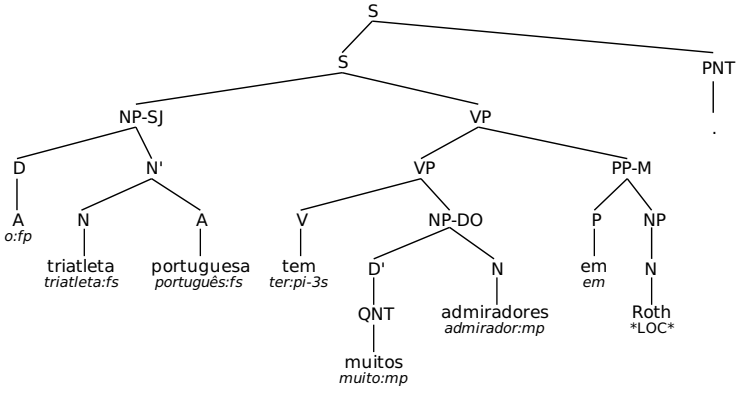

Fig. 2. A sample of CINTIL Treebank

the results from [2]. However, keeping in mind that our interest is in testing to what extent the parser can be used out-of-the-box, we did not adapt the parser for Portuguese.

We ran 10-fold cross-evaluation, by iteratively training over a randomly selected 90% portion of the treebank and testing over the remaining 10% to get a baseline score. The average labeled PARSEVAL f-score obtained was 76.18%.

This represents an improvement of more than 20% over the best result (63.2%) in [2] with Bosque. Significantly, when comparing the two baseline scores, the improvement over the Bosque-supported parser is more than 100%.

Despite the fact that the results in [2] were obtained with a parser fine-tuned for Portuguese, and a much larger training dataset, Bikel's parser fared much better over CINTIL than over Bosque.[4] Accordingly, we opted for using the CINTIL treebank for the remainder of this paper given this dataset not only ensures better parser performance as it adheres to the *de facto* standard format and can actually be used out-of-the-box.

3 Parsers

Having cleared the issue on which dataset, if any, would support out-of-the-box parsing, we proceeded with a detailed experimentation of several packages.

In order to gather as many as possible freely available packages that were good candidates for training out-of-the-box parsers we searched the ACL software repository[5] and collected the replies to a post (September 2009) on CorporaList[6] querying on this type of packages.

The following five appeared as good candidates: Bikel [3], Stanford [8], Berkeley [1], BitPar [9] and Reranking/Charniak [10] parsers.

[4] It is known that annotation errors adversely affect the performance of classifiers [6,7].

[5] http://www.aclweb.org/

[6] http://gandalf.aksis.uib.no/corpora/

3.1 Bikel

Bikel's parser [3] is a language-independent, head-driven, statistical parsing engine. Language-independence is achieved through "language packages" which encapsulate all the procedures that are specific to a language or treebank. The package comes with specific support for English, Chinese and Arabic.

For English, running in a mode emulating Model 2 of Collins' parser [11], and when evaluating over sentences up to 40 words in length from Section 00 of the Penn Treebank, it performs at 90.01% PARSEVAL f-score [3, p. 181].

3.2 Stanford

The Stanford package [8] allows training a factored parser, which models phrase-structure and lexical dependencies separately. Phrase-structure is modeled using a probabilistic context-free grammar similar to that in [11], which includes features such as parent annotation of nodes and English-specific splits of certain features.[7] The model for lexical dependencies takes into account direction, distance and valence between a constituent and its dependents. The probability of a tree is then given by the product of the probabilities that the phrase-structure model and the lexical dependencies model assign to that tree. This software package comes with specific support for English, Chinese, Arabic and German.

For English, the parser scores 86.7% PARSEVAL f-score over sentences up to 40 words in length from Section 23 of the Penn Treebank [8, p. 8].

3.3 Berkeley

The Berkeley parser [1] iteratively refines a base X-bar grammar by repeatedly splitting and merging non-terminal symbols. For instance, the symbol NP could be split into the symbols NP_0 (NP in subject position) and NP_1 (NP in object position). The base X-bar is obtained directly from the training dataset by a binarization procedure that introduces left-branching X-bar nodes in order to ensure that each node has two children. In each iteration, every symbol is split in two. Since this would quickly become unwieldy, a merging step checks which splits can be undone without a great loss in likelihood.

For English, this parser achieves 90.15% PARSEVAL f-score over the sentences up to 40 words in length from Section 23 of the Penn Treebank [1, p. 440].

3.4 Other Packages

BitPar [9] is a bit-vector implementation of the CYK parsing algorithm. The compact bit-vector representation greatly minimizes the memory and runtime costs during parsing, allowing BitPar to more easily represent the parse forest of all possible analysis of a sentence.

Reranking/Charniak [10] uses a maximum entropy model to select the best parse from the 50-best parses delivered by Charniak's coarse-to-fine n-best generative parser [12].

[7] Infinitive VPs are marked, the tag for preposition is split into subtypes, etc.

Like the three packages presented above, these two packages were originally developed for English. However, they eventually revealed to be much more language dependent. A great deal of their source code has hard-coded features for English and in particular for the English specific tag set and annotation conventions used in the Penn Treebank. Moreover, to a considerable extent, for the training phase of the parser to be induced, they resort to rule-based preprocessing of this dataset in order to extract or make available linguistic features specific of English but not explicitly present in the Penn treebank.

Their adaptation to train a parser for Portuguese, from whatever training dataset, would thus require a very large effort to change the source code. Accordingly, we realized that BitPar and Reranking/Charniak cannot be considered for *out-of-the-box* parsing of Portuguese. For this reason, they ended up not being further used in the research exercise described here.

3.5 Evaluation Results

As there is nowadays a specific and productive research area on dependency parsing, our focus here will be on parsers for syntactic constituency alone. Accordingly, the evaluation results reported from this point onwards were obtained with one of the available versions of CINTIL that contains constituency information and no syntactic function tags.

We maintain the same evaluation methodology used in the quick experiment ran for assessing the datasets in Section 2. We use 10-fold cross-evaluation, iteratively training over a random 90% portion of the treebank and evaluating over the remaining 10%, and averaging the scores. The following performance metrics were computed:

Parseval is the classic metric of bracketing correctness [13]. We use a subsequent adaptation [14] that also takes into account whether the constituent label is correct. This metric provides labeled recall and labeled precision, from which the labeled f-score[8] is calculated.

Evalb is similar to PARSEVAL in that it is also a metric of bracketing correctness, providing, among other metrics, labeled f-score [15]. The most important difference, however, is that pre-terminals nodes are taken separately, allowing for a separate measure of tagging (part-of-speech) accuracy.

LeafAncestor is argued to mirror more closely our intuitive notions of parsing accuracy [16]. Instead of looking at bracketing correctness, it checks the *lineage* of terminal elements, i.e. the sequence of nodes connecting a terminal element to the root of its tree.

Three parsers were trained with each one of the packages described above that were suitable for out-of-the-box parser induction. Each one of the resulting parsers was evaluated along the metrics just outlined. The results are summarized in Table 1.

[8] F-score is defined as the harmonic mean of precision and recall: $f = \frac{2pr}{p+r}$

Table 1. Performance scores of parsers for Portuguese

	$f_{Parseval}$	f_{Evalb}	POS acc.	LeafAnc.
Bikel	84.97%	73.08%	88.82%	90.48%
Stanford	88.07%	78.75%	92.91%	91.87%
Berkeley	89.33%	80.79%	91.62%	93.72%

The first point worth noting is that the results obtained with the PARSEVAL metric are within the 85–90% window that has been consistently obtained for top-performing parsers, running over English.

These are extremely encouraging scores, specially if one takes into account that they were obtained over a small treebank and with out-of-the-box parsers, with no effort applied to adapt them to Portuguese or to the training dataset.

Moreover, to our knowledge, these are the best published results for the probabilistic parsing of Portuguese.

Coming now to a comparison among parsers, Berkeley's parser has the best overall performance, which is in line with it also being one of the best parsers for English [1, p. 440]. A possible contribution for its better score may come from the fact that this is possibly the least language-dependent parser, since it does not use head-finder rules. As such, it was not as severely penalized as the other parsers for running out-of-the-box over a language that is not English.

4 Assessing Enhanced Performance

In this section we seek to assess, at least in part, to which extent the performance of these out-of-the-box parsers can be expected to be improved.

Improvement will likely come by pursuing two lines of action, each moving away from the out-of-the box status of the parsing task supported by the induced parsers. On the one hand, each parser can be put under systematic testing so that their parameters can be progressively fine-tuned in order to set a running configuration that support optimal performance. On the other hand, each parser can be aided by pre-processing modules so that the resulting parsing pipeline has better performance than the stand-alone parser.

In this section, we will concentrate on the second line of action as it can be implemented under a straightforward approach. It is immediate to test the *upper bound* for the improvement of performance in parsing due to the contribution of a pre-processing module. It is enough to use the correct data already present in the treebank as if they were the output of a perfect, 100% accurate pre-processing module.

Taking the values from Table 1 as a baseline, we test different ways of pre-processing the input delivered to the parser, typically in view of reducing data-sparseness, and measure how that improves the overall parsing performance. We test each of the following pre-processing procedures:

Lemmatization. Portuguese has a rather rich morphology, in particular in what concerns verbal inflection. Abstracting away from the variations caused by inflection should help mitigate the problems caused by data-sparseness.

To test this, the parser is trained and evaluated over a treebank where each noun, adjective and verb has been replaced by its lemma.

Named-entity recognition. Named-entities and multi-word expressions are a well-known problem for natural language processing [17]. Recognizing these sequences of words as an entity is very helpful for the parsing process since the parser can handle them as being a syntactic atom.

In CINTIL, certain types of named entities (NE) are marked and classified.[9] This allowed us to create a variant of the treebank, we termed NE-joined, where those named entities appear as one syntactic unit with their parts concatenated (separated by a plus sign) into a single token. For instance, `Hong Kong` is represented as the single node `(N Hong+Kong)`.

In addition, given that the named entities in CINTIL are semantically classified, we created another variant, we termed NE-sem, where each named entity expression was replaced by its semantic type. For instance, `Hong Kong` is represented as `(N *LOC*)` (for location), while `David Maia` is represented as `(N *PER*)` (for person).

POS tagging. As can be seen in Table 1, the POS tagging accuracy of each parser is below what is attainable by a state-of-the-art, dedicated POS tagger.[10] This is likely due to the small size of the training treebank we are using here.

It is worth noting that even if the parser could assign POS tags with as much accuracy as a POS tagger, such pre-processing by a dedicated, stand-alone POS tagger is often used and coupled to robust parsers as a way to reduce ambiguity and speed up the parsing process [20, p. 240].

To test a contribution by an hypothetical optimal POS tagger, we assume that the input to the parser has been previously annotated by a faultless tagger. This is quite easy to simulate since the parsers we are using already support a mode where they accept POS-tagged input. All that we need to do is to keep the pre-terminal tags from the treebank in the text that is going to be annotated.

The results obtained when pipelining the above different pre-processing modules with each parser are displayed in Table 2. Each pre-processing procedure is able to improve on the baseline scores, albeit by different amounts.

The relative ranking of the parsers (Berkeley > Stanford > Bikel) is generally preserved across the different experiments. Also, for each parser, the relative differences in the scores obtained through each evaluation metric ($f_{Parseval}$, f_{Evalb}, POS accuracy and LeafAncestor) are also generally preserved across the experiments.

[9] Expressions for persons, organizations, locations, events, works (e.g. movies, books, paintings, etc.) and miscellaneous.

[10] For instance, for Portuguese, an accuracy of 97% has been obtained for POS tagging also with the CINTIL tagset [18,19].

Table 2. Performance scores for each pre-processing procedure

		baseline	lemmas	NE-joined	NE-sem	POS
Bikel	$f_{Parseval}$	84.97%	87.68%	85.16%	87.68%	92.34%
	f_{Evalb}	73.08%	74.35%	73.48%	74.71%	79.72%
	POS acc.	88.82%	92.03%	90.26%	91.13%	n.a.
	LeafAnc.	90.48%	91.79%	90.49%	91.03%	94.06%
Stanford	$f_{Parseval}$	88.07%	89.49%	88.69%	88.91%	93.69%
	f_{Evalb}	78.75%	80.74%	79.63%	80.41%	84.60%
	POS acc.	92.91%	94.05%	94.48%	94.63%	n.a.
	LeafAnc.	91.87%	92.81%	92.06%	92.27%	94.97%
Berkeley	$f_{Parseval}$	89.33%	90.21%	89.55%	90.34%	95.61%
	f_{Evalb}	80.79%	81.11%	81.60%	83.15%	87.42%
	POS acc.	91.62%	92.82%	92.74%	93.29%	n.a.
	LeafAnc.	93.72%	94.63%	94.16%	94.36%	96.55%

Given this, and for the sake of simplicity, the results in Table 2 are commented taking the Berkeley parser and the $f_{Parseval}$ metric as a reference. This is also the parser that is being released as an outcome of the present research exercise, at `http://lxparser.di.fc.ul.pt/`

NE-joined provides the least improvement (89.55%). While it does simplify the phrase-structure, it does not help in reducing the lexicon.

If the entities are replaced by their types, as in NE-sem, the improvement becomes more apparent (90.34%), gaining 1 percentage point over the baseline. When every entity is mapped into a small set of semantic types, it greatly compacts the lexicon and any new named entity, when classified with a semantic type seen in training, will not be considered an unknown word.

Using lemmas instead of inflected forms brings about a similar amount of improvement (90.21%). This improvement, however, was not caused by better handling of named entities. Therefore, assigning lemmas as a pre-processing step could be applied together with NE-sem for an even bigger increase in scores.

Providing the parsers with correct POS yields by far the largest improvement, of 6 percentage points over the baseline, bringing the score to 95.61%.

5 Conclusions

With the experiment reported in this paper we showed that it is possible to apply out-of-the-box, state-of-the-art software packages for training parsers for Portuguese. More importantly, not only is this possible, as the results are in line with those obtained for English with top-performing parsers given the best treebank available, namely the CINTIL Treebank.

Given that the results described above were obtained over a modestly sized treebank and with out-of-the-box parsers it will be possible to obtain new parsers

with improved performance. The improvements achieved by experimenting with different pre-processing procedures confirm this.

The best parser for Portuguese, obtained with the Berkeley package, is released at http://lxparser.di.fc.ul.pt/

As future work, our goal will be to retrain this parser over upcoming and larger versions of CINTIL Treebank and find the configuration of parameters that eventually support its optimal performance.

References

1. Petrov, S., Barrett, L., Thibaux, R., Klein, D.: Learning accurate, compact, and interpretable tree annotation. In: Proceedings of the 44th ACL, pp. 433–440 (2006)
2. Wing, B., Baldridge, J.: Adaptation of data and models for probabilistic parsing of Portuguese. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 140–149. Springer, Heidelberg (2006)
3. Bikel, D.: Design of a multi-lingual, parallel-processing statistical parsing engine. In: Proceedings of the 2nd Human Language Technology Conference (2002)
4. Branco, A., Costa, F.: A computational grammar for deep linguistic processing of Portuguese: LXGram, version A.4.1. Technical Report DI-FCUL-TR-08-17, University of Lisbon (2008)
5. Branco, A., Costa, F.: A deep linguistic processing grammar for portuguese. In: Pardo, T.A.S., et al. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 83–86. Springer, Heidelberg (2010)
6. Padró, L., Màrquez, L.: On the evaluation and comparison of taggers: The effect of noise in testing corpora. In: Proceedings of the 17th COLING, pp. 997–1002 (1998)
7. Dickinson, M., Meurers, D.: Detecting inconsistencies in treebanks. In: Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (2003)
8. Klein, D., Manning, C.: Fast exact inference with a factored model for NLP. Advances in Neural Language Processing Systems 15, 3–10 (2003)
9. Schmid, H.: Efficient parsing of highly ambiguous context-free grammars using bit vectors. In: Proceedings of the 20th COLING, pp. 162–168 (2004)
10. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and maxent discriminative reranking. In: Proceedings of the 43rd ACL (2005)
11. Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. PhD thesis, University of Pennsylvania (1999)
12. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American Chapter of the ACL, pp. 132–139 (2000)
13. Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Marcus, M., Santorini, B.: A procedure for quantitatively comparing the syntactic coverage of English grammars. In: Proceedings of the Workshop on the Evaluation of Parsing Systems, pp. 306–311 (1991)
14. Magerman, D.: Statistical decision-tree models for parsing. In: Proceedings of the 33rd ACL, pp. 276–283 (1995)
15. Sekine, S., Collins, M.: Evalb website, http://nlp.cs.nyu.edu/evalb/
16. Sampson, G., Babarczy, A.: A test of the leaf-ancestor metric for parse accuracy. Natural Language Engineering 9(4), 365–380 (2003)

17. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Proceedings of the 3rd Conference on Intelligent Text Processing and Computational Linguistics, pp. 1–15 (2002)
18. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: Proceedings of the 4th Language Resources and Evaluation Conference (LREC), pp. 507–510 (2004)
19. Silva, J.: Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, University of Lisbon (2007); Published as Technical Report DI-FCUL-TR-07-16
20. Bangalore, S., Joshi, A.: Supertagging: An approach to almost parsing. Computational Linguistics 25(2), 237–265 (1999)

LXGram: A Deep Linguistic Processing Grammar for Portuguese

Francisco Costa and António Branco

Universidade de Lisboa
fcosta@di.fc.ul.pt
Antonio.Branco@di.fc.ul.pt

Abstract. In this paper we present LXGram, a general purpose grammar for the deep linguistic processing of Portuguese that delivers high precision grammatical analysis and detailed meaning representations. We present the main design features and evaluation results on the grammar's coverage as well as its ability to produce correct grammatical analyses.

Keywords: Deep linguistic processing, unification grammars, parsing.

1 Introduction

We present what is, to the best of our knowledge, the first general purpose grammar, distributed under an open-source license, for the deep linguistic processing of Portuguese, that delivers a thorough and principled linguistic analysis of sentences, including their formal semantic representation. LXGram is part of the DELPH-IN Consortium, an international group of researchers working on deep linguistic processing for a variety of languages. In Section 2 the main design features of this grammar are described. Evaluation results are presented in Section 3, based on an experiment consisting of parsing spontaneous text that was not seen during the development phase. Finally, we conclude in Section 4.

2 Scope and Design Features

LXGram is based on hand coded linguistic generalizations supplemented with a stochastic model for ambiguity resolution of parses. It follows the grammatical framework of Head-Driven Phrase Structure Grammar (HPSG [1]), one of the most prominent linguistic theories used in natural language processing.

HPSG is a linguistic framework for which there is a substantial amount of published work. This allows for the straightforward implementation of well known grammatical analyses, which are linguistically grounded and have undergone scientific scrutiny. It also has a positive impact in reusability and extendibility, because more people can understand it immediately. The HPSG literature has produced very accurate analyses of long distance dependencies, and a general strong point of computational HPSGs, among many others, is precisely the implementation of this key phenomenon of natural language syntax.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 86–89, 2010.

HPSGs associate grammatical representations to natural language expressions, including the formal representation of their meaning. Like several other computational HPSGs, LXGram uses Minimal Recursion Semantics (MRS [2]) for the representation of meaning. An MRS representation is a description of a set of possible logic formulas that differ only in the relative scope of the relations present in these formulas. In other words, it supports scope underspecification. Semantic representations provide an additional level of abstraction, as they completely abstract word order and language specific grammatical restrictions. Additionally, the MRS format of semantic representation that is employed is well defined in the sense that it is known how to map between MRS representations and formulas of second order logic, for which there is a set-theoretic interpretation. Because of space limitations, it is impossible to provide a detailed account of MRS representations here. [2] provides a very clear description of it.

LXGram is developed in the Linguistic Knowledge Builder (LKB) system [3], an open-source development environment for constraint-based grammars. This environment provides a GUI, debugging tools and very efficient algorithms for parsing and generation with the grammars developed there. Several broad-coverage HPSGs have been developed in the LKB; the largest ones are for English [4], German [5] and Japanese [6]. The grammars developed with the LKB are also supported by the PET parser [7], which allows for faster parsing times due to the fact that the grammars are compiled into a binary format.

LXGram is in active development, but it already supports a wide range of linguistic phenomena, such as long distance dependencies, coordination, subordination, modification and many subcategorization frames. and its lexicon contains 25000 entries. At the moment, LXGram contains 64 lexical rules, 101 syntax rules, around 850 lexical leaf types (determining syntactic and semantic properties of lexical entries), and 35K lines of code (excluding the lexicon). LXGram supports both European and Brazilian Portuguese. It contains lexical entries that are specific to either of them, and it covers both European and Brazilian syntax ([8]). A statistical disambiguation model was also trained, in order to automatically select the most likely analysis of a sentence when the grammar produces multiple solutions. This model was trained from a dataset comprising 2000 sentences of newspaper text, using a maximum entropy algorithm. The linguistic analyses that are implemented in the grammar are documented in a report that is updated and expanded with each version of the grammar. The grammar is available for download at `http://nlx.di.fc.ul.pt/lxgram`, together with its documentation.

3 Evaluation

We conducted an experiment to assess the coverage of LXGram's current version on spontaneous text. We used a subset of the Portuguese Wikipedia, as well as part of two publicly available corpora: CETEMPúblico and CETENFolha, which contain newspaper text from "O Público" and "Folha de São Paulo" respectively.

Table 1. Evaluation data and grammar coverage

	Wikipedia	CETEMPúblico	CETENFolha	Total
Sentences	66304	30000	30000	126304
Avg. words/sentence	25	27.5	18.6	24
Avg. seconds/sentence	2.6	4.7	2	3
Parsed sentences	20995	8455	11173	40623
Parsed percent	32%	28%	37%	32%
Avg. readings/parsed sentence	67	87	75	73
Avg. words/parsed sentence	11	13	11	11

The Wikipedia corpus consists of a selection of articles downloaded from the Portuguese Wikipedia, by following the links on the page "Artigos Destacados" ("Featured Articles"). 318 pages were obtained in this way and preprocessed in order to remove HTML markup.

As for the two newspaper corpora, we randomly selected 30000 sentences from each of them. We removed all XML-like tags (such as `<s>` for sentence boundaries) but kept each sentence in its own line, to be processed separately.

Before parsing these texts, we fed each sentence to a part-of-speech tagger [9] and a morphological analyzer [10,11], in order to handle out-of-vocabulary words and to constrain the parser search space. For each sentence, we kept the 250 most likely analyses, as determined by the disambiguation model presented.

LXGram was able to successfully parse 32% of the sentences in the Wikipedia sample, 28% of the CETEMPúblico sentences and 37% of the CETENFolha sample. Table 1 summarizes our results, using a 2,5 GHz Intel processor.

The fact that the average length of parsed sentences is very similar for both CETEMPúblico and CETENFolha indicates that the large difference in coverage on these two datasets may be more related to average sentence length than to differences between European and Brazilian Portuguese.

When comparing these results to the other computational HPSGs, it should be mentioned that [12] reports values of 80.4% coverage on newspaper text for the English grammar, 42.7% for the Japanese grammar and 28.6% for the German grammar.[1] All of these grammars have been in development for over 15 years now, and they are all substantially older than LXGram, with 4 years of development. A more recent HPSG Grammar, for Spanish—a language quite similar to Portuguese—, is the Spanish Resource Grammar [13], approximately as old as LXGram. The SRG is reported in [12] to have a coverage of 7.5%.

In order to assess the accuracy of the grammar, we inspected a sample with the first 50 parsed sentenced in the CETENFolha subcorpus. 20 sentences were correctly parsed, and furthermore the preferred reading was the one chosen by the disambiguation model. Another 10 sentences also received a correct parse, although the disambiguation model did not choose the preferred reading for these sentences as the best one. From the 20 sentences that did not receive a correct

[1] However, the German grammar has close to 40% coverage on newspaper text (personal communication by Berthold Crysmann) using a more recent method to integrate information coming from preprocessing tools.

parse, 12 sentences were affected by errors from the part-of-speech tagger or the morphological analyzer, and 8 of them were due to genuine limitations in the grammar or the disambiguation model (for instance, lack of some subcategorization frames for some words in the lexicon).

4 Conclusions

We presented a resource grammar for Portuguese which is based on HPSG. To the best of our knowledge, it is the only deep linguistic parser for Portuguese that outputs fine-grained semantic representations.

The grammar keeps being developed, but it already features interesting coverage of unrestricted text, achieving over 30% coverage on newspaper text, which is usually hard to parse by symbolic systems. Additionally, a sample of those parsed sentences was manually evaluated, and it indicates that a good portion of the parsed sentences got a correct representation (60%) and are disambiguated correctly (40%), while 60% the of parse failures were due to preprocessing errors.

Our ongoing work includes grammar expansion, and also the creation of a treebank of sentences parsed with the grammar and manually disambiguated.

References

1. Pollard, C., Sag, I.: Head-Driven Phrase Structure Grammar. Chicago University Press and CSLI, Stanford (1994)
2. Copestake, A., Flickinger, D., Sag, I.A., Pollard, C.: Minimal Recursion Semantics: An introduction. Journal of Research on Language and Computation 3(2–3) (2005)
3. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI, Stanford (2002)
4. Copestake, A., Flickinger, D.: An open-source grammar development environment and broad-coverage English grammar using HPSG. In: LREC 2000, Athens (2000)
5. Crysmann, B.: Local ambiguity packing and discontinuity in German. In: ACL Workshop on Deep Linguistic Processing, Prague (2007)
6. Siegel, M., Bender, E.M.: Efficient deep processing of Japanese. In: The 3rd Workshop on Asian Language Resources and International Standardization. Coling 2002 Post-Conference Workshop, Taipei, pp. 31–38 (2002)
7. Callmeier, U.: PET— A platform for experimentation with efficient HPSG processing techniques. Natural Language Engineering 6(1), 99–108 (2000)
8. Branco, A., Costa, F.: Accommodating language variation in deep processing. In: King, T.H., Bender, E.M. (eds.) GEAF 2007, pp. 67–86. CSLI, Stanford (2007)
9. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: LREC 2004, pp. 507–510. ELRA, Paris (2004)
10. Branco, A., Silva, J.: Very high accuracy rule-based nominal lemmatization with a minimal lexicon. In: APL XXI, Lisbon (2007)
11. Branco, A., Nunes, F., Costa, F.: The processing of verbal inflection ambiguity: Characterization of the problem space. In: APL XXI, Lisbon (2007)
12. Zhang, Y., Wang, R., Oepen, S.: Hybrid multilingual parsing with HPSG for SRL. In: CoNLL 2009, Boulder, USA (2009)
13. Marimon, M., Bel, N., Espeja, S., Seghezzi, N.: The Spanish Resource Grammar: pre-processing strategy and lexical acquisition. In: ACL Workshop on Deep Linguistic Processing, Prague (2007)

InferenceNet.Br: Expression of Inferentialist Semantic Content of the Portuguese Language

Vladia Pinheiro[1], Tarcisio Pequeno[2], Vasco Furtado[2,3], and Wellington Franco[2]

[1] Departamento de Ciências da Computação – Universidade Federal do Ceará
Campus do Pici-UFC, Fortaleza, Ceará, Brasil
[2] Mestrado em Informática Aplicada – Universidade de Fortaleza (UNIFOR)
Av. Washington Soares, 1321, Fortaleza, Ceará, Brasil
[3] ETICE – Empresa de Tecnologia da Informação do Ceará
Av. Pontes Vieira 220, Fortaleza, Ceará, Brasil
vladia@lia.ufc.br, {vasco,tarcisio}@unifor.br,
jwellingtonfranco@gmail.com

Abstract. Often, the information necessary for a complete understanding of texts is implicit, which requires drawing inferences from the use of concepts in the linguistic praxis. We consider that the usual semantic reasoners of natural language systems face difficulties in capturing this knowledge, due mainly to the lack of linguistic-semantic resources that support reasoning of this nature. This paper presents a new linguistic resource that expresses semantic-inferentialist knowledge for the Portuguese language – InferenceNet.Br – containing a base of concepts and a base of sentence patterns. These bases provide content for a top layer of semantic reasoning in natural language systems, where semantic relations are considered according to their roles in inferences, as premises or conclusions. This linguistic resource was used in a system for extracting information about crime, and the results of this proof of concept are discussed.

Keywords: Linguistic Resource, Semantic, Portuguese Language.

1 Introduction

Often, the information necessary for a complete understanding of texts by Natural Language Processing (NLP) systems is implicit, which requires drawing inferences from the use of concepts[1] in the linguistic praxis. For instance, when we read the news "João murdered his wife by shooting her to death after an argument on Solon Pinheiro Street", we are able to refute an assertion that the type of weapon used in the crime was a "cold weapon" (non-firearm) and to argue that the type of crime was "homicide". This is possible because we, users of natural language, know the conditions in which the concepts "to shoot" and "to murder" can be used and the commitments we

[1] "Concept" is the semantic content expressed by a term (n-grams) of a natural language. So when we refer to "concept" we are expressing the content or semantic value of a term in natural language.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 90–99, 2010.

make when using them in a sentence. Such inferences do not come only from individual content of the concepts, but are generated by a holistic thinking about using these concepts in a particular sentence structure.

We consider that the usual semantic reasoners of natural language systems face difficulties in capturing this knowledge, mainly because of the lack of linguistic-semantic resources that support a complete understanding of concepts and sentences, considering their use within the linguistic practice. Simple taxonomies or more complex ontologies – traditionally used in the semantic analysis of sentences – seek to give shape to the content of concepts expressing the organization of the universe into genera, families and species, and through relationships that express the physical and even the functional and causal characteristics represented by the concepts. In short, a representation of the world *a priori* is sought in order to express the content of concepts, disregarding the linguistic praxis (i.e. the uses of natural language). We argue that it is within the linguistic practice that the circumstances to use a word and the consequences thereof can be grasped, and – by disregarding them – much of what could be inferred is lost. Even simple ontologies that define taxonomies, typical lexical-semantic bases such as WordNet[1,2,3] and CYC[4], or more complex semantic bases, expressing causal knowledge, functional knowledge, etc., such as FrameNet[5] and ConceptNet[6], should be considered from the inferential and pragmatic viewpoint.

Philosophers such as Sellars[7], Dummett[8] and Brandom[9] proposed semantic-inferentialist theories, which present a different approach to define the content of concepts. According to these theories, the expression of the semantic value of concepts should favor the role that such concepts play in reasoning, as premises and conclusions, rather than their referents and representational characteristics. Sellars asserts that "to grasp or understand such a concept is to have practical mastery over the inferences it is involved in – to know what follows from the applicability of a concept, and what it follows from" [7, p.48]. Following this inferentialist vision, Pinheiro et AL.[10,11,12] propose the *Semantic Inferentialism Model* (SIM) – a computational model that defines the main requirements for expression and reasoning about semantic knowledge of a natural language – to enable NLP systems to carry out inferences considering the pragmatic bias of the language in a systematic way.

In this work, we present a new linguistic resource that expresses semantic-inferentialist knowledge for the Portuguese language – InferenceNet.Br – and its collaborative portal (www.inferencenet.org), which allows the dissemination, use and development of this resource by the community. InferenceNet.Br consist of two bases: the conceptual base and sentence-patterns base, which provide content for a top layer of semantic reasoning in NLP systems, where semantic relations are considered according to their roles in inferences, as premises or conclusions. The conceptual base originated and was evolved from the American commonsense base – ConceptNet[6]. The Sentence-Patterns Base was generated from a linguistic corpus, CRIMES2008. InferenceNet.Br was used in an Information Extraction (IE) system for extracting information about crime, described on the Web, and it was possible to carry out a proof of concept of the advantages of using this resource in NLP tasks.

2 InferenceNet

One motivation for a new linguistic resource is the lack of linguistic resources with large scale semantic knowledge for the Portuguese language. Lexical-semantic bases in Portuguese – for example, WordNet.Pt [2], WordNet.Br [3] and OMCS-Br[2] – paradigmatically mirrored on their counterparts in English, are currently under construction and as yet unavailable, or are restricted to specific domains. Another motivation is the non-existence of a linguistic resource with inferentialist semantic knowledge, either for Portuguese or other languages. InferenceNet.Br contains two semantic bases included in the SIM: the *Conceptual Base* and the *Sentence-Patterns Base.*

The *Conceptual Base* contains the inferential content of concepts of a natural language, defined and agreed upon in a community or area of knowledge. According to the inferentialist view, the content of a concept must be expressed, becoming explicit, through the use of it (the concept) in inferences, as premises or conclusions of reasoning. Moreover, what determines the use of a concept in inferences or potential inferences in which this concept may participate are: (i) its pre-conditions or premises of use: what gives someone the right to use the concept and what could exclude such a right, serving as premises for utterances and reasoning; and (ii) its post-conditions or conclusions of use: what follows or what are the consequences of using the concept, which let one know what someone is committed to by using a particular concept, serving as conclusions from the utterance *per se* and as premises for future utterances and reasoning. The *Sentence-Patterns Base* contains generic sentences, called sentence-patterns, which have a syntactic structure and function as templates, whose slots can be filled in with terms from a natural language. For example, '*X ser assassinar por Y*' ['*X be murder by Y*'] follows the structure <sentence> ::= <NP> <VP> <PP>[3]. Its noun phrase (NP), represented by *X*, can be filled in by '*uma mulher*' ['*a woman*'] and the prepositional phrase (PP), represented by '*por Y*' ['*by Y*'] can be complemented by '*seu amante'* ['*her lover*'], generating the sentence '*uma mulher ser assassinar por seu amante*' ['*a woman be murder by her lover*']. The importance of a sentence-patterns base for semantic analysis is the following: something that can be inferred by reading a sentence doesn't come directly and uniquely, at least in an efficient manner, from the content of the concepts of the sentence, but rather from these concepts together and articulated under a particular sentence structure. This base provides a holistic approach for grasping the meaning of sentences: to define the semantic value of a concept, one must consider other related concepts in the sentence and how all of them are structured.

2.1 Construction of InferenceNet.Br's Conceptual Base

The process of constructing InferenceNet.Br's conceptual base included the following steps:

[2] http://www.sensocomum.ufscar.br, accessed on 09/29/2009.
[3] NP – Noun Phrase; VP – Verb Phrase; PP - Prepositional Phrase.

1. **Translation of the American ConceptNet into the Portuguese language.** With financial resources of the research project[4], the American ConceptNet[6] (v2.1) was translated by a team of specialized human translators. We considered this the only alternative due to a lack of Portuguese-language resources available on a large scale, which could serve as a basis for the construction of a semantic-inferentialist base. In this paper, we refer to the translated ConceptNet base as "ConceptNet.Tr".
2. **Formalization of the base.** The conceptual basis is a directed graph *Gc (V,E)* where V = non-empty set of concepts c_i (vertices of the graph) and E = set of edges labeled by a variable that indicates (binary) relationships between two concepts in V. The types of relationships between concepts are pre-defined as two types: pre-condition or post-condition for use of the origin concept of the relationship. As this is a digraph, there are two functions s and t, where s: $E \rightarrow V$ is a function that associates an edge of E with its origin concept in V and t:$E \rightarrow V$ is a function that associates an edge of E with its target concept in V. A concept in V can be represented in the conceptual base by simple terms (unigrams), which belong to open classes of words – nouns, verbs, adjectives, adverbs (e.g.: 'crime', 'death'; in Portuguese: '*crime*' and '*morte*'); or by expressions consisting of more than one term (n-grams), whether or not linked to words from the closed classes: prepositions and conjunctions (e.g. 'mathematical proof', 'bank mugging'; in Portuguese: '*prova de matemática*' and '*saidinha bancária*').
3. **Generation of Inferential Content based on ConceptNet.Tr.** We needed to express pre-conditions and post-conditions for use of concepts, not just a base with traditional semantic relationships. In order to do so, we implemented two heuristics for generating inferential content based on ConceptNet.Tr:

   - For each type of ConceptNet.Tr relationship of the form *type_rel*(c_1,c_2) (based on the category of relationship), the type of inferential relationship it would express was analyzed: a pre-condition or a post-condition for use of the origin concept c_1. Table 1 shows the type of inferential relation expressed by each type of relationship of the ConceptNet.Tr, according to the category thereof. For example, the "*usedFor*" type of relationship was defined as expressing a pre-condition of use, because it denotes a function for determining a condition in which the origin concept can be used in sentences.
   - For ConceptNet.Tr relationships of the type *IsA* (c_1,c_2) and *DefinedAs* (c_1,c_2) inferential content of the parent concept c_2 was transcribed to the child concept c_1. These types of relationships define a hereditary relationship between c_1 and c_2. So, given that the pre-conditions for use of c_2 define which c_2 conditions can be used in sentences and that *IsA(c1,c2)*, we can assume that c_1 can also be used under the same preconditions as c_2. The same reasoning is valid for post-conditions of use of c_2. Therefore, for every relationship of the type *IsA*(c_1,c_2) and *DefinedAs*(c_1,c_2) we generate all of the pre-conditions and post-conditions of c_2 as pre-conditions and

[4] Research project funded by FUNCAP (*Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico Ceará* [Foundation for Support to Scientific and Technological Development], Proc. 9145/08 – Edital Nº 06/2008 – Information Technology.

post-conditions of c_1. For example, given that there is a pre-condition *CapableOf(crime,have victim)* of c_2=*crime* and that *IsA(murder,crime)*, we generate a pre-condition *CapableOf(crime, have victim)* for the concept c_1=*murder* (in Portuguese: '*assassinato*').

Table 1. Type of inferential relationship (pre-condition or post-condition) expressed by each category of relationship of ConceptNet.Tr

Category	Type of relation (ConceptNet.Tr)	Type of inferential relation (InferenceNet.Br)
THINGS	**PropertyOf; partOf; MadeOf; IsA; DefinedAs**	Pre-condition
SPATIAL	**LocationOf**	Pre-condition
EVENTS	**PrerequisiteEventOf; FirstSubeventOf; SubeventOf; LastSubeventOf**	Pre-condition
CAUSAL	**EffectOf; DesirousEffectOf**	Post-condition
AFFECTIVE	**MotivationOf; DesireOf**	Post-condition
FUNCTIONAL	**UsedFor; CapableOfReceivingAction**	Pre-condition
AGENTS	**CapableOf**	Pre-condition

InferenceNet.Br's Conceptual Base currently contains 186,047 concepts related inferentially through 842,392 relationships of the form *type_rel*(c_1,c_2), with each relationship type (*type_rel*) defining a pre- or post-condition of use of c_1. Figure 1 presents part of the conceptual base for the concept **crime**.

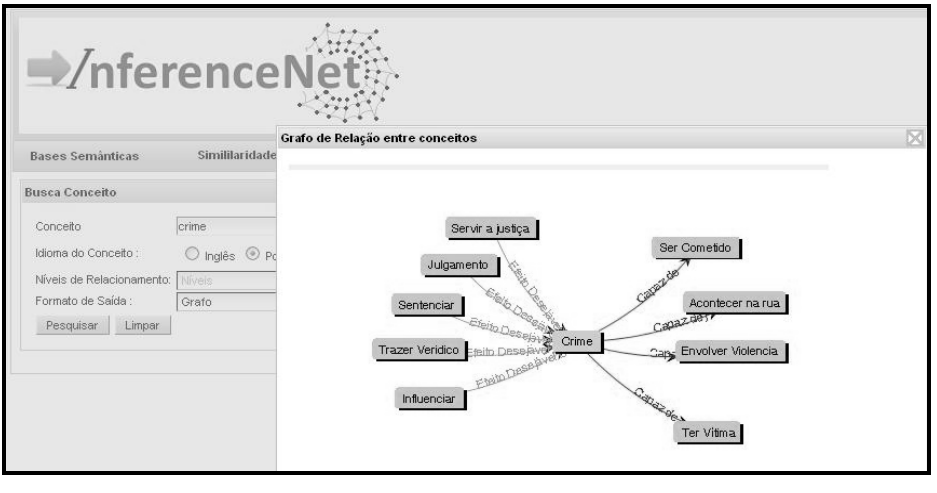

Fig. 1. Part of the conceptual base for the concept **crime** with some pre-conditions (incoming arrows) and post-conditions (outgoing arrows)

2.2 Construction of InferenceNet.Br's Sentence-Patterns Base

The process of constructing InferenceNet.Br's sentence-patterns base included the following steps:

1. **Extraction of Verb+Preposition occurrences based on the corpus.** The CRIMES2008 corpus – which was compiled with police news published on web pages of Brazilian newspapers and contains 158,664 words – was noted syntactically by the morphosyntactic parser PALAVRAS[13]. A parser was implemented on the treebank that, in every sentence, researched the occurrence of main verbs, auxiliary verbs, and all dependent prepositions of the verb(s) (dependencies were retrieved from the syntactic dependency tree generated by PALAVRAS). Each distinct occurrence (verbs+preposition) was generated as a sentence pattern and its frequency was calculated: the sentences with highest frequencies were "X ser Y" ["X be Y"] (1063), "X ter Y" ["X have Y"] (518) and "X fazer Y" ["X do Y"] (245). For example, the following sentence from the corpus: "*Segundo a Polícia, o bancário foi assassinado por dois homens que ocupavam uma moto*" ["*According to police, the bank employee was murdered by two men who were riding a motorcycle*"] generates the pattern "*X ser assassinar por Y*" ["X be murder by Y"], because the verbs "ser" ["be"] and "assassinar" ["murder"] were retrieved as the auxiliary verb and main verb (respectively) and the preposition "por" ["by"] was retrieved as a preposition directly dependent on the verb "assassinar" ["murder"]. We generated 5907 sentence patterns out of the 22,085 sentences in the corpus, i.e. 73% of the sentences in the corpus are structurally equal to at least one of the sentence patterns. The Sentence Patterns Base thereby generated follows Zipf's Law[5] and the Pareto Principle[6].
2. **Formalization of the base.** The sentence-patterns base consists of a directed graph $Gs(V,E)$, where V = non-empty set of sentence-patterns s_j and concepts c_i (vertices of the graph); E = set of edges labeled by a variable that indicates the pre- or post-condition relationship of a sentence-pattern s with concepts c in V, always in the direction from sentence s to concept c, in V. As this is a digraph, there are two functions s and t where $s:E\rightarrow V$ is a function that associates an edge of E with its origin sentence-pattern s in V, and $t:E\rightarrow V$ is a function that associates an edge of E with its target concept c in V.
3. **Generating the inferential content of sentence patterns.** Our need was to express the semantic-inferentialist content of sentence-patterns. There is no learning mechanism (automatic or semi-automatic) of knowledge of this nature and the first experiments and heuristics adopted were the following:

   - Generation of post-conditions according to the circumstance expressed by the adverbial complement of main verbs and auxiliary verbs: The parser PALAVRAS [13] does not identify the circumstance expressed in adverbial complements (for example, circumstance of cause, place or time). Therefore, the strategy used was to analyze the sentence-patterns with prepositions and the sentences from the corpus that follow the pattern, to infer what

[5] Zipf's law states that given a linguistic corpus with sentences in natural language, the most frequent word occurs approximately twice as much as the second most frequent, which, in turn, occurs approximately twice as much as the third, and so on. The Sentence-Patterns base follows the same frequency distribution.

[6] According to the Pareto principle, in many events, 80% of effects come from 20% of causes. *Mutatis mutandis*, it is observed that 80% of sentences from a linguistic corpus follow the structure of 20% of the sentences of the corpus.

circumstance it would express. Three groups of circumstances were defined (cause, place and time), in which all of the propositions were grouped. For each sentence-pattern whose preposition belongs to the group, post-conditions of the same nature were generated. For instance, for the sentence-pattern *s*= "*X ser assassinar em frente a Y*" ["X be murder in front of Y"] was generated the post-condition *IsA(complement(s),local*), meaning that "Y" is the local of crime.

- Generation of post-conditions of sentence patterns with verbs related to the act of committing crimes: We identified the verbs, from the Conceptual Base, that relate to the concept *crime* (we used the Inferential Relatedness Measure, described in [10,11] with a cutoff value of 50%). For each sentence pattern *s* containing the selected verbs, the post-conditions *IsA(subject(s),criminal)* or *IsA(subject(s),victim)* were generated, according to whether the verb voice is active or passive, respectively.

InferenceNet.Br's Sentence-Patterns Base currently contains 5907 sentence patterns and 1061 of relationships of the form *type_rel(<subject,complement>(s),*c_1*)*, defined as a pre- or post-condition of the sentence-patterns.

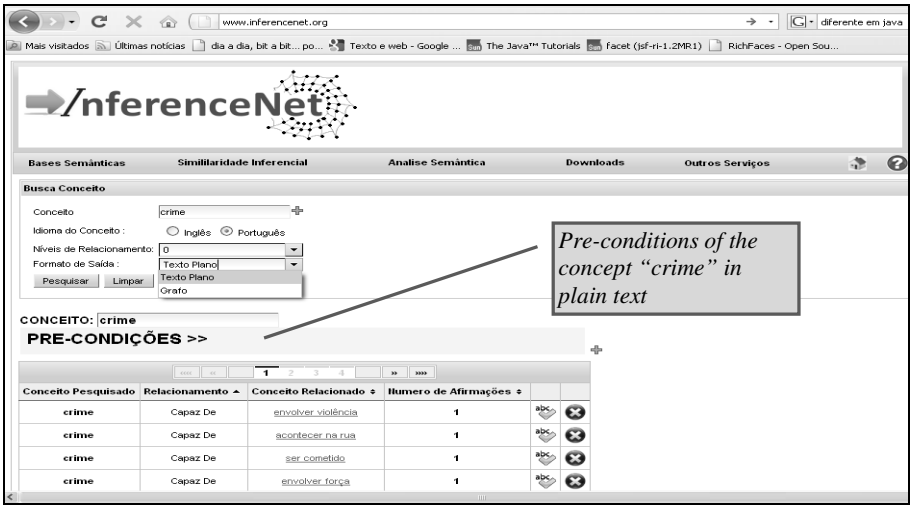

Fig. 2. The InferenceNet portal. Note the options of services available and a partial view of the query of the inferential content of the concept "*crime*".

2.3 The InferenceNet Portal

The InferenceNet portal can be accessed at www.inferencenet.org with services that enable the dissemination, use, and development of InferenceNet.Br in a collaborative manner, by the NLP community (Figure 2). The main services provided are the following:

- **Semantics Bases:** Enables the query and evolution of the content of InferenceNet's semantic bases. One can view, through plain text or graph, the pre-conditions/post-conditions for the use of concepts and sentence patterns. Figure 2 shows a partial view of the inferential content of the concept *crime* (pre-conditions). One can also include inferential content in the semantic bases, which are incorporated into InferenceNet.Br.
- **Inferential Similarity:** Performs the calculation of semantic similarity between two concepts, through the Inferential Relatedness Measure, described in [10,11].
- **Semantic Analyzer:** Allows the User to upload texts to be analyzed semantically, generating the premises and conclusions thereof, in plain text or graph.
- **Download:** This functionality allows user to download the latest version of InferenceNet.Br available and its documentation.

3 InferenceNet's Proof of Concept

The WikiCrimes system[7] [14] aims to provide a common area of interaction among people so that they can make the reports and monitor the locations where crimes are occurring. A need in the WikiCrimes project is to provide users with a tool that assists in the registration of crimes from news reports on the Web. For that, an information extraction system was developed – WikiCrimesIE – to extract information about crimes described in the Portuguese language and generate the records in the WikiCrimes database. The semantic bases of InferenceNet.Br were used as a linguistic resource of the WikiCrimesIE system.

The semantic reasoner of the WikiCrimesIE and its results with respect to precision and recall measures are described in [10]. The use of InferenceNet.Br in this task of extracting information allowed us to analyze how the use of resources with semantic-inferentialist content – as provided by InferenceNet.Br – is fundamental to NLP tasks. The aim is to prove, through a use case, the importance of resources of this nature for NLP tasks that involve an understanding of natural language. In particular, for extracting information about the *type of crime*, it was manually analyzed that in 95% of the texts, this information was not explicit, requiring complex inferences authorized by the semantic content expressed in terms of pre- and post-conditions for the use concepts and sentences (inferential content). For example, based on the sentence "*João assassinou sua esposa com vários tiros após uma discussão na Rua Solon Pinheiro*" ["*João murdered his wife with several gunshots after an argument on Solon Pinheiro Street*"] and on the inferential content of the concepts "assassinar" ["to murder"] and "esposa" ["wife"], the sufficient pre-condition of the type of crime "homicide" (PreRequisiteEvent(homicide, death of person)) was possible be inferred:

1.	IsA(*esposa*, *pessoa*)	[IsA(wife, person)]
2.	EffectOf(*assassinar*, *morte de pessoa*)	[Effect(murder,death of person)]
3.	IsA(complemento("X assassinar Y"),morte de Y)	[IsA(complement(("X assassinar Y"),death of Y)]
4.	IsA(*esposa do João*, *pessoa*), from (1)	[IsA(João's wife, person]
5.	(morte de esposa do João), from (3)	[death of João's wife]

[7] www.wikicrimes.org, accessed on 09/29/2009.

We argue that in order to extract information about type of crime, motive of the crime, and/or type of weapon, for instance, it is necessary to make more complex inferences, since such information is not explicit in the text, such as the address, thus requiring more ability to handle the semantic knowledge. The expressed semantic knowledge in InferenceNet.Br was fundamental for the levels of precision achieved.

4 Related Works

WordNet.Pt[2] and WordNet.Br[3] provide lexical-semantic resources for the Portuguese language; however, they express basic semantic hierarchy relationships (hyperonymy/hyponymy), inclusion (holonymy/meronymy), equivalence (synonymy) or opposition (antonymy), and are not available for large-scale use. The OMCS-Br project aims to capture commonsense knowledge in the Portuguese language and generate a base of relationships that are causal, functional, planning-related, etc., like the American ConceptNet[6]. However, it also has relationships only in specific themes (children's universe, colors, emotions, objects, etc.). One recent work discusses the challenges to constructing some frames specific to the Portuguese language [15]. The PAPEL base [16] contains semantic relationships that are causal, functional, synonymy-related, etc., extracted from the dictionary of the Portuguese language published Porto Editora.

The main differential of InferenceNet regarding these linguistic resources for the Portuguese language is its pragmatic bias. InferenceNet is innovative because it expresses situations of use of concepts and sentences in the Portuguese language, consisting of a base with semantic-inferentialist content theretofore non-existent for either English or Portuguese. By expressing pre- and post-conditions for use of concepts and sentence patterns, we provide NLP tasks (information extraction, text generators, summarizers, automatic translators, etc.) with an inferential base that allows improvement of the quality of reasoning at the semantic-pragmatic level. The network of [potential] inferences in which concepts can participate consists of a base for material and holistic reasoning in grasping the meaning of sentences. This is referred to as material reasoning because we have at hand the inferential content of the concepts, which allows inferences authorized by the content (e.g.: from "a bolt of lightning is seen now" to "thunder will soon be heard", authorized by the content of the concepts of "thunder" and "lightning"), and argument to refute and validate inferences (e.g.: "Water is red" is refuted by the pre-condition for use of the concept of "water", which defines that this concept can only be used in sentences in which no color is associated with it). Holistic reasoning, in turn, is based on the inferential content of the concepts and on how such concepts are related in the sentence, according to its syntactic structure (sentence-pattern).

5 Conclusion

This paper presents a new linguistic resource for the Portuguese language – InferenceNet.Br – and its collaborative portal www.inferencenet.org that will allow the use and subsequent development by the NLP community. Currently, this resource consists of a base of concepts in the Portuguese language related to one another through pre- or post-conditions of use, defining an inferential network, as well as a base of

sentence-patterns extracted from the linguistic practice of police journalism. The use of InferenceNet.Br in an information extraction system about crimes (WikiCrimesIE) enabled us to evaluate the usefulness and effectiveness of semantic-inferentialist content in NLP systems that involve an understanding of natural language. Plans for future work include an evaluation of the "PAPEL" bases and the OCMS-Br project as resources to be incorporated into InferenceNet.Br, and the use thereof in other NLP tasks such as text generators and summarizers.

References

1. Fellbaum, C. (ed.): WordNet: An electronic lexical database. MIT Press, Cambridge (1998)
2. Marrafa, P., et al.: WordNet.PT – Uma Rede Léxico-conceptual do Português online. In: XXI Encontro da Associação Portuguesa de Linguística, Porto, Portugal (2005)
3. Dias da Silva, B.C., Di Felippo, A., Hasegawa, R.: Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 120–130. Springer, Heidelberg (2006)
4. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM 38(11) (1995)
5. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: Proceedings of COLING-ACL (1998)
6. Liu, H., Singh, P.: ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal 22(4) (2004)
7. Sellars, W.: Inference and meaning (1950); Reprinted in: Sicha, J. (ed.): Pure Pragmatics and Possible Worlds. Ridgeview Publishing Co, Reseda (1980)
8. Dummett, M.: Frege's Philosophy of Language. Harvard University Press, Cambridge (1973)
9. Brandom, R.B.: Articulating Reasons. In: An Introduction to Inferentialism. Harvard University Press, Cambridge (2000)
10. Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D.: Semantic Inferentialist Analyser: Um Analisador Semântico de Sentenças em Linguagem Natural. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology. STIL, Brasil (2009)
11. Pinheiro, V., Pequeno, T., Furtado, V., Nogueira, D.: Information Extraction from Text Based on Semantic Inferentialism. In: Andreasen, T., et al. (eds.) FQAS 2009. LNCS (LNAI), vol. 5822, pp. 333–344. Springer, Heidelberg (2009)
12. Pinheiro, V., Pequeno, T., Furtado, V., Assunção, T., Freitas, E.: SIM: Um Modelo Semântico-Inferencialista para Sistemas de Linguagem Natural. In: VI Workshop em Tecnologia da Informação e da Linguagem Humana -TIL, WebMedia, Brasil (2008)
13. Bick, E.: The Parsing System "Palavras". In: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)
14. Furtado, V., et al.: Collective intelligence in law enforcement – The WikiCrimes system. Information Sciences (in Press), Corrected Proof, Available online (August 2009) doi:10.1016/j.ins.2009.08.004.
15. Bertoldi, A., Chishman, R.: Desafios para a Criação de um Léxico baseado em Frames para o Português: um estudo dos frames Judgment e Assessing. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), São Paulo, Brasil (2009)
16. Oliveira, H.G., Santos, D., Gomes, P.: Avaliacão da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário. PAPEL. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), São Paulo, Brasil (2009)

Comparing Verb Synonym Resources for Portuguese

Jorge Teixeira, Luís Sarmento, and Eugénio Oliveira

Universidade do Porto - LIACC
Rua Dr. Roberto Frias, s/n - Porto, Portugal
{jft,las,eco}@fe.up.pt

Abstract. In this paper we compare verb synonym information contained in four public-available lexical-semantic resources for Portuguese: TeP, PAPEL, Wiktionary and OpenThesaurusPT. We quantify the extent to which verb synonymy information in four resources overlaps, and we quantify how much novelty each resource in comparison to the others. We demonstrate that the four resources vary *significantly* in respect to verb synonymy information. Also, we show that by merging the four resources we can obtain a more comprehensive verb thesaurus. Finally, we suggest that resource merging may actually be required in order to avoid *performance* and *evaluation bias* that arise from coverage problems when using only one of these resources.

Keywords: Language Resources, Synonymy, Evaluation.

1 Introduction

In the last few years there has been a significant increase in the number of lexical-semantic resources for Portuguese. This results, on one hand, from the products of several finished or on-going projects, such as TeP [1], WordNet.PT [2] and, more recently, MultiWordNet.PT (the Portuguese branch of the MultiWordNet project [3]) and PAPEL [4]. Essentially, these projects aim at producing *high-quality* WordNet-like resources for Portuguese, supported either by *human specialists* or by previously compiled *reference dictionaries*. On the other hand, community-edited resources, such as the Wikipedia, Wiktionary or OpenThesaurus, have become an extremely attractive option for obtaining *broad-coverage* lexical-semantic knowledge, at almost no cost.

As the number of choices of lexical-semantic resources increases, several questions immediately arise. The first is: *how do these resources relate to each other*? One typical claim is that manually edited resources - such as TeP or, indirectly, PAPEL - will have high-quality information at the cost of a lower coverage (or over-specialization), while community-edited resources tend to have much broader coverage (e.g. at the level of the semantic domains included) but provide no quality guarantees. The second question is: *how much benefit is there from merging the information of all resources*? Merging is an interesting option

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 100–109, 2010.

not only because of *quality vs. coverage* issues but also because, quite paradoxically, resources tend to *diverge* in many points. This is so because when building lexical-semantic resources, developers need to follow a given linguistic / conceptual framework, which inevitable imprint a specific *semantic bias* [5] to the corresponding resource. Since there are several possible linguistic / conceptual frameworks, the resulting resources will tend to be significantly different in points where the frameworks do not agree.

In the current work, we compare the *verb synonymy* information offered by four freely-available lexical-semantic resources for Portuguese: TeP, PAPEL, Wiktionary and OpenThesaurusPT. These four resources result from four distinct approaches to resource building and are thus expected to posses only partially overlapping information. We will focus on comparing the information regarding one specific semantic relation - synonym - for one word class only: verbs. We will compute the degree to which the verb synonymy information in four resources overlaps, and quantify how much novelty each resource has in relation to the others. Finally, we will present statistics about the result of merging the verb synonym information of the four resources into a more comprehensive verb thesaurus.

2 Related Work

Despite the significant increase in the number of available lexical-semantic resources, including community-edited resources such as Wikipedia or Wiktionary, in the last few years there has not been a corresponding effort by the research community in *explicitly* comparing those resources. There are, however, a few works in which a set of lexical-semantic resources is compared with the purpose of (i) achieving cross-validation, (ii) measuring the benefit of merging their information, or (iii) testing their usefulness in a specific application settings.

For example, Oliveira et al [6] evaluate PAPEL, a lexical resource built by extracting relations from a large commercial Portuguese dictionary [4], and perform a comparative study against TeP thesaurus [1] by computing the overlap between synsets of both the resources. Results obtained show that 50% of the synonym pairs in PAPEL were found in TeP, while only 39% of the pairs in TeP were found in PAPEL.

Navarro et al [7] evaluate how two lexical-semantic resources, WordNet and Dicosyn (a compilation of synonym relations extracted from seven dictionaries) can be used to augment synonymy information in Wiktionary. They compute two overlap measures: *recall* - the fraction of the common entries between the resources, and *precision* - the fraction of common synonyms for the common entries. The comparison between the English Wiktionary and WordNet yields recall values from 0.45 (for verbs) to 0.12 (for nouns), while precision ranges from 0.23 (for nous) to 0.08 (for verbs). As for the comparison between the French Wiktionary and Dicosyn, the recall values ranges from 0.35 for nouns to 0.32 for verbs, while precision values varies from 0.08 (for nouns) to 0.04 (for verbs).

In [8] the authors compare the performance of several word selection procedures that use WordNet-like resources for computing two semantic relatedness

measures - *path length* and *concept vector*. Four different resources were used to ground the computation of such measures: Wiktionary, Wikipedia, GermaNet and WordNet. Results achieved using Wiktionary were found to be *comparable*, and sometimes *better*, than those achieved when using expert-made resources (i.e. GermaNet and WordNet).

Muller et al [9] compare the effectiveness of using Wiktionary, Wikipedia and WordNet in reducing the vocabulary gap between query terms and documents in a Information Retrieval setting. They showed that all the three resources contain sufficient information for achieving such goal by helping to expand queries with related words (obtained by computing a word relatedness metric presented in [8]). However, results differed depending on the semantic relation between the words in query and best words for expansion. For example, while WordNet was especially effective when hypernyms/hyponyms were the optimal expansion keywords, Wiktionary was the most effective in providing synonyms and co-hyponyms, and Wikipedia was the best in providing words related by non-classical relations (e.g. functional).

3 Resources Being Compared

In this section we describe the four resources being compared: Wiktionary, PAPEL, TeP and OpenThesaurusPT. Among other information, all the four resources store synonym information about verbs (which are the focus of our current study), and are freely available for download. There are other lexical-semantic resources for Portuguese, such as WordNet.PT and MultiWordNet.pt, which provide Web interfaces available for users to query[1]. However, as far as we know, none of these two resource is available on the Web *for download*.

3.1 Wiktionary

Wiktionary is a comunited-edited resource that stores a wide variety of information about words, such as grammar, pronunciation, etymology, definitions, a list of synonyms, a list of related terms, a list of derived terms, and also translations to other languages. Wiktionary project started in 2002, and it currently has more than 6.9 million entries for 172 languages. The Portuguese version of Wiktionary was started in 2004 and contains approximately 92,000 entries. The structure of a Wiktionary page for verbs is generally divided in two sections from which synonymy information can be obtained. The first is the *synonyms list* that explicitly presents synonyms, but is only available for a relatively small number of words. The second section is the *word definitions* that may be found in a larger number of pages, and provide extensive information regarding synonyms. The version of the Wiktionary for Portuguese we used[2] contains 92,144 pages,

[1] WordNet.PT: http://cvc.instituto-camoes.pt/wordnet/
MultiWordNet.pt: http://mwnpt.di.fc.ul.pt/

[2] http://download.wikimedia.org/ptwiktionary/20090809/ptwiktionary-20090809-pages-articles.xml.bz2

each containing a word/expression. We found 4,818 pages addressing verbs in Portuguese, of which 1,656 have synonym information.

3.2 PAPEL

PAPEL (*Palavras Associadas Porto Editora Linguateca*) [4] is a freely available[3] lexical database for Portuguese, automatically constructed by extracting relations from a large commercial Portuguese dictionary. The process consisted in parsing 237,246 dictionary definitions (using a chart parser), and using a set of manually-developed lexical-syntactical patterns to mine 10 semantic relations: (i) *hypernym*, (ii) *cause_of*, (iii) *part_of*, (iv)*means_to_end*, (v) *place_of*, and its corresponding inverse relations. PAPEL contains information about 125,871 words: 50,201 are nouns, 17,932 verbs, 14,025 adjectives and 43,713 adverbs (however, this value seems to be too large and is probably a spelling mistake in [4]). It contains 80,429 pairs of *synonym* relations, of which we will focus on the ones that connect pairs of verbs.

3.3 TeP

TeP (*Thesaurus Electrônico para o Português do Brasil*) [1] is a freely-available[4] electronic thesaurus for the Brazilian Portuguese, strongly inspired by WordNet. TeP development was based on four dictionaries used as reference corpus and relies on three main concepts: (i) *synsets*, i.e. sets of synonyms, (ii) the *lexical matrix*, which defines a biunivocal correspondence between synsets and *senses*, (iii) and the *index*, which establishes the antonymy relation by using indexes. TeP contains information about approximately 44,000 words: 17.000 nouns, 15,000 adjectives, 11,000 verbs and 1,000 adverbs. There are 19,885 synsets, 4,145 of which related to verbs. As far as we know, there is only one study [10] (unpublished at the time of this writting) that compares *TeP* thesaurus with *WordNet.pt* and *PAPEL*.

3.4 OpenThesaurusPT

OpenThesaurusPT is the official thesaurus used for OpenOffice and is freely available online[5]. The OpenThesaurusPT project [11] started as an effort to migrate the German OpenThesaurus project to the Portuguese. The OpenThesaurus project is based on Wordnet, and also uses synsets to hold information about synonyms and antonyms. It also stores some information about superordinate/subordinate (i.e. hypernymy/hyponymy) relations between synonym sets. The current available version of OpenThesaurusPT dates from 2006-08-17, and contains 4,010 synsets for 12,941 words, 2,815 of them verbs.

[3] http://www.linguateca.pt/PAPEL/PAPELv1.0.zip
[4] http://www.nilc.icmc.usp.br/tep2/download.htm
[5] http://openthesaurus.caixamagica.pt/index.php

4 Experimental Set-Up

4.1 Pre-processing and Normalization

The resources being compared provide information about synonymy in different formats: TeP and OpenThesaurusPT provide *synsets*, PAPEL provides a list with *pairs* of synonyms and Wiktionary provides information about the synonyms of a word embedded in the corresponding page, formatted according to the standard Wiki formatting style. For Wiktionary, we developed our own parsing procedures, highly optimized for extracting synonyms from lists and from available definitions. However, because Wiktionary is a community-edited resource, several pages have formatting inconsistencies which lead to extraction problems. We thus, manually evaluated a sample of 20% of the synonym data extracted. We concluded that our extraction procedures achieve over 98% of *precision* and over 90% *recall*, which is compatible with our evaluation purposes.

Additionally, each resource deals with the symmetry of synonymy in different ways. For example, in Wiktionary if we find from the page of verb v_A that v_B is one of its synonyms, there is no guarantee that we find in the page of v_B information indicating that v_A is one of its synonyms. We have thus inverted all synonym relations extracted from Wiktionary to obtain explicit representations for *both* directions of the relation. In TeP and OpenThesaurusPT, information is given in synsets, so we established bidirectional synonymy links between all possible pair in the set. In PAPEL, there is one explicit synonymy link for each direction of the synonym relation so no further processing was needed. Finally, for each of the four resources, generically denoted by $\mathcal{R}$, the (symmetric) synonymy pairs for each verb were aggregated in order to create mappings between verbs $v^{\mathcal{R}}(i)$ the corresponding list of synonyms, $S^{\mathcal{R}}(i)$:

$$v^{\mathcal{R}}(i) \longrightarrow S^{\mathcal{R}}(i) = [s_1^{\mathcal{R}}, s_2^{\mathcal{R}} ... s_j^{\mathcal{R}}] \tag{1}$$

We will denote, $v^{\mathcal{R}}(1)$, $v^{\mathcal{R}}(2)$, ... $v^{\mathcal{R}}(n)$, the set of all verbs for which synonym mappings were obtained as $\mathcal{V}^{\mathcal{R}}$. Thus, for Wiktionary, PAPEL, OpenThesaurus PT and TeP we will have, $\mathcal{V}^{wk}$, $\mathcal{V}^{pp}$, $\mathcal{V}^{ot}$ and $\mathcal{V}^{tp}$, respectively. It is important to clarify that information about the different senses of verbs was available only in TeP and in the OpenThesaurusPT thesaurus (i.e. the synsets provided address different senses). However, while producing the synonym mappings we ignored information about multiple senses. Thus, for all verb considered in this work, the set of synonyms includes synonyms for all senses found, so as to compare all resources in equal conditions. Table 1 presents some statistics about the number of verbs, $|\mathcal{V}^{\mathcal{R}}|$, and the average number of synonyms gathered for each verb, $\overline{|S^{\mathcal{R}}(i)|}$, for the four resources under study. Also, we present an indicative measure of amount of synonymy information found in each resource, by computing the number of synonymy links it stores i.e. $|\mathcal{V}^{\mathcal{R}}| \times \overline{|S^{\mathcal{R}}(i)|}$.

Table 1. Verb Synonyms in each resource after normalization

Resource	$\lvert\mathcal{V}^{\mathcal{R}}\rvert$	$\overline{\lvert S^{\mathcal{R}}(i)\rvert}$	$\lvert\mathcal{V}^{\mathcal{R}}\rvert \times \overline{\lvert S^{\mathcal{R}}(i)\rvert}$
OpenThesaurusPT	2,815	1.67	4,701
PAPEL	4,142	5.38	22,284
Wiktionary	1,964	2.34	4,596
TeP	10,827	3.28	35,512

4.2 Comparison Measures

We are mainly interested in comparing two parameters over all resources. The first is the *overlap* between them, i.e. the amount of information that resources have in common. This implies measuring two more specific parameters: (i) the number of common verb entries extracted, and (ii) for each common verb entry, the number of common synonyms found. The second parameter is the relative *novelty* between resources, i.e. the amount of information that one resource has while the others do not. Again this can be seen both at the level of verb entries or, for each verb entry in common, at the level of synonyms.

Let us assume that $\mathcal{A}$ and $\mathcal{B}$ are two generic verb synonymy mappings to be compared. Let us define *Entry Overlap*, $\mathcal{O}_{ent}$, as the number of common verb entries between resources $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{O}_{ent}^{\mathcal{A}:\mathcal{B}} = |\mathcal{V}^{\mathcal{A}} \cap \mathcal{V}^{\mathcal{B}}| = |\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}| \tag{2}$$

We can also define *Entry Novelty from* $\mathcal{A}$ *to* $\mathcal{B}$ to as the number of entries in $\mathcal{A}$ for which there is *no* corresponding entry in $\mathcal{B}$:

$$\mathcal{N}_{ent}^{\mathcal{A}\rightarrow\mathcal{B}} = |\mathcal{V}^{\mathcal{A}} - \mathcal{V}^{\mathcal{A}\cap\mathcal{B}}| \tag{3}$$

Let us now consider only the set of verbs that have synonyms in both $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}$. For all verbs $v(i)$ in $\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}$ we can measure the *synonym overlap* $O_{syn}^{\mathcal{A}:\mathcal{B}}(i)$ between the corresponding synonyms sets, i.e. the number of common synonyms between the set of synonyms found in $\mathcal{A}$, $S^{\mathcal{A}}(i)$, and in the set found in $\mathcal{B}$, $S^{\mathcal{B}}(i)$:

$$O_{syn}^{\mathcal{A}:\mathcal{B}}(i) = |S^{\mathcal{A}}(i) \cap S^{\mathcal{B}}(i)| \tag{4}$$

A resource-wide synonym overlap measure, $\mathcal{O}_{syn}^{\mathcal{A}:\mathcal{B}}$ can be computed by averaging the values of overlap for all the verbs at stake, $\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}$:

$$\mathcal{O}_{syn}^{\mathcal{A}:\mathcal{B}} = \frac{\sum_{i}^{|\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}|} O_{syn}^{\mathcal{A}:\mathcal{B}}(i)}{|\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}|} \tag{5}$$

For any $v(i) \in \mathcal{V}^{\mathcal{A}\cap\mathcal{B}}$ we can also define *synonym novelty from* $\mathcal{A}$ *to* $\mathcal{B}$, $N_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}(i)$, as the number synonyms found in $\mathcal{A}$, $S^{\mathcal{A}}(i)$, that are not found in $\mathcal{B}$, $S^{\mathcal{B}}(i)$:

$$N_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}(i) = |S^{\mathcal{A}}(i) - S^{\mathcal{A}}(i) \cap S^{\mathcal{B}}(i)| \tag{6}$$

Likewise, the resource-wide *average synonym novelty from* $\mathcal{A}$ *to* $\mathcal{B}$, $\mathcal{N}_{syn}^{\mathcal{A}\to\mathcal{B}}$, can be computed by averaging the values of $N_{syn}^{\mathcal{A}\to\mathcal{B}}(i)$ for all verbs in $\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}$:

$$\mathcal{N}_{syn}^{\mathcal{A}\to\mathcal{B}} = \frac{\sum_{i}^{|\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}|} N_{syn}^{\mathcal{A}\to\mathcal{B}}(i)}{|\mathcal{V}^{\mathcal{A}\cap\mathcal{B}}|} \tag{7}$$

5 Results and Analysis

Table 2 presents the values of *entry overlap*, $\mathcal{O}_{ent}^{\mathcal{A}:\mathcal{B}}$, between all resources. Wiktionary has the lowest entry overlap among all the evaluated resource, even-though its overlap with TeP is the highest. TeP, which is the largest resource being evaluate, is, as expected, the one with the highest entry overlap.

Table 2. Entry Overlap $\mathcal{O}_{ent}^{\mathcal{A}:\mathcal{B}}$ between resources

$\mathcal{A}:\mathcal{B}$	PAPEL	Wiktionary	TeP
OpenThesaurusPT	2,119	1,407	2,664
PAPEL	-	1,816	3,647
Wiktionary	-	-	1,883

Entry Novelty results are presented in Table 3 and show that Wiktionary is the resource with less novel information, while PAPEL and TeP are the ones the highest number of novel entries.

Table 3. Entry Novelty, $\mathcal{N}_{ent}^{\mathcal{A}\to\mathcal{B}}$, between the resources

$\mathcal{A}\setminus\mathcal{B}$	OpenTh.PT	PAPEL	Wiktionary	TeP
OpenTh.PT	-	696	1,408	151
PAPEL	2,023	-	2,326	495
Wiktionary	557	148	-	81
TeP	8,163	7,180	8,944	-

Figure 1 presents perspective that combines both the relative sizes of the four resources (already presented in Table 1) and the entry overlap between all combinations of resources (including the aggregation of all resources). As it can be seen, the entry overlap between *all* resources is relatively low: there are only 1339 common entries between all resources in an universe of 11,450 verbs. It is also interesting to confirm that the two largest resources exhibit a relatively large number of exclusive entries, while the two smallest (and community-edited resources) have much fewer novel entries (i.e. Wiktionary with only 29 novel verbs and OpenThesaurusPT with 98 novel verbs).

Tables 4 and 5 present statistics about the average synonym overlap, $\mathcal{O}_{syn}^{\mathcal{A}:\mathcal{B}}$ and the average synonym novelty $\mathcal{N}_{syn}^{\mathcal{A}\to\mathcal{B}}$. There are two interesting facts that can be observed in these tables. The first is that the synonym overlap between

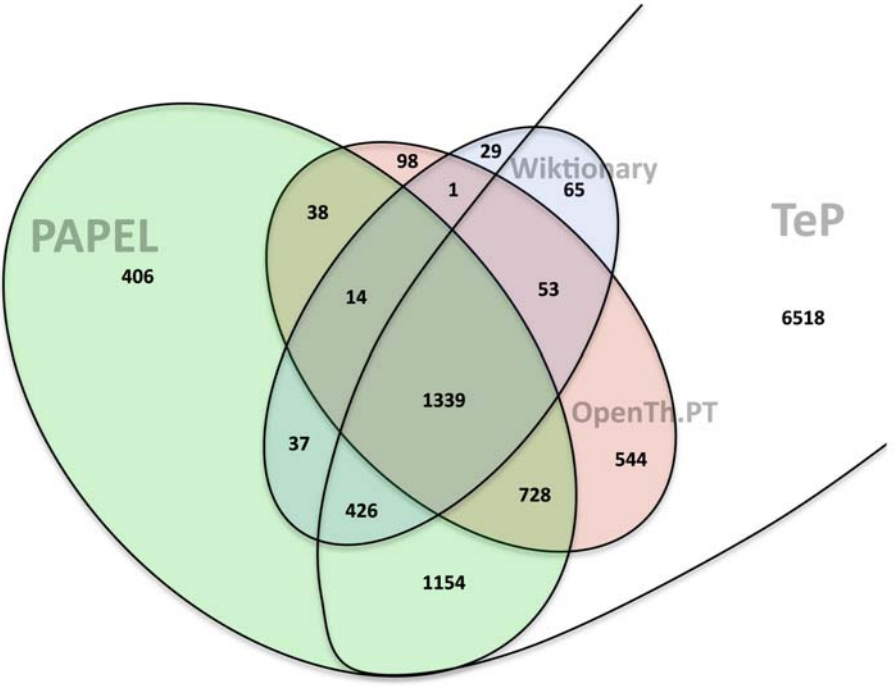

Fig. 1. Perspective of all the resources and their overlap

resources is *very low*. In most cases, the average synonym overlap is less than 1 meaning that there are many verbs for which the resources do not share any synonym. There is a relatively higher synonym overlap between PAPEL on one side and OpenThesaurusPT and Wikitionary on the other, but not between these two. This suggests that PAPEL aggregates part of the synonymy information that can be either on OpenThesaurusPT or on Wikitionary. Notably TeP has a very low synonymy overall with all other resources despite the fact that it is the resource with larger number of verb entries.

The second interesting fact observed is that most values of average synonymy novelty, $\mathcal{N}_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}$, are relatively high. For example, for shared verbs PAPEL stores 6.97 synonyms not found in the OpenThesaurusPT, 6.54 not found in Wiktionary and 5.00 not found in TeP. Notably, the inverse is also true: for the shared verbs TeP has 3.90 synonyms not found in PAPEL, and OpenThesaurusPT has 1.05 synonyns not found in PAPEL. Of all resources, OpenThesaurusPT seems to be

Table 4. Synonym Overlap $\mathcal{O}_{syn}^{\mathcal{A}:\mathcal{B}}$ between resources

$\mathcal{A}:\mathcal{B}$	PAPEL	Wiktionary	TeP
OpenThesaurusPT	0.68	0.42	0.44
PAPEL	-	1.62	0.87
Wiktionary	-	-	0.50

Table 5. Synonymy Novelty $\mathcal{N}_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}$ between the resources

$\mathcal{A} \setminus \mathcal{B}$	OpenTh.PT	PAPEL	Wiktionary	TeP
OpenTh.PT	-	1.05	1.43	1.22
PAPEL	6.97	-	6.54	5.00
Wiktionary	2.12	0.80	-	1.88
TeP	4.92	3.90	5.80	-

the one with less average synonymy novelty in relation to the others but, still, the corresponding $\mathcal{N}_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}$ values are always higher than one.

The low values of $\mathcal{O}_{syn}^{\mathcal{A}:\mathcal{B}}$ combined with the relatively high values of $\mathcal{N}_{syn}^{\mathcal{A}\rightarrow\mathcal{B}}$ suggest that the synonym information differs *considerably* among the four resources. Additionally, since entry novelty among resources is also quite high (check Figure 1) one can say that the four resources are *significantly different*. By merging the information of the four resources we found that it us possible to obtain synonym mappings for 11,450 verb with 5.49 synonyms in average, which results in 62,902 synonymy links. When comparing these values with the corresponding values for each of the four resources (see Table 1)one can see that there is a very significant increase in all the parameters, especially in the number of synonym links. This confirms that merging existing resources is, in fact, a good option for reducing the coverage and recall gaps of each individual resource.

6 Conclusion and Future Work

We compared the overlap and novelty of verb synonymy information contained in four freely-available resources. Results show that there are *significant differences* among *all* the resources both at the level of *verb coverage* and at the level of *synonym links* between shared verbs. Globally, as far as verb synonymy is concerned, TeP and PAPEL are clearly much more comprehensive resources than Wiktionary or OpenThesaurusPT. However, the verb synonymy information contained in these two community-edited resources is non-negligible. In the particular case of Wiktionary, we believe that it is a very interesting resource for developers to follow since it is constantly growing (contrary to OpenThesaurusPT which has not been updated recently) and has the potential to efficiently cover more the frequently used lexicon.

Overall, these results should draw our attention to the fact that applications using any of these resources may obtain rather different performances depending on the actual resource chosen. Likewise, when using any of these resources as *gold-standard* data to evaluate automatic procedures, evaluation figures may vary significantly depending on the resource used as reference. Additionally, since each resource covers different parts of the lexicon, it is unsafe to compare automatic procedures using only one of these resources. It is possible that both procedures being evaluated produce different yet *equally valid results*, although only one set of results intersects with the specific gold-standard chosen. By merging information all the four resources we reduce the chances of these problematic events.

In future work, we wish to extend the comparative work performed on synonymy verbs both to other parts-of-speech, such as nouns and adjectives, and to other semantic relations such as antonymy, hypernymy/hyponymy and meronymy. However, not all resources available (e.g. the OpenThesaurusPT thesaurus) provide information about all these semantic relations. In the case of Wiktionary, new extraction procedures have to be developed to mine information regarding additional semantic relations.

Acknowledgments

This work was partially supported by grant SFRH/BD/23590/2005 FCT-Portugal, co-financed by POSI and by grant SAPO/BI/UP/2009 from Portugal Telecom.

References

1. da Silva, B.D.: Construcão de um thesaurus electrónico para o português do brasil. Processamento Computacional do Português Escrito e Falado (PROPOR) 4, 1–10 (2000)
2. Marrafa, P., Amaro, A., Chaves, P., Lourosa, S., Martins, C., Mendes, S.: Wordnet.pt - uma rede léxico-conceptual do português online. In: XXI Encontro da Associação Portuguesa de Linguística, 28-30 de Setembro, Porto, Portugal (2005)
3. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet, Mysore, India, January 21-25 (2002)
4. Oliveira, H., Santos, D., Gomes, P., Seco, N.: Papel: A dictionary-based lexical ontology for portuguese. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 31–40. Springer, Heidelberg (2008)
5. Velardi, P., Pazienza, M.T., Fasolo, M.: How to Encode Semantic Knowledge: A Method for Meaning Representation and Computer-Aided Acquisition. Computational Linguistics 17, 153–170 (1991)
6. Oliveira, H., Santos, D., Gomes, P.: Avaliação da extracção de relações semânticas entre palavras portuguesas a partir de um dicionário. In: 7th Brazilian Symposium in Information and Human Language Technology, STIL 2009 (2009)
7. Navarro, E., Sajous, F., Gaume, B.: Wiktionary and nlp: Improving synonymy networks. In: Proceedings for the Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (2009)
8. Zesch, T., Müller, C., Gurevych, I.: Using wiktionary for computing semantic relatedness. In: Proceedings of AAAI, pp. 861–867 (2008)
9. Muller, C., Gurevych, I.: A study on the semantic relatedness of query and document terms in information retrieval. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (2009)
10. Santos, D., Barreiro, A., Costa, L.: O papel das relações semânticas em português: Comparando o tep, o mwn.pt e o papel. To appears in XXV Encontro Nacional da Associação Portuguesa de Linguistíca, Lisboa, Portugal (2009)
11. Naber, D.: Openthesaurus: Building a thesaurus with a web community. Technical report (2004)

Auxiliary Verbs and Verbal Chains in European Portuguese

Jorge Baptista[1,3], Nuno Mamede[2,3], and Fernando Gomes[2,3]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
[2] Universidade Técnica de Lisboa, Instituto Superior Técnico
[3] L2F Spoken Language Laboratory - INESC ID Lisboa
jbaptis@ualg.pt, Nuno.Mamede@l2f.inesc-id.pt, fmfg@ist.utl.pt

Abstract. This paper describes auxiliary verb constructions in European Portuguese in view of their correct parsing in a fully integrated NLP chain. The paper provides data on these constructions over a large-sized corpus and evaluates the parsing system performance.

Keywords: Auxiliary verb, verbal chain, European Portuguese, parsing, NLP.

1 Introduction

Besides its rich verbal morphology, Portuguese shows a copious system of verbal auxiliaries (*Vaux*), used to express different, mainly aspectual and modal, but also some few temporal, grammatical values. Traditional grammar has for long included in the description of verbal inflexion paradigms the so-called compound tense formed by the verb *ter* 'to have' (or, more rarely, *haver* 'there be') and the past participle. Other *Vaux* combinations (e.g. *ir* 'to go'+ infinitive, *estar* 'to be'+ gerund, *poder* 'may/can'+ infinitive) were described as «verbal periphrasis» without any further commentary on the relation between them and that canonical 'compound tense' (see for example [8]). This modest grammatical tradition contrasts with the frequency of auxiliary verbs in texts, when compared with the use of simple verb forms. In fact, we estimate that no less than 30% of all verbal predicates in large-sized corpora of texts present an auxiliary verb construction (see below). In view these figures, the correct identification of auxiliary verb constructions is essential not only for the sake of linguistic adequacy but also to develop many NLP applications. For example, if two verbs form an auxiliary construction, their correct parsing will make them count as a single verb domain (i.e. auxiliary and main verb); the identification of the syntactic structure and the word sense will then hinge on the main verb, with only one set of dependent arguments. In the case of reduced (zeroed or pronoun) arguments, an anaphora resolution system/module [15] may then be activated, in order to look for the antecedent of those reduced arguments; the strategy (and results) may differ if the two verbs do not form a single verbal domain but two, with the corresponding argument domains. A syntactic-semantic interface can produce an operator-argument domain representation [7], corresponding to the syntactic expression of a single semantic predicate (that of the main verb), only if the correct parsing is available. In

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 110–119, 2010.

other words, a single verb domain should be determined for every auxiliary verb construction. None of the above would be possible if no information is available for auxiliary constructions.

The aim of this paper is to describe auxiliary verb constructions in European Portuguese in view of their correct parsing in the fully integrated NLP chain of L2F [14,15]. The paper provides data on these constructions over a large sized corpus and presents a preliminary evaluation of the parsing system performance.

The paper is structured as follows: First, section 2 presents related work on parsing systems dealing with auxiliary verbs. Then, section 3 proceeds with the basic definitions and the classification of auxiliary verb constructions. Next, section 4 describes the implementation of these linguistic aspects in the parsing system is described. Finally, results from a preliminary evaluation are presented and discussed in sections 5 and 6. The paper concludes with the mapping of future developments of current work (section 7).

2 Related Work

In romance languages, auxiliary verbs are an unavoidable problem for many NLP tasks. While the precise listing and syntactic properties of auxiliary verbs in Portuguese still constitute a matter of linguistic debate [5,6,8,10,17,18], it is possible to contemplate a reasonable consensus around a significant number of auxiliary verbal constructions [3,5,8,20]. Some work emphasizes the formal difficulty associated to the identification of auxiliary verbs in texts, namely insertions and clitic attachment. For example [19] proposes a finite-state approach to their identification, but without integration in any fully developed parser. In fact, there are, to our knowledge, very few, fully integrated NLP systems that parse auxiliary verb constructions in Portuguese. One of such cases is [4]. [2] also addresses the auxiliary verbal chains' problem in a transfer machine translation context using finite-state methods. Other developments for Portuguese relate to semantic representation of complex tense and aspect values involved in auxiliary construction have been attempted [9] in view of sentence generation. A brief reference to *Vaux* is made in [11] on the context of multilanguage word alignment. To our knowledge, no systematic evaluation for the problem of parsing Portuguese auxiliary verbs has been presented so far.

3 Basic Definitions

The concept of auxiliary encompasses a wide range of linguistic phenomena [3,20], including copula verbs (VCOP) and support verbs (VSUP). Certain nouns [18] and adjectives [5] can also be considered to function as auxiliary constructions. This is the case of aspectual nouns (*Nasp*) such as *fase* 'phase' in *A empresa está em fase de recuperação* 'The company is in phase of recovering'; or certain adjectives like *prestes* 'ready, to be on the verge of' in *A empresa está prestes a recuperar da crise* 'The company is on the verge of recovering from the crisis'.

Auxiliary verbs are to be considered as grammatical elements (as opposed to lexical, or meaningful, or distributional verbs) [12,13], not only because of their loosely

distributional constraints and lack of characteristic distribution, but also because they convey grammatical meanings, falling (mainly) under the scope of grammatical categories expressing the concepts of mode (VMOD), aspect (VASP) and tense (VTEMP) [5,17].

In this paper, only auxiliaries of verbs will be considered. Even so, certain syntactical structures, such as *O Pedro começou o livro* = *O Pedro começou a ler o livro* 'Peter begun the book = begun reading the book', which can be analyzed as resulting from appropriate reductions of distributional verbs (in this example *ler* 'read') in front of certain auxiliary verbs (in this case, *começar a* 'begin to') [18], were also not considered in this paper. Passive auxiliary verbs were assimilated to copula verbs. Therefore, verbal chains on passives do not include the past participle.

3.1 Auxiliary Verb Construction

Following [12, 13], we consider that auxiliary verbs are transparent to the selection restrictions of the main verb (*Vmain*); hence they do not have a characteristic distribution of their own and constitute with the main verb a single verbal domain:

O Pedro [*leu*] $_{\text{VDOMAIN}}$ *o livro* 'Peter read the book'
O Pedro [*acabou por ler*]$_{\text{VDOMAIN}}$ *o livro* 'Peter ended up reading the book'

O prédio [*ruiu*] $_{\text{VDOMAIN}}$ 'The building fell down'
O prédio [*acabou por ruir*] $_{\text{VDOMAIN}}$ 'The building ended up falling down'

Each auxiliary has, however, a characteristic construction, defined by: (a) the preposition (*Prep*) that links the auxiliary to the main verb; in some cases, there is no preposition at all and the *Vaux* is directly linked to the *Vmain*; (b) the non-finite form of the main verb: *Vaux* (Prep) *Vmain*; in Portuguese this non-finite form, can be: a past participle (VPP), a gerund (GER), or a (non inflected) infinitive (VINF)[1]. Therefore, an auxiliary verb construction is to be referred to by these structural elements, for example: *ter* VPP: *ter feito* 'to have done'; *estar* GER: *estar fazendo* 'to be doing'; *acabar por* VINF: *acabar por fazer* 'to end up doing'.

Table 1 illustrates some of the most common auxiliary verb constructions that we have established for European Portuguese. In this table, a literal translation of the auxiliary verb is provided. The next column shows the preposition linking the auxiliary verb to the main verb; if there is no preposition, the symbol <E> is used instead; the non-finite form of the main verb follows, namely the infinitive (W), the gerund (G) or the past participle (K). Finally, the next column identifies approximately the main type of auxiliary, namely, aspectual (VASP), modal (VMOD) and tense (VTEMP) auxiliaries. For each type, a semantic feature is also added, corresponding to the basic grammatical value that the auxiliary conveys. A short example is also given.

[1] Besides bare (or impersonal, or non-inflected) infinitive (e.g. *cantar* 'to sing'), Portuguese also presents an inflected (or personal), v.g. *cantar* (eu, ele/ela) 'to sing_1st,3rd/SG', *cantares* (tu) 'to sing_2nd/SG', etc. The inflected (personal) infinitive cannot have auxiliaries when it functions as a main verb.

Table 1. Auxiliary verb constructions in European Portuguese (extract)

Vaux	*Prep*	*Vmain*	*GRAM*	*Gram:Feature*	*Example*
acabar 'end'	<E>	G	VASP	terminative	*acabou fazendo isso*
acabar 'end'	*a*	W	VASP	terminative	*acabou a fazer isso*
acabar 'end'	*de*	W	VASP	terminative	*acabou de fazer isso*
acabar 'end'	*por*	W	VASP	terminative	*acabou por fazer isso*
andar 'walk, go'	<E>	G	VASP	durative	*anda fazendo isso*
andar 'walk, go'	*a*	W	VASP	durative	*anda a fazer isso*
desatar 'untie'	*a*	W	VASP	inchoative	*desatou a fazer isso*
dever 'must/should'	<E>	W	VMOD	deontic/epistemic	*deve fazer isso*
estar 'be'	<E>	G	VASP	durative	*está fazendo isso*
estar 'be'	*a*	W	VASP	durative	*está a fazer isso*
ficar 'stay'	<E>	G	VASP	durative	*ficou fazendo isso*
ficar 'stay'	*a*	W	VASP	durative	*ficou a fazer isso*
ficar 'stay'	*de*	W	VMOD	obligation	*ficou de fazer isso*
haver 'there be'	<E>	K	VTEMP	past	*havia feito isso*
ir 'go'	<E>	G	VASP	durative	*vai fazendo isso*
ir 'go'	<E>	W	VTEMP	future	*vai fazer isso*
passar 'pass'	*a*	W	VASP	inchoative	*passou a fazer isso*
poder 'may'	<E>	W	VMOD	deontic/epistemic	*pode fazer isso*
pôr-se 'put PROself'	*a*	W	VASP	inchoative	*pôs-se a fazer isso*
ter 'have'	<E>	K	VTEMP	past	*tinha feito isso*
ter 'have'	*de*	W	VMOD	obligation	*teve de fazer isso*
ter 'have'	*que*	W	VMOD	obligation	*teve que fazer isso*
tornar 'do again'	*a*	W	VASP	inchoative /iterative	*tornou a fazer isso*
vir 'come'	<E>	G	VASP	durative	*vem fazendo isso*
vir 'come'	*a*	W	VASP	terminative	*veio a fazer isso*
voltar 'return'	*a*	W	VASP	inchoative /iterative	*voltou a fazer isso*

3.2 Identifying Auxiliary Verb Constructions

Because of its syntactic status, it becomes necessary to distinguish when a sequence of verbs (eventually linked by a preposition) constitutes a single verbal domain, composed by an auxiliary and a main verb, or, else, it is to be dealt with as two distinct verbal domains, each one with a lexical verb [5,10,12,13]. The problem becomes even more interesting for some verbs may function both as lexical and auxiliary verbs. For example, *desatar* 'to untie' is a lexical verb in *O Pedro desatou o nó da gravata* 'Peter untied is tie knot', while it functions as an aspectual (inchoative) auxiliary verb in *O Pedro desatou a correr* 'Peter started running'.

As *Vaux* are transparent to the selection constraints of the main verb, the subject of the auxiliary is the same as the subject of the main verb ([13] approach to English is slightly different from [12] and ours in this respect). By 'same' one does not mean a mere coreferent, although deleted, noun phrase (*NP*), but exactly the same *NP*, which functions as subject both of the *Vaux* and the *Vmain*. Therefore, in order to identify an ordinary juxtaposition of two distinct verb domains (1a-1b) and to distinguish it from

an auxiliary verb construction (2a), it must be possible to insert as subject of the second verb, that is, the *Vmain* candidate, a different subject *NP*, non-coreferent to the sentence's subject (or to the subject of the first verb), that is the *Vaux* candidate (2b-c):

(1a) *O Pedro* [*gostou*]$_{\text{VDOMAIN}}$ *de* [*ler*]$_{\text{VDOMAIN}}$ *o livro*
'Peter loved reading the book'

(1b) *O Pedro* [*gostou*]$_{\text{VDOMAIN}}$ (*de*) *que a Maria* [*lesse*]$_{\text{VDOMAIN}}$ *o livro*
'Peter loved that Mary reading the book'

(2a) *O Pedro* [[*acabou por*]$_{\text{VASP}}$ *ler*]$_{\text{VDOMAIN}}$ *o livro*
'Peter ended up reading the book'

(2b) **O Pedro acabou por a Maria ler o livro*
'Peter ended up Mary reading the book'

(2c) **O Pedro acabou por que a Maria* (*lesse, lia*) *o livro*
'Peter ended up Mary reading the book'

Naturally, for the test sentence to become acceptable, one may have to convert a non-finite (1a) into a finite subordinate clause, either in the subjunctive (1b) or indicative mode, depending on the main verb. In a sentence with a candidate auxiliary verb, one must systematically test the non-coreferent *NP* with the infinitive (2b) and the non-finite subordinate clauses (2c), the latter both in the indicative and the subjunctive mode, as exemplified above.

3.3 Verbal Chains

Auxiliary verbs can operate on another auxiliary verb. This is a recursive, even if limited, process. The string of the auxiliary verbs and the main verb constitute a verbal chain. In a verbal chain, only the first auxiliary is in an inflected form; all subsequent *Vaux* comply with the auxiliary construction definition of the previous *Vaux*:

O Pedro [[*acabou por*] $_{\text{VASP}}$ [*ter de*]$_{\text{VMOD}}$ [*começar a*]$_{\text{VASP}}$ [*ler*]$_{\text{VINF}}$]$_{\text{VDOMAIN}}$ *o livro*
'Peter ended up having $_{\text{VASP}}$ to start reading the book'

As shown in Table 2 below, the most common case consists of a single *Vaux* and the *Vmain*. However, longer chains are often found. So far, the longer verbal chains appearing in the corpus consist of a string of three *Vaux* operating on a copula verb (VCOP), which operates in its turn on a past participle; 22 strings of this length were found, which correspond to passive constructions, such as, for example the following strings:

[[*dever*]$_{\text{VMOD}}$ [*poder*]$_{\text{VMOD}}$[*começar a*]$_{\text{VASP}}$ [*ser*] $_{\text{VCOP}}$] $_{\text{VDOMAIN}}$ [*utilizar*]$_{\text{VCPART}}$
'must/should/would' 'may' 'begin to' 'be' 'used'

[[*dever*]$_{\text{VMOD}}$ [*ter*]$_{\text{VTEMP}}$ [*começar a*]$_{\text{VASP}}$ [*ser*]$_{\text{VCOP}}$]$_{\text{VDOMAIN}}$ [*utilizar*] $_{\text{VCPART}}$
'must/should/would' 'have 'begin to' 'be' 'used'

4 Implementation of *Vaux* in L2F NLP Chain

The identification of *Vaux* in the L2F NLP processing chain [14] is achieved following three steps: (i) auxiliary verbs are tagged according to Table 1; (ii) next, verbal chunks (VASP, VMOD, VTEMP, etc.; see below) are identified and built; and (iii), finally, verbal chains (a main verb and its auxiliaries) are grouped in a VDOMAIN dependency. The chunking and the dependency extraction phases are carried out using the Xerox Incremental Parser [1] with a Portuguese rule-based grammar [15]. Next, this process will be presented in more detail.

Each auxiliary verb is tagged with a syntactic feature representing the different ways it can be used as an auxiliary verb. For example, all tokens associated with the lemma of verb *acabar* 'to end' are tagged with 4 extra tags, which correspond to the information contained in first 4 lines of Table 1:

```
acabar+=[noGerTer:+,aInfTer:+,deInfTer:+,porInfTer:+].
```

The tag `noGerTer` is interpreted by the chunking rules as the syntactic structure of the auxiliary, directly linked to the main verb, that is, without any preposition (`no`); this main verb must be in a gerund form (`Ger`); and the resulting auxiliary construction adds a grammatical (aspectual) terminative (`Ter`) value to the main verb. The remaining tags involve different prepositions (*a* 'to', *de* 'of' and *por* 'by').

Verbal chunks [15] are grouped into 9 categories, which are built in the following order:

VMOD – modal auxiliary verbal chunk;
VTEMP – temporal auxiliary verbal chunk;
VASP – aspectual auxiliary verbal chunk;
VPP – past participle verbal chunk;
VCOP – copulative verbal chunk;
VCPART – past participle verbal chunk (at the end of a verbal chain);
VGER – gerund verbal chunk;
VINF – infinitive verbal chunk;
VF – finite verbal chunk.

For example, the following chunking rules are responsible for building the VTEMP chunks:

```
VTEMP[pass=+] =verb[noPpPas]
  | (pron), advp*, pastpart[masc,sg];vmod{verb[inf]} |.
VTEMP[futu=+] =verb[noInfFut]
  | (pron), advp*, verb[inf];vmod{verb[inf]} |.
VTEMP[futu=+] =verb[aInfFut]
  |(pron),advp*, prep[lemma:a],verb[inf];vmod{verb[inf]}|.
```

The first one builds a VTEMP chunk with a verb tagged with the `noPpPas` feature, if it is followed by one optional pronoun, that can be followed by zero or more adverbs, and must be followed by a verb that is a past participle in the masculine-singular form or a VMOD chunk that has as its first constituent an infinitive verb. The components

that are inside vertical bars do not belong to the VTEMP chunk, but their existence is a requirement to fire the rule.

Afterwards, when dependencies between chunks are extracted, the following dependencies are computed:

HEAD – a binary dependency linking each chunk node to its most important constituent, the verb when referring to verbal chunks;

VLINK – a binary dependency that connects the heads of the following chunks:
(VASP;VPP)→(VINF;VGER;VCOP), VCOP→VCPART,
(VMOD;VTEMP)→Vchunk, (VINF;VGER)→VINF
(whereas Vchunk represents any verbal chunk, and the symbol ';' stands for disjunction). VLINKS are used to compute VDOMAIN dependencies; and

VDOMAIN – A binary dependency between the head of the first verb and the head of the last verb of a verbal chain. This dependency is built by recursively joining VLINKs. Fig. 1 shows the verbal chunks, their heads and the dependencies extracted from the sentence *O Pedro acabou por ter de começar a ler o livro* 'Peter end up having to start reading the book':

O Pedro [*acabou por*]$_{\text{VASP}}$ [*ter de*]$_{\text{VMOD}}$ [*começar a*]$_{\text{VASP}}$ [*ler*]$_{\text{VINF}}$ *o livro*

```
HEAD(acabou, acabou por)
HEAD(ter, ter de)
HEAD(começar, começar a)
HEAD(ler, ler)
VLINK(acabou,ter)
VLINK(ter,começar)
VLINK(começar,ler)
VDOMAIN(acabou,ler)
```

Fig. 1. Verbal chunks and dependencies

5 Results

Table 2 provides the counts of strings formed of a *Vaux* chain with a main verb on the 180 million word *CETEMPúblico* corpus[2]. The first line of this table shows the number of simple verbs (i.e. verb domains composed of only the main verb and without any auxiliary), while the remaining lines show strings of auxiliaries and a main verb by crescent length. These combinations, which we call a verb domain, total 2,748,061 occurrences, that is, about 30% of all verbal structures.

It is possible to assess the different patterns formed by sequences of several *Vaux* types. For example, Table 3 shows the number of patterns with two *Vaux* and a *Vmain* by type sequence. Naturally, the particular *Vaux* in each combination could also be retrieved. In-depth analysis of these results may contribute to model the syntactic-semantic combinations of different *Vaux* grammatical values, necessary for idiomatic text generation.

[2] http://www.linguateca.pt/cetempublico/

Table 2. Number of verb chains in the CETEMPúblico corpus with string of auxiliaries by length (or number) of *Vaux* in the string

string length	*count*	*%*
1 verb (only the main verb)	6,399,481	0.699,585
2 verbs (one auxiliary and the main verb)	2,538,296	0.277,484
3 verbs (string of auxiliaries and the main verb)	206,556	0.022,580
4 verbs	3,187	0.000,348
5 verbs	22	0.000,002
Total	9,147,542	

Table 3. Number of patterns with different combinations of *Vaux* types (extract)

Vaux1-Vaux2-Vmain	*Count*
VMOD-VCOP-VPP	87,889
VASP-VCOP-VPP	55,560
VTEMP-VCOP-VPP	19,034
VMOD-VTEMP-VPP	14,935
VMOD-VASP-VINF	13,416
VTEMP-VASP-VINF	12,856
…	…
Total	206,556

6 Evaluation

It would be out of the scope of this paper to fully evaluate the performance of the NLP chain, for it would mean assessing 2,8M verb chains, resulting from the task of identifying and extracting auxiliary verb chains on the entire CETEMPúblico corpus. Therefore, we only present a preliminary evaluation of the system performance on a relatively small text (20,296 words), namely the Portuguese blog *De Rerum Natura* [3]. This a scientific divulgation/discussion weblog, which includes very carefully written texts, from several authors, about different, science-related, subject matters and which consist of above-average ratio of scientific vocabulary.

The corpus is composed of 672 sentences, from which 1,464 VDOMAIN dependencies were extracted (notice that a sentence may contain several VDOMAIN). These included 192 VLINK. The corpus was manually checked for parsing errors. All VLINK dependencies were correctly extracted, however, 9 VLINK were not found, and therefore the system attained a 0,955 recall. As a result, the number of VDOMAIN dependencies extracted (1,463) is higher than the real number (1,455), but this still represents a 0,994 precision for this dependency.

[3] http://blogs.publico.pt/dererumnatura/

Some of the missed VLINK dependencies are due to incorrect POS tagging. See, for example, sentence #140, below, where the infinitive *ser* 'to be' has been wrongly tagged as the noun *ser* 'being':

```
VLINK(tem,vindo), VDOMAIN(tem,vindo)
140> […] NP{um trabalho} SC{que VTEMP{tem}} VPP{vindo} NP{a ser}
AP{desenvolvido} PP{por NOUN{Gil Alterovitz}}[…]
```

Other errors are due to insertions, like in the following sentence, where the (compound) adverbial phrase between comas prevent the linking of VTEMP *vão* to the main verb *fazer* 'to do':

…factores proteicos que, interagindo entre si, ***vão****, em última instância,* ***fazer*** *com que umas células se diferenciem…* 'protein factor that, by interacting among themselves, will, ultimately, make some cellules to differentiate'

Notice, however, that other type of insertions and clitic pronouns do not affect the correct parsing of the verbal chains, as it is illustrated by the following (shortened) sentences:

```
VLINK(poder,pisar), VDOMAIN(poder,pisar)

207>TOP{Para SC{quem NP{toda a vida} VF{viu}} NP{aquilo} PP{a o
abandono} , VMOD{poder} ADVP{agora} VINF{pisar} NP{o seu pátio}
AP{magnífico} , […]


VLINK(posso,deixar), VLINK(deixar,indignar)

VDOMAIN(posso,indignar)

647>TOP{" ADVP{não} VMOD{posso} VASP{deixar de} NP{me} VINF{indignar}
PP{com algumas} PP{de as suas determinações} , […]
```

7 Conclusions and Future Work

This paper presented a solution fully integrated in a NLP processing chain for the correct parsing of auxiliary verbal chains in European Portuguese. The system adequately chunks and correctly retrieves the set of dependencies that allow for the precise identification of verbal domains, i.e. strings composed of a main verb and its entire verbal chain of auxiliaries. In general, insertions and different clitic pronouns' attachment do not hinder the parser's performance. Preliminary results over a small 20K word corpus are most encouraging. In the near future, the authors propose to extend further the list of auxiliaries already identified, evaluate a significant sample of the CETEMPúblico corpus on the context of different NLP tasks. Extension to the Brazilian variety of Portuguese, with minor adaptations, is also being envisaged.

References

1. Ait-Mokhtar, S., Chanod, J., Roux, C.: Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering 8(2/3), 121–144 (2002)
2. Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: An FST grammar for verb chain transfer in a Spanish-Basque MT System. In: Proceedings of the FSMNLP Workshop. ACL, Helsinki (2005)
3. Anderson, G.: Auxiliary Verb Constructions. Oxford University Press, Oxford (2006)
4. Bick, E.: Automatic Parsing of Portuguese. In: PROPOR 1996, Curitiba (1996)
5. Brito, A.M.: O Sintagma Verbal. In: Mateus, M.H., et al. (eds.) Gramática da Língua Portuguesa, pp. 403–417. Caminho, Lisboa (2003)
6. Casteleiro, J.: Sintaxe transformacional do adjectivo. INIC, Lisboa (1981)
7. Coheur, M.: Uma interface sintaxe/semântica baseada em regras hirarquicamente organizadas, PhD. Thesis. IST/UTL, Lisboa (2004)
8. Cunha, C., Cintra, L.: Nova Gramática do Português Contemporâneo. Ed.João Sá da Costa, Lisboa (1984)
9. Gagnon, M., Coelho, E., Finger, R.: Towards a formalization of Tense and Aspect for the Generation of Portuguese Sentences. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 229–232. Springer, Heidelberg (2006)
10. Gonçalves, A.: Para uma sintaxe dos verbos auxiliaries em Português Europeu, M.A.Thesis, FLUL, Lisbon (1992)
11. Graça, J., Pardal, J., Coheur, L., Caseiro, D.: Building a golden collection of parallel Multi-Language Word Alignments. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco (2008)
12. Gross, M.: Compound Tenses in English. Lingvisticae Investigationes 22, 71–122 (1999)
13. Harris, Z.S.: A Teory of Language and Information: A Mathematical Approach. Clarendon, Oxford (1991)
14. Mamede, N.: A cadeia de processamento de língua natural do L2F. Relatório Técnico Interno do L2F. INESC-ID Lisboa, Lisboa (to appear)
15. Mamede, N., Baptista, J., Vaz, P., Hagège, C.: Nomenclature of chunks and dependencies in Portuguese XIP grammar 2.1. Relatório Técnico Interno do L2F. INESC-ID Lisboa, Lisboa (2007)
16. Mitkov, R.: Anaphora Resolution. Pearson Longman, UK (2002)
17. Pontes, E.: Verbos Auxiliares em Português. Ed. Vozes, Petrópolis (1973)
18. Ranchhod, E.: Remarks on the Complementation of Aspectual Verbs. Lingvisticae Investigationes Supplementa 24, 423–438 (2004)
19. Ranchhod, E., Carvalho, P., Mota, C., Barreiro, A.: Portuguese Large-scale Language Resources for NLP Applications. In: Proceedings of LREC 2004, pp. 1755–1758 (2004)
20. Steele, S.: Auxiliaries. In: Brown, K., Miller, J. (eds.) Concise Encyclopedia of Grammatical Categories. Elsevier/Pergamon, Cambridge (1999)

P-AWL: Academic Word List for Portuguese

Jorge Baptista, Neuza Costa, Joaquim Guerra,
Marcos Zampieri, Maria Cabral, and Nuno Mamede

Universidade do Algarve, Campus de Gambelas P 8005-139 Faro, Portugal
Spoken Language Laboratory, INESC ID Lisboa, R. Alves Redol 9, P 1000-029 Lisboa, Portugal
{jbaptis,ncosta,jguerra,mcabral@ualg.pt},
marcos_zampieri@uol.com.br, Nuno.Mamede@l2f.inesc-id.pt

Abstract. This paper presents and discusses the methodology for the construction of an Academic Word List for Portuguese: PAWL, inspired in its English equivalent. The aim of this linguistic resource is to provide a solid base for future studies and applications on Computer Assisted Language Learning, while maintaining comparability with other comparable resources.

Keywords: Portuguese, academic word list, Computer Assisted Language Learning.

1 Introduction

Computer Assisted Language Learning (CALL) is a growing area of research, in the intersection of Computational Engineering, Linguistics and Language Teaching [8]. One of the important applications developed on this field is the (automatic) assessment of students' language proficiency or of his/her lexical knowledge [16]. The later is usually based on carefully selected subset of the lexicon, which is deemed relevant for a specific purpose; for example, in view of enrolling the students in an adequate level at language courses [13].

The English Academic Word List [5], henceforward E-AWL, is the academic vocabulary most widely used today in language teaching, testing and materials development. E-AWL defines a subset of the English lexicon whose mastery is considered necessary for students at University level. E-AWL has been extensively used in many Natural Language Processing (NLP) studies [9,19], and more recently, in CALL research on reading practice [7,11,15]. To our knowledge, there is still no equivalent resource for Portuguese. Still, there is a significant awareness that university students are required to master a basic 'academese' vocabulary as an indispensable tool for their reading and writing practices [14]. The lack of such standard resource also makes difficult a consensual assessment of students' proficiency/lexical knowledge, especially, but not exclusively, for Portuguese as Second Language courses.

The purpose of this paper is to present and discuss the methodology followed in the construction of Portuguese-Academic Word List (P-AWL). P-AWL is a general purpose vocabulary, with current (but not colloquial) words, which has been designed

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 120–123, 2010.

for immediate application on a CALL web-based environment, currently devoted to improve students' reading practice and vocabulary acquisition [4,13].

2 Issues on Selecting Vocabulary

There is probably no consensus about how an academic word list should be drawn [12,17]. Corpus-based approaches may contribute to define a vocabulary intersecting different academic genres and subject matters [3,19]. For European Portuguese, there is still no freely available, digital repository of graduation/post-graduation monographs (theses). Therefore, establishing a (balanced) corpus for deriving the vocabulary intersection, across scientific domains and genres, may well be an impossible mission. Even if a particular list could be obtainable by some corpus-based methods, an inevitable selection would still have to be done, since frequency-based vocabulary is very unlikely to produce the balanced and homogeneous word lists adequately targeted for such specific purposes, such as language learning and vocabulary assessment. For example, consider French Dubois-Buise Scale [18]. Because of the (quantitative) methods used to produce this scale, certain relatively homogenous grammatical categories, such as numerals or the names of months, have been attributed not only different ranks in the scale, but they were deemed to belong to different teaching levels. This is direct contradiction with well established evidence (often reflected in teaching practices) that people structure their vocabulary in an associative manner, by grasping semantically or domain-related related words at the same [1, 6].

This paper assumes in full the manually crafting of an academic word list with the inevitable subjective nature of any vocabulary selection. It proposes, however, a set of provisions on procedures and methodology to minimize individual bias in that selection [10]. Our goal is that P-AWL may be used not only as a means of assessment of the students but also a tool to aid in the acquisition of the Portuguese language, hence permitting students to further develop their proficiency in the various levels of the language acquisition process and therefore become increasingly successful as Portuguese speakers.

3 Methods

Taking E-AWL as a starting point, and in order to select the adequate vocabulary for P-AWL, an initial list of 2,145 (2,136 different) entries was drawn adopting the following main criteria: 1. As far as possible, all meaning units of E-AWL were kept in the P-AWL; 2. As far as possible, the Portuguese entry or node-word is cognate of the English entry; 3. Where multiple senses or synonyms were involved, several entries were established for Portuguese; Brazilian Portuguese (BR) and European Portuguese (LUS) orthographic variants were systematically contrasted (*gênero*/*género* 'gender, type, kind') but were not split into different entries; 4. The morphologic family of each entry was systematically drawn; 5. For each word (lemma), a part-of-speech (POS) and inflectional codes (FLEX) were given.

A team of 4 annotators included linguists and language teachers, from different scientific backgrounds, including a native-speaker Brazilian linguist applied the

selection criteria to the initial list, through a two-round selection procedure. After this, a 85,6% inter-annotator agreement was achieved. The final list for P-AWL was obtained by selecting all words with less than 2 'discard' marks (i.e. only consensual or with only minimal disagreement vocabulary) and by carefully checking the consistency of that list (in order to avoid word/headword repetitions or incorrect pairings).

Table 1. Breakdown of P-AWL by PoS

PoS	*Count*
Noun	754
Adjective	451
Verb	409
Adverb	203
Conjunction	4
Preposition	2
Total	1812

In its current form, P-AWL contains 1,823 entries. Each entry of P-AWL is classified for its part-of-speech (PoS) category. Table 1 shows the breakdown of P-AWL by part-of-speech category. Words from the same morphological family (derivates) are grouped together under a headword. Currently, there are 814 headwords and each morphological family consists of an average of 2.23 words. Different word senses were distinguished using [2] semantic tags for nouns. The task of semantic annotation of P-AWL entries is still ongoing.

4 Conclusions and Future Work

This paper presented a new linguistic resource — Portuguese Academic Word List (P-AWL) — specifically built to be used in language tutoring system, but that can also be of interest for the Portuguese NLP community. It was directly inspired in English AWL [5], but its content and detail make it more comprehensive in scope and aim. Future work will include the completion of the semantic annotation of P-AWL nominal entries and calibrating the list by using frequency information from different sources.

Acknowledgments. Research was funded by project REAP.PT ref. CMU-PT/HumMach/0053/2008.

References

1. Bauer, L., Nation, P.: Word Families. International Journal of Lexicography 6(4), 253–279 (1993)
2. Bick, E.: Noun Sense Tagging: Semantic Prototype Annotation of a Portuguese Treebank. In: Hajic, J., Nivre, J. (eds.) Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic, December 1-2, pp. 127–138 (2006)

3. Chen, Q., Guang-Chun, G.: A Corpus-Based Lexical Study on Frequency and Distribution of Coxhead's AWL Word Families in Medical Research Articles (RAs). English for Specific Purposes 26(4), 502–514 (2007)
4. Collins-Thompson, K., Callan, J.: Information retrieval for language tutoring: An overview of the REAP project. In: Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK (2004)
5. Coxhead, A.: A New Academic Word List. TESOL, Quarterly 34(2), 213–238 (2000)
6. Fry, E.: The Vocabulary Teacher's Book of Lists. Jossey Bass (2004)
7. Heilman, M., Zhao, L., Pino, J., Eskenazi, M.: Retrieval of Reading Materials for Vocabulary and Reading Practice. In: 3rd Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics (2008)
8. Hubbard, P.: Call and future of language teacher education. CALICO Journal (2008)
9. Kulkarni, A., Heilman, M., Eskenazi, M., Callan, J.: Word Sense Disambiguation for Vocabulary Learning. In: Ninth International Conference on Intelligent Tutoring Systems (2008)
10. Laporte, E.: Lexicons and Grammars for Language Processing: Industrial or Handcrafted Products? In: Rezende, L., Dias da Silva, B., Barbosa, J. (eds.) Léxico e gramática: dos sentidos à construção da significação, Cultura Académica, São Paulo Brazil, pp. 51–83 (2009)
11. Liou, H., Chang, J., Kuo, C., Chen, H., Chang, C.: Web-based Academic English Course Design and Materials Development. In: International English Teachers' Association Symposium, Chieh-Tan Activity Center, Taipei, November 12 (2005)
12. Nation, P.: Learning Vocabulary in another Language. Cambridge University Press, Cambridge (2001)
13. Marujo, L., Lopes, J., Mamede, N.J., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J., Viana, C.: Porting REAP to European Portuguese. In: ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009), ISCA, Wroxall Abbey Estate, Warwickshire (September 2009)
14. Paquot, M.: Towards A Productively-Oriented Academic Word List. In: PALC 2005 Practical Applications in Language and Computers, Łódź, Poland, April 7-9 (2005)
15. Pino, J., Eskenazi, M.: An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings. In: Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL/HLT (2009)
16. Read, J.: Assessing Vocabulary. Cambridge University Press, New York (2000)
17. Schmitt, N.: Vocabulary In Language Teaching. Cambridge University Press, New York (2000)
18. Ters, F., Mayer, G., Reichenbach, D.: L'Echelle Dubois-Buyse. Editions M.D.I (1995)
19. Vongpumivitcha, V., Huang, J., Chang, Y.: Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. English for Specific Purposes 25(1) (2008)

Automatic Phone Clustering Based on Confusion Matrices

Carla Lopes[1,2], Arlindo Veiga[1], and Fernando Perdigão[1,3]

[1] Instituto de Telecomunicações
[2] Instituto Politécnico de Leiria-ESTG
[3] Universidade de Coimbra – DEEC, Pólo II, P-3030-290 Coimbra, Portugal
{calopes,aveiga,fp}@co.it.pt

Abstract. Phone recognition experiments give information about the confusions between phones. Grouping the most confusable phones and making a multilevel hierarchical classification should improve phone recognition. In this paper a clustering method is investigated, based on phone confusion matrix, for the data-driven generation of phonetic broad classes (PBC) of the Portuguese language. The method is based on a statistical similarity measurement rather than acoustical/phonetic knowledge. Results are presented for two phone recognisers (TIMIT corpus and Portuguese TECNOVOZ database).

Keywords: Phone recognition, phone clustering.

1 Introduction

Besides the phone is the main unit of recognition on Automatic Speech Recognition (ASR) systems, there are phones with similar acoustic/phonetic characteristics that are easily confused. It has been pointed out [1-3], that intermediate representations (Phone Broad Classes) between the speech signal and the corresponding phonetic units, may be used to help classification and improve recognition. PBC classification refers to grouping phones into categories. This classification may be done manually (knowledge-driven) or automatically (data-driven). In the case of an expert-based approach, the categories are selected according to acoustic-phonetic properties that derive from articulatory knowledge or human hearing perception. In data-driven approaches the classes usually have origin in clustering algorithms that uses some similarity or distance measure between phones. PBC for the English language is an issue that has been widely addressed by the scientific community, [2-4]. Its definition is usually related to manner and place of articulation, so that all the classes show strong agreement within some phonetic, articulatory and/or acoustic properties. The efficient construction of small phone sets is not trivial, what is confirmed by the lack of consensus between several proposals, [1],[5]. The subjective influence of the expert affects the number of classes and the set of phones of each category. In order to overcome this problem, this paper explores an automatic data-driven method where the broad classes are defined according to the output of a phone recognition system. The similarity between phones derives from the confusion matrix using a similarity measure. A clustering method to group the most similar phones is also proposed.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 124–127, 2010.

2 Similarity Measure

All clustering algorithms depends on a proximity or distance measure that can be obtained from the acoustic models or rely on the confusion matrix [6],[7]. In confusions-driven methods, the similarity measure approach comes from the phone confusion matrix, which is computed from the output of a phone recogniser by aligning the recognised sentence with the reference one, using a dynamic programming algorithm. Zgank, [7], define a phone similarity measure based on the confusions between phones, but dependent of a threshold. In [6] the confusion matrix is converted into a symmetric similarity matrix using the so called *Houtgast* measure. It measures the similarity between models *i* and *j* using the confusions between phones *i* and *j* to all other phones *k*. So, if *N* is the total number of classes (phones), the *Houtgast* similarity between phones *i* and *j*, s_{ij} is given by $s_{ij} = s_{ji} = \sum_{k=1}^{N} \min\left(f_{ik}, f_{jk}\right)$ with $1 \le i, j \le N$. f_{ij} is the number of confusions between phone *i* and phone *j*. According to this measure, two phones *i* and *j* are similar if they both have many confusions with the same phones and its similarity is zero only when phone *i* and *j* have no simultaneous confusions with all phones.

3 Phone Distance Measure and Clustering Method

In this section we define a phone distance based on the confusion matrix. The out-of -diagonal values represent misclassified phones. Phones with similar acoustic/phonetic characteristics tend to be more confused, so we start from these misclassifications to define a phone similarity measure. Since the number of occurrence of each phone is quite different, we normalise the confusion matrix by dividing the frequency counts f_{ij} by the total number of occurrences in the speech data of the phone *i*:

$$p_{ij} = f_{ij} \Big/ \sum_{n=1}^{N} f_{in} = P\left(\hat{c}_j \mid c_i\right), \tag{1}$$

resulting on a estimative of $P(\hat{c}_j \mid c_i)$, the probability of recognising a phone c_j when the phone class is c_i. The previous *Houtgast* similarity measure become $s'_{ij} = \sum_{n=1}^{N} \min\left(p_{in}, p_{jn}\right) = \sum_{n=1}^{N} \min\left(P(\hat{c}_i \mid c_n), P(\hat{c}_j \mid c_n)\right)$ where $s'_{ij} \le 1$ and $s'_{ii} = 1$. A distance measure can therefore be defined as

$$d_1(c_i, c_j) = 1 - s'_{ij} = \tfrac{1}{2} \sum_{n=0}^{N} \left| p_{in} - p_{jn} \right|. \tag{2}$$

This results since $\min(a,b) = \frac{1}{2}\left(a + b - |a - b|\right)$ and $\sum_{1}^{N} p_{in} = 1$. This distance forms a metric because it is based on the L_1 norm applied to row differences of matrix P.

Using the proposed phone distance measure, a clustering method is applied grouping phones in a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. Clustering can be done using the following steps, where initially each phone will be considered as a distinct cluster.

Step 1→Compute matrix P using (1).
Step 2→Find the distance between every pair of phones using (2).
Step 3→Compute the single distances between all clusters. The distances between cluster *r* and *s* can be computed with several criterions, being the simplest the nearest neighbour: $d(r,s) = \min\left(d\left(c_{ilr}, c_{jls}\right)\right), i = 1..n_r; j = 1..n_s$ where n_k is the number of phones in cluster *k* and c_{ilk} is the i^{th} phone in cluster *k*.
Step4→Create a new cluster grouping the two nearest clusters.
Repeat from Step 3 until the number of desired clusters or a distance threshold is reached.

4 Experiments

In order to evaluate the clustering capabilities of the proposed data-driven clustering method, phone recognition results from two different recognisers were used: a hybrid MLP/HMM system, trained with English speech data from TIMIT database,[8] and a HMM system trained with European Portuguese TECNOVOZ speech data, [9].

In TIMIT phone recognition task**,** data-driven PBC generation starts from a confusion matrix obtained by testing all the TIMIT original training set. Evaluation, on the 39 phones, was done by means of the *Correctness* and *Accuracy* rates that reached 81.37% and 79.74%, respectively. Following the clustering steps proposed, and using the nearest neighbour distance to clusters, we arrive at a hierarchical phone clustering, depicted as a binary cluster tree (dendrogram) on Figure 1a). The labels along the horizontal axis represent the phones in the original data set and vertical axis refers to the distance between the phones, $d_1(c_i,c_j)$.

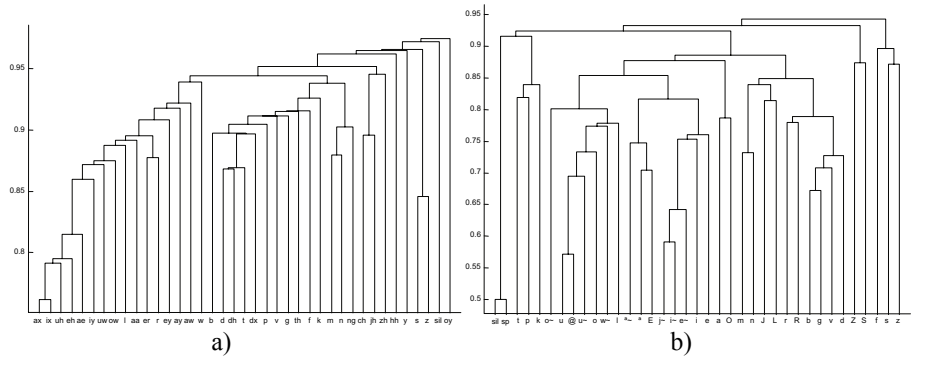

Fig. 1. Hierarchical phone clustering dendrogram for: a) TIMIT and b) TECNOVOZ

It is notorious the consistence of the clusters, namely the ones involving vowels. If we compare it with the knowledge-based division, proposed in [3], only [y], [w] and [oy] are in a separated cluster (not because it is acoustically different but due to the few confusions with other phones). Another strong cluster corresponds to nasals where data-driven division and knowledge-based division is the same. Affricatives are well set in a separate cluster. Regarding fricatives and stops, the method rely on more that two clusters. The number of confusions between some of them suggests that acoustically they present similarities.

TECNOVOZ is a European Portuguese speech database, [9], collected in 2007, that includes commands and phonetically rich read sentences. The taken confusion matrix derives from a context free phone recogniser, trained with the sentence utterances (20,364) from the database. To describe the Portuguese language 37 phones were used including a silence and a short pause model. In the test set *Correctnes* reached 63.85% and *Accuracy* 59.46%. The dendogram, using the average distance to clusters, is in figure 1b). It is interesting to note that nasals and other vowels are not separated. Fricatives /s/, /S/, /f/ and /Z/ remain in separated clusters, again because the corresponding models are accurate and do not need to be clustered.

5 Conclusions and Discussion

A new confusion-driven method for the generation of PBC was defined. Phones are classified into particular classes according to their distance, determined by a phone confusion matrix. In this proposal no expert knowledge is needed, which is often unavailable and can introduce subjective influence, and do not depend on empirically defined thresholds. The method can easily be generalized and applied to other languages or to multilingual systems. The results, regarding TIMIT database show strong agreement in vowels and nasals with knowledge-driven division. In relation to TECNOVOZ database nasals and some vowels tend to cluster.

Besides the promising results of the proposed data-driven clustering method, the method is highly dependent on the performance of the recogniser. The quality of the acoustic models must be good enough to provide a warrantable confusion matrix.

Acknowledgments. Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant (SFRH/BD/27966/2006).

References

[1] Halberstadt, A., Glass, J.: Heterogeneous Acoustic Measurements for Phonetic Classification. In: Eurospeech (1997)
[2] Lopes, C., Perdigão, F.: A Hierarchical Broad-Class Classification To Enhance Phoneme Recognition. In: Proc. EUSIPCO 2009, Glasgow, UK (August 2009)
[3] Scanlon, P., Ellis, D., Reilly, R.: Using Broad Phonetic Group Experts for Improved Speech Recognition. IEEE TASLProcessing 15(3), 803–812 (2007)
[4] Rose, R.C., Momayyez, P.: Integration of multiple feature sets for reducing ambiguity in ASR. In: Proc of ICASSP 2007 (April 2007)
[5] Ali, A.M.A., Van der Spiegal, J., Mueller, P.: Acoustic-Phonetic Features for the Automatic Classification of Stop Consonants. IEEE TSAP 9(8) (November 2001)
[6] Shih, P.-Y., et al.: Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition. In: SUTC 2008, pp. 500–506 (2008)
[7] Zgank, A., Horvat, B., Kacic, Z.: Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. Speech Commun. 47(3), 379–393 (2005)
[8] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N.: DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST (1990)
[9] Lopes, J., et al.: Development of a Speech Recognizer with the Tecnovoz Database. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 260–263. Springer, Heidelberg (2008)

An Open-Source Speech Recognizer for Brazilian Portuguese with a Windows Programming Interface

Patrick Silva, Pedro Batista, Nelson Neto, and Aldebaro Klautau

Universidade Federal do Pará, Signal Processing Laboratory,
Rua Augusto Correa. 1, 660750110 Belém, PA, Brazil
{patrickalves,pedro,nelsonneto,aldebaro}@ufpa.br
http://www.laps.ufpa.br

Abstract. This work is part of the effort to develop a speech recognition system for Brazilian Portuguese. The resources for the training and test stages of this system, such as corpora, pronunciation dictionary, language and acoustic models, are publicly available. Here, an application programming interface is proposed in order to facilitate using the open-source Julius speech decoder. Performance tests are presented, comparing the developed systems with a commercial software.

Keywords: Continuous speech recognition, Brazilian Portuguese resources, application programming interface.

1 Introduction

This contribution describes an on-going work concerning the development of a speech recognition system for Brazilian Portuguese (BP). The goal is to implement a large-vocabulary continuous speech recognition system (LVCSR), capable of operating in real-time using Julius [1]. The Julius speech recognition engine is a high-performance LVCSR decoder for speech-related applications. However, incorporating Julius to an application targeting the Windows operating system is not trivial. This work presents a simple application programming interface (API) to ease the task of developing applications based on speech recognition.

2 Developed Resources

In [2], the authors have described a grapheme-to-phoneme converter with stress determination for BP, based on a set of rules. The resulting phonetic dictionary has over 65 thousand words. It is well-known that the amount of training data can drastically influence the performance of an automatic speech recognition system. This motivates the inclusion of a brief description of our BP speech and text corpora [3]. The LapsStory is a corpus based on audiobooks with 6 speakers and more than 14 hours of audio. The files were manually segmented to create

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 128–131, 2010.

smaller audio files. The acoustic environment in audiobooks is very controlled, so the audio files have no audible noise and high signal to noise ratio. The Spoltech corpus was also used, but it required revision and correction of multiple files, as described in [4]. After revision the modified Spoltech was composed by 477 speakers, which corresponds to 4.3 hours of audio. The Spoltech's acoustic environment was not controlled, in order to allow for background conditions that would occur in application environments. To complement Spoltech and LapsStory and get a benchmark reference corpus for testing BP systems, the LapsBenchmark is under construction. Currently, such corpus has data from 35 speakers with 20 sentences each, which corresponds to approximately 54 minutes of audio. The LapsBenchmark's recordings were performed on computers using common desktop microphones and the acoustic environment was not controlled.

3 Baseline System

The front-end consists of the widely used 12 mel-frequency cepstral coefficients (MFCCs). The initial acoustic models for the 39 phones (38 monophones and a silence model) used 3-state left-to-right HMMs. After that, cross-word triphone models were built from the monophone models and a decision tree was designed for tying triphones with similar characteristics. After each step, the models were reestimated using the Baum-Welch algorithm via HTK tools [5]. A more detailed description of the HMMs training process can be seen in [4].

The SRILM tools [6] were used to build the n-gram language models. The number of sentences used to train the language models was 1.6 million with approximately 65 thousand distinct words from CETENFolha and our BP text corpora. A trigram language model with 25.8 million words and perplexity 169 was designed with Kneser-Ney smoothing. The amount of 10 thousand sentences, unseen during the training phase, were used to measure the perplexity.

The HTK was used and two adaptation techniques are adopted [5]. The first is the maximum likelihood linear regression (MLLR), which computes a set of transformations that aim to reduce the mismatch between an initial model set and the adaptation data. The second technique uses the maximum a posteriori (MAP) approach. In the HTK implementation, the mean of each Gaussian is updated by MAP using the average of the prior distribution, the weights of the Gaussian and the data of the speaker.

4 Proposed Application Programming Interface

While promoting the widespread development of applications based on speech recognition, the authors noted that it was not enough to make available resources such as language models. These resources are useful for speech scientists but most programmers demand an easy-to-use API. Hence, it was necessary to complement the documentation and code that is part of the Julius package. An API for Windows was developed in the C++ programming language with

the Common Language Runtime specification, which enables communication between the languages supported by the .NET platform. The proposed API allows the real-time control of the Julius and the audio interface. Julius is responsible for producing the speech parameterization to perform the recognition and was chosen to be the decoder used by the proposed API because of its flexible license.

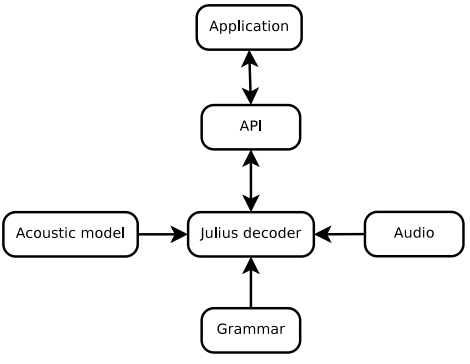

Fig. 1. API interaction model

Since the API supports the component object model automation, most high level languages (e.g., C#, Visual Basic, and others), can be used to write applications. As shown in Fig. 1, the API enables applications to control aspects of the Julius decoder, from which the application can load the acoustic and language models to be used, start and stop recognition, receive events and recognition results with an associated confidence measure. In essence, a speech application must create, load, and activate a grammar, which essentially indicates what type of utterances to recognize, i.e., dictation or context-free grammar.

5 Experiments Results

The acoustic model was initially trained using the LapsStory corpus and then it was adapted using the MLLR and MAP techniques with the Spoltech corpus [3]. Both techniques were used in supervised training (offline). A comparison was made with another decoder: HDecode (part of HTK) and the commercial software IBM ViaVoice [7]. The LapsBenchmark corpus was used to evaluate the systems. The used performance measures were the correct words rate (CWR) and real-time factor (xRT). The evaluation process was carried out in two stages: speaker independent and dependent models.The results are shown in Table 1.

The IBM ViaVoice requires a session of speaker adaptation. Hence, for the first stage, the speaker adaptation process for ViaVoice was carried out using the voice of six speakers, 3 men and 3 women, which corresponds to 10 minutes of audio. We could not measure the xRT value from ViaVoice due to the adopted procedure for invoking the recognizer in batch. Note that HDecode and ViaVoice

Table 1. Systems comparison using speaker independent and dependent models

Decoder	Independent models		Dependent models	
	CWR(%)	xRT	CWR(%)	xRT
Julius	60.42	0.7	77,7	0.7
HDecode	70.63	0.9	84,6	0.8
IBM ViaVoice	70.71	-	82.7	-

had almost the same performance, while Julius had worse performance, but it was faster. Two speakers were used in the dependent process, which corresponds to 10 minutes of each voice. The MLLR and MAP adaptation techniques were again used for the models adopted in Julius and HDecode. In the case of IBM ViaVoice, the standard adaptation process (guided by the software) was adopted. HDecode showed satisfactory results when compared to ViaVoice. Although the Julius decoder presented the worst performance in both tests, if tuned more carefully, it can eventually outperform HDecode, as described in [8].

6 Final Considerations

This paper presented a speech recognition system for BP, including the resources, a Julius-based engine and an API for Windows. The resources were made publicly available [9] and allow for reproducing results across different sites. Future work includes making Julius complaint to the Microsoft Speech API while using BP.

References

1. http://julius.sourceforge.jp/en/ (Visited in May 2009)
2. Siravenha, A., Neto, N., Macedo, V., Klautau, A.: Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em português brasileiro. In: 7th International Information and Telecommunication Technologies Symposium (2008)
3. Silva, P., Neto, N., Klautau, A.: Novos recursos e utilização de adaptação de locutor no desenvolvimento de um sistema de reconhecimento de voz para o português brasileiro. In: XXVII Simpósio Brasileiro de Telecomunicações (2009)
4. Silva, P., Neto, N., Klautau, A., Adami, A., Trancoso, I.: Speech recognition for brazilian portuguese using the spoltech and OGI-22 corpora. In: XXVI Simpósio Brasileiro de Telecomunicacões - SBrt 2008 (2008)
5. Young, S., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, Cambridge (2006)
6. Stolcke, A.: SRILM an extensible language modeling toolkit. In: Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado (2002)
7. http://www.ibm.com/software/speech/ (Visited in September 2009)
8. Rotovnik, T., Maucec, M.S., Horvat, B., Kacic, Z.: A comparison of HTK, ISIP and julius in slovenian large vocabulary continuous speech recognition. In: 7th International Conference on Spoken Language Processing, ICSLP (2002)
9. http://www.laps.ufpa.br/falabrasil (Visited in October 2009)

A Baseline System for Continuous Speech Recognition of Brazilian Portuguese Using the West Point Brazilian Portuguese Speech Corpus

Fabiano Weimar dos Santos[1], Dante Augusto Couto Barone[1], and André Gustavo Adami[2]

[1] Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil
[2] Universidade de Caxias do Sul (UCS)
Rua Francisco Getúlio Vargas, 1130 – CEP 95070-560 – Caxias do Sul – RS – Brasil

Abstract. Despite the availability of several speech corpora that can be used to build automatic speech recognition systems, there are only a few corpora for the Brazilian Portuguese (BP) language. This lack of corpora does not allow an extensive and deep research on continuous speech recognition systems for BP. In this work, we present a baseline system for continuous speech recognition for BP and its results using the West Point Brazilian Portuguese Corpus. In addition to the results, the resources developed to build the system are made available for continuing the research on such systems for BP.

1 Introduction

The lack of Brazilian Portuguese corpora and resources has been an obstacle for research and development of continuous speech recognition systems for BP. In fact, to the best of our knowledge, there is no baseline system in BP to compare results. Besides the lack of corpora, the few available corpora have some orthographic and phonetic transcription problems. Some corpora, like OGI-22 and Spoltech, have been criticized [1] with regard to their high level of inconsistency, poor recording quality, erroneous or nonexistent orthographic transcriptions. Other researchers who faced the same problems just decided to create its own corpus [2] or even to change their research by using English language corpora [3].

The goal of this paper is to present the process of building a continuous speech recognition system for BP using the West Point Brazilian Portuguese corpus. The recognition system was built using a well known toolkit, HTK [4], and other open source tools. This way the results can be easily reproducible by any researcher. We show that the default parameters defined for the toolkit, recommended for building speech recognition systems in general, are not acceptable for BP.

The paper is organized as follows. Section 2 presents the concepts used to evaluate continuous speech recognition systems. Section 3 presents the West Point Brazilian corpus. Section 4 and 5 describe the pronunciation dictionary

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 132–141, 2010.

and its phoneme set. Section 6 and 7 describe the language and acoustic models, respectively. Section 8 presents experiments using CETEN-Folha [5], a large vocabulary textual corpora, to build more representative language models. Finally, the paper conclusions and future work are described in Section 9.

2 Continuous Speech Recognition System

The goal of continuous speech recognition systems (CSR) is to identify the sequence of words that are continuously spoken (usually, the words are separated by very short or almost no pauses) by a human. This is one of the main problems that affect the performance of such systems.

A CSR system can be divided into two main components: an acoustic model and a language model. The acoustic model produces the more likely words uttered in a given speech signal. Usually, such component is built using Hidden Markov Models on some language sound unit (e.g., phoneme, triphones). The words are estimated using pronunciation models or dictionaries that allow recognize words from the sequences of the language sound units. The language model produces the more likely sequence of sounds (and words) of a given language from the possible words recognized by the acoustic model. Usually, such models are represented by the probabilistic models, also known as n-grams, estimated on the sequences of words.

The performance of CSR systems are usually expressed in terms of word error rate. This article investigates configurations that yield the lowest error rates for continuous speech recognition system using the West Point Brazilian Portuguese corpus and the large vocabulary CETEN-Folha corpus.

3 The West Point Brazilian Portuguese Speech Corpus

The West Point Brazilian Portuguese Speech [6] (WPBP) corpus is a collection of digital recordings of spoken Brazilian Portuguese, collected to develop acoustic models for speech recognition systems. The data was collected in Brazil, and the speakers were recruited from the Brazilian Military Academy. Some of the speakers in the corpus are non-native speakers. This work only used data from the native speakers to avoid errors caused by the mispronunciation of some words and insertion of foreign language words. Thus, only data from 99 native speakers (53 males and 46 females) were used from the corpus.

The recordings are short sentences, based on prompts read by the speakers. A sentence was presented to the speaker through a visual display. The speaker pressed a key and spoke the sentence. The recording was played back for reviewing the spoken sentence, allowing the utterance to be re-recorded. A member of the data collection team was available during the recording session to verify the recordings and to provide technical assistance in case of malfunctioning equipment.

There are 296 sentences, but every speaker recorded about 80 sentences. In fact, after a manual inspection, it was observed that some of those sentences

were duplicated and others were spoken only by one speaker. Therefore, we selected those sentences that were spoken by more than one speaker. This selection created a subset of the original corpus with 200 sentences.

Despite the apparent well controlled recording process, the corpus has some recording issues. Speech data was collected using 4 different laptop computers. Three laptops recorded speech on a sampling rate of 22050 Hz; the other laptop recorded speech on a sampling rate of 11025 Hz. To avoid sampling rate and quantization problems in the recognition process, the data with the highest sampling rate was re-sampled at 11025 Hz.

4 Pronunciation Dictionary

Instead of using the pronunciation dictionary supplied with the WPBP corpus, this work uses a bigger dictionary called UFPADIC [1]. This dictionary uses a 32 phoneme set.

Since the original WPBP corpus only provided a simple pronunciation dictionary, this research decided to follow the same approach and do not do a careful hand made orthographic to phonetic transcription, in order to take into account the different pronunciations produced by different speakers when reading the same written sentences. We assume that there is a single transcription per word, since the hand made orthographic to phonetic transcription is a complex and very time demanding process.

Since the UFPADIC does not contain all the words in the WPBP corpus, the dictionary was extended using the method described in [7] and the tool Sequitur G2P [8]. On Table 1 and 2, the word (E1) and phoneme (E2) errors obtained with different n-grams sizes are described. Similar tables were published recently in [9], but this paper shows better results on bigger vocabularies.

Table 1. G2P 'Test' (133,555 words and 114,145 phonemes)

model	E1	E1 %	E2	E2 %	Virtual Memory (MB)	Resident Memory (MB)	Time (s)
1-gram	13555	100%	34701	30.40%	98.8	56.5	24
2-gram	1251	9.23%	1420	1.24%	169.4	127.4	164
3-gram	408	3.01%	455	0.40%	209.2	162.5	109
4-gram	457	3.37%	510	0.45%	259.9	198.9	117

E1 = Word Error, E2 = Phoneme Error.

Table 1 shows the recognition error for different models created with the tool Sequitur G2P applied to UFPADIC. In this test, we used 80% of UFPADIC for training and 20% for evaluation. Note that the unigram-based model has low recognition rates (with only some phonemes recognized and no words recognized at all), whereas more complex models provide a better modeling of the BP language. It also shows the computational resources (amount of memory and

Table 2. G2P 'Final' (67,743 words and 570,331 phonemes)

model	E1	E1 %	E2	E2 %	Virtual Memory (MB)	Resident Memory (MB)	Time (s)
1-gram	67740	100%	172700	30.28%	107.3	64.7	60
2-gram	6160	9.09%	7005	1.23%	191.3	149.0	190
3-gram	1553	2.29%	1656	0.29%	185.1	138.0	72
4-gram	1002	1.48%	1075	0.19%	302.1	224.5	96

E1 = Word Error, E2 = Phoneme Error.

processing time) required to create the models, which is proportional to the order of the model and size of the vocabulary. Table 2 shows results using the entire UFPADIC dictionary and WPBP words (the missing pronunciations produced using the models computed on Table 1). This final model has been used to produce automatic word pronunciation for other corpora [10].

5 Phoneme Set

Even though it is not claimed that the WPBP corpus has a balanced representation of the BP language, we compared the phoneme distribution of this corpus against a much bigger corpus: the CETEN-Folha [5]. CETEN-Folha includes text from the Folha de São Paulo newspaper published in 1994 (all the 365 editions). For this work, CETEN-Folha text was normalized using a script by removing sentences with invalid words (e.g. words that contain special characters or numbers). The removal reduced the corpus to 682,722 sentences and 8.58 million words (the original corpus has more than 24 million words). The goal is to verify whether the WPBP contain a phoneme distribution that is representative of the BP language.

Figure 1 shows the frequency, in log scale, of each phoneme for the CETEN-Folha (continuous line) and the WPBP (dotted line). The CETEN-Folha curve is surprisingly similar to the WPBP curve (correlation coefficient equal to 0.9824 with a p-value near zero - 1.198e-13). This comparison shows that the sentences used in the WPBP corpus were carefully built so that they could be considered balanced.

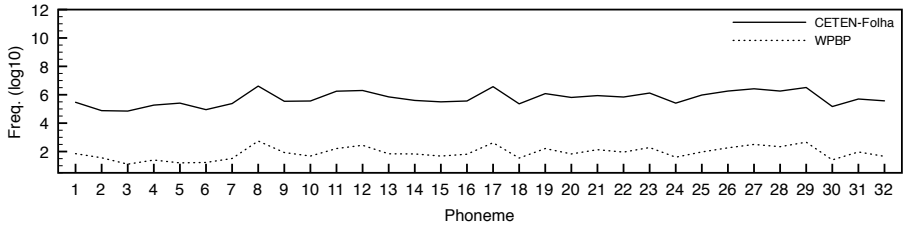

Fig. 1. Number of occurrences (in log scale) of the 32 phonemes in the CETEN-Folha and WPBP corpora

6 Language Model

Several experiments with language models were performed to verify the best combination of discount and interpolation algorithms. As expected, some algorithms, like the Kneser-Ney, did not produce one of the best results because of the small amount of sentences on the WPBP corpus (this algorithm is evaluated using large corpora CETEN-Folha on section 8).

Table 3. WPBP (sentences: 200, words: 1016)

model	1-gram	2-gram	3-gram	4-gram
SRILM	181	30	29	29
SRILM-W	189	9	8	8
SRILM-IW	189	8	8	8

WPBP perplexity measures for different discount and backing-off algorithms.

Table 3 shows the perplexity of the most used language models estimated using WPBP and different configurations of the SRILM toolkit [11]. The SRILM represents the default HTK algorithm for discount and backing-off: the Good-Turing algorithm. The SRILM-W represents an alternative algorithm, valuable for smaller sets of sentences: the Witten-Bell algorithm. Finally, the SRILM-IW represents the same Witten-Bell algorithm with interpolation. Usually, language models with smaller perplexity measures delivers better recognition results since the recognizer will have more hypotheses to evaluate when using language models with higher perplexity.

Since WPBP is a small corpus, the language model generated using only the WPBP sentences is not representative of the Brazilian Portuguese language. However, its use is valuable to verify the effectiveness of the acoustic models created with WPBP audio data. On section 8, a more representative vocabulary is used to create a language model with higher perplexity.

7 Acoustic Model

In search of a reliable and neutral analysis, experiments were performed using a 10-fold cross validation. That is, data from 90% of the speakers are used for training and the remaining 10% are used for evaluation. The sets were created using a balanced proportion of male and female speakers for training and testing.

The acoustic model was estimated using different parameter configurations. As widely used by CSR systems, the feature set was composed by 13 MFCC (Mel-Frequency Cepstral Coefficients) and their respective dynamic features: delta and delta-delta coefficients (39-dimensional feature vector). Training was performed using the HTK Toolkit and evaluation was performed using HVite and AVite, a part of the ATK toolkit [12] that supports 3-grams models. The results obtained are shown in Figures 2, 3 and 4.

Figure 2 shows the word recognition rate for monophones. The worst result is achieved with the default configuration for the HTK and HVite. A better result is obtained by replacing the language model with the Witten-Bell discounting method, yielding a relative improvement of 25% over the default configuration (Student-t Test with df=18 and p-value=0.0). Finally, the best result is obtained using trigrams (3-grams) instead of the default bigram (2-grams). A Student-t Test obtained a p-value 0.0587354 and suggests that state tying delivers significant improvement only at 90% of confidence level. However, it is expected that this observed small improvement would be more evident for acoustic models trained with spoken sentence with higher triphone variability.

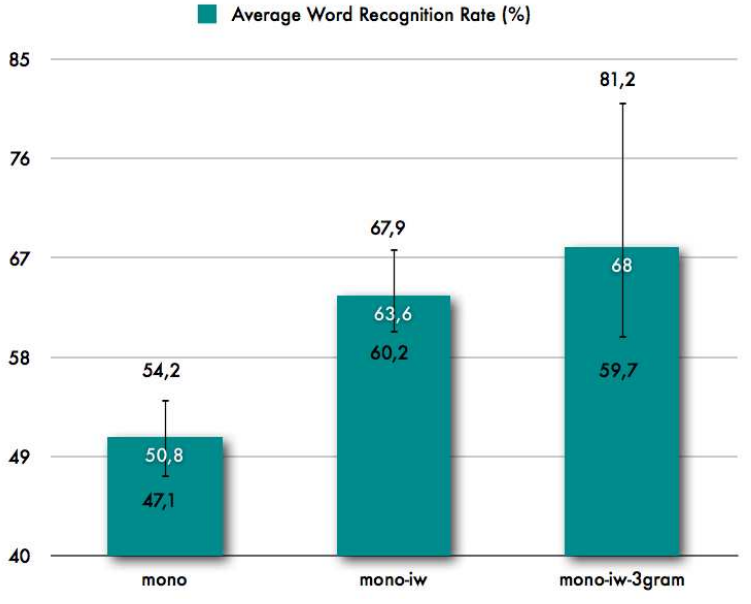

Fig. 2. Recognition rate of systems using monophones with: Good Turing discount (mono), Witten-Bell discount (mono-iw) and Witten-Bell discount and 3-grams (mono-iw-3gram)

Figure 3 used similar notation to represent near state-of-the-art configurations [1,9]: 1) tied state triphones using bigrams, 2) bigrams with a special discount algorithm and 3) trigrams and the same discount algorithm used on 2. Note that previous publications on Brazilian Portuguese recognition system have not used trigrams on speaker independent systems [1].

More improvement can be achieved by using multiple Gaussian mixtures, as shown by the results presented in Figure 4. The best average word recognition rate for the bigram system is 91.2% (3.2% relative improvement over default tied state triphones "tied-iw") and it is achieved by using 6 Gaussian mixtures. The best average word recognition rate obtained for the trigram system is 93.3%

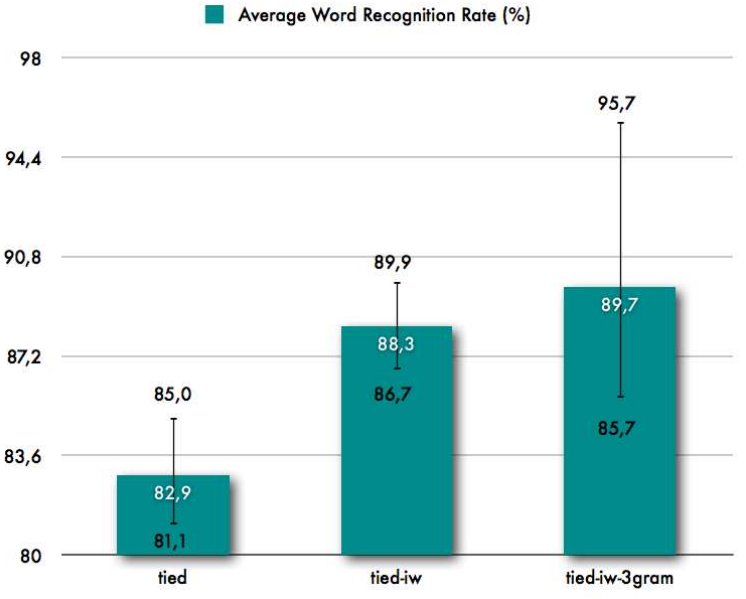

Fig. 3. Recognition rate of systems using Tied State Triphones with: Good Turing discount (tied), Witten-Bell discount (tied-iw) and Witten-Bell discount and 3grams (tied-iw-3gram)

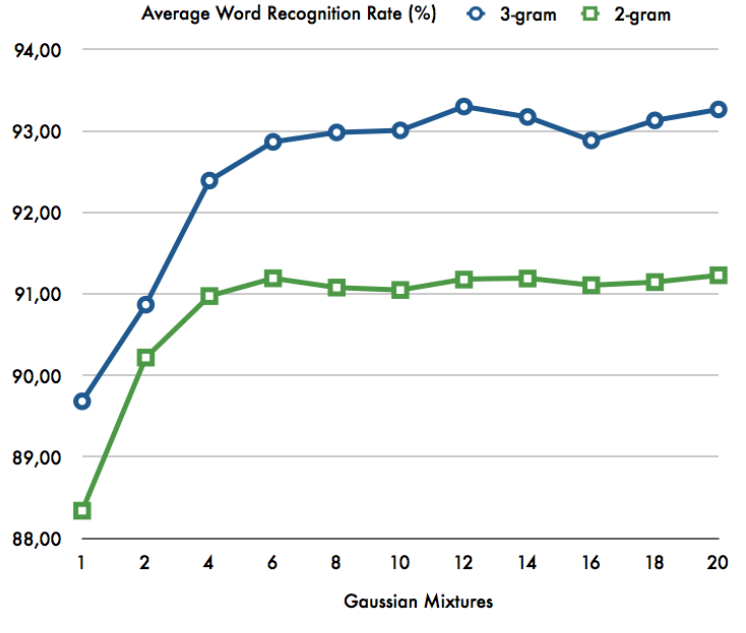

Fig. 4. Average word recognition for bigram and trigram systems for different number of Gaussian mixtures per triphone state

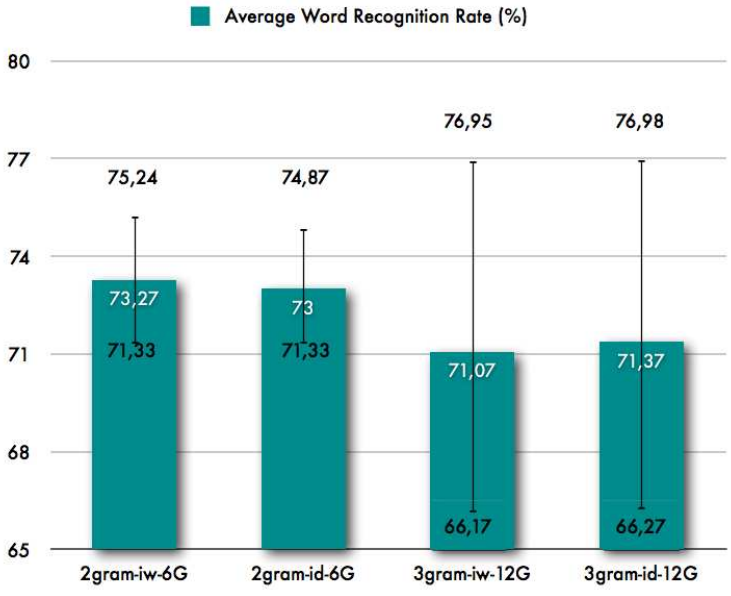

Fig. 5. Average word recognition for bigram and trigram systems for 6 and 12 of Gaussian mixtures using Witten-Bell and Kneser-Ney algoritms

(4.0% relative improvement over 3gram tied state triphones "tied-iw-3gram") and it is achieved by using 12 Gaussian mixtures. Using more Gaussian mixtures does not provide any significant improvement (in fact, it could yield worse result because of the overfitting of the models).

8 Large Vocabulary

Despite the promising results obtained on the experiments using the WPBP corpus, it is important to understand the associated limitations. Since the WPBP corpus has a small amount of triphones when compared to a large corpus like CETEN-Folha, we proceeded with experimental evaluations that use the same acoustic models described in Section 7 and different language models. These language models were created using sentences from the CETEN-Folha corpus, where only sentences that contain the same triphone set trained for WPBP were considered, but now following the probability distributions estimated on a much larger corpus (with 43.9 times more words than WPBP).

Two language models were created with different discounting algorithms. The first model used the same discount algorithm that delivered the best results for the WPBP corpus: the Witten-Bell algorithm. The second model used the Kneser-Ney algorithm, an alternative that usually delivers better recognition rates for large language models. We evaluated the 2 best configurations found for the WPBP on Section 7: 2-gram with 6 gaussian mixtures and 3-gram with 12

gaussian mixtures. Each one was tested with the 2 different options of language model. Figure 5 shows the word recognition rates of the combinations.

As expected, using large vocabularies for training affects the performance by decreasing the recognition rates. However, note that there is not a significant difference between the Witten-Bell and Knesser-Ney versions in our experiments.

9 Conclusions

This paper describes a new baseline system for Brazilian Portuguese continuous speech recognition research and shows its results. The results showed that the default configuration of the HTK toolkit delivers poor results for BP. The improvements observed on 3-gram models showed that better results can be achieved (25% relative improvement) by giving special attention to the language model and using recognition engines that support 3-gram models. The results show that there is still room for improvement, especially on more complete and validated pronunciation dictionaries and on large vocabulary models. Future work could focus on language models based on 4-grams and even larger vocabulary. This work already has started to expand the WPBP corpus by adding more speakers in collaboration with the open source project VoxForge [13]. Sentences, similar to the ones used in the WPBP, have been collected under a free license. All scripts used in this paper are available through the website http://xiru.org/propor2010, expecting that this effort become valuable to other researchers.

References

1. Sampaio Neto, N., Patrick, C., Adami, A.G., Klautau, A.: Spoltech and ogi-22 baseline systems for speech recognition in brazilian portuguese. In: Teixeira, A., de Lima, V.L.S., de Oliveira, L.C., Quaresma, P. (eds.) PROPOR 2008. LNCS (LNAI), vol. 5190, pp. 256–259. Springer, Heidelberg (2008)
2. Teruszkin, R., Junior, F.: Implementation of a Large Vocabulary Continuous Speech Recognition System for Brazilian Portuguese. Journal of Communication and Information Systems 21(3), 204–218 (2006)
3. Neto, N.S., Sousa, E., Macedo, V., Adami, A.G., Klautau, A.: Desenvolvimento de software livre usando reconhecimento e síntese de voz: O estado da arte para o português brasileiro. In: 6 Workshop Software Livre, Anais da Trilha Nacional do Workshop Software Livre, Porto Alegre, vol. 1 (2005)
4. Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. Entropic Cambridge Research Laboratory (1997)
5. Linguateca: Corpus de extractos de textos electrónicos nilc/folha (2008), http://www.linguateca.pt/cetenfolha/
6. Morgan, J., Ackerlind, S., Packer, S.: West Point Brazilian Portuguese Speech (2008), http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008S04
7. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication (2008)
8. Sequitur G2P: Sequitur G2P - A trainable Grapheme-to-Phoneme converter (2008), http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

9. Santos, F., Barone, D., Adami, A.: Validação de Corpus para Reconhecimento de Fala Contínua em Português Brasileiro. In: Proc. V Workshop em Tecnologia da Informação e da Linguagem Humana, TIL 2008 (2008)
10. dos Santos, F.W.: Validação de corpus para reconhecimento de fala contínua em português brasileiro. Master's thesis, Universidade Federal do Rio Grande do Sul (2009)
11. Stolcke, A.: SRILM-an Extensible Language Modeling Toolkit. In: Seventh International Conference on Spoken Language Processing, vol. 2, pp. 901–904. ISCA, Denver (2002)
12. Young, S.: ATK-An Application Toolkit for HTK (2007)
13. VoxForge: Read Prompts and Submit Recordings (2008), http://www.voxforge.org/pt_br/read

Voice Quality of European Portuguese Emotional Speech

Ana Nunes[1], Rosa Lídia Coimbra[2], and António Teixeira[3]

[1] Universidade de Aveiro, Portugal
[2] Dep. Línguas e Culturas, Universidade de Aveiro, Portugal
[3] Dep. Electrónica, Telec. & Informática/IEETA, Univ. Aveiro, Portugal

Abstract. In this paper we investigate parameters related to voice quality in European Portuguese (EP) emotional speech. Our main objectives were to obtain, to our knowledge for the first time, values for the parameters commonly contemplated in acoustic analyses of emotional speech and investigate if there is any difference for EP relative to the results obtained for other languages. A small corpus contemplating five emotions (joy, sadness, despair, fear, cold anger) and neutral speech produced by a professional actor was used. Parameters investigated include fundamental frequency, jitter, shimmer and Harmonic Noise Ratio. In general, results were in accordance with the consulted literature regarding F0 and HNR. For jitter and shimmer our results were, in certain aspects, similar to the ones reported in a study of emotional speech for Spanish, another Latin language. From our analyses, and taking into consideration the reduced size of our corpus and the use of an actor as informant, no clear EP characteristic emerged, except for a possible, needing confirmation, difference regarding joy, with values similar to neutral speech.

1 Introduction

Emotions influence physiological state, with important effects on speech production and especially on the phonation process. These effects are reflected in varied and complex voice quality related parameters, such as fundamental frequency (F_0) and jitter. Some parameters are identical in several (or all) languages; others are part of the specificities of a language or speaker.

"Increased emotional arousal is accompanied by greater laryngeal tension and increased subglottal pressure which increases a speaker's vocal intensity". For example, Darwin observed that angry utterances sound harsh and unpleasant because they are meant to strike terror into an enemy [1].

Anger is usually associated with an increase in mean F_0 and energy. Anger also includes "increases in high frequency energy and downward-directed F_0 contours. The rate of articulation usually increases" [2].

The increase of F_0 mean and range is also a characteristic of fear, also with high frequency energy; sadness shows a decrease in mean F_0, F_0 range and mean energy; joy, a positive emotion (one of the few that are usually studied), has an increase in mean F_0, F_0 range, F_0 variability, mean energy, and an increase in high frequency energy [2].

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 142–151, 2010.

"Understanding a vocal emotional message requires the analysis and integration of a variety of acoustic cues" [3].

1.1 Voice Quality and Emotion

Johnstone & Scherer (1999) [4] present studies in which emotional vocal recordings were made using a computer emotion induction task and an imagination technique. Voice quality acoustic parameters included F_0 minimum, F_0 range, jitter and spectral energy distribution. The emotions studied were: tense, neutral, irritated, happy, depressed, bored and anxious.

The authors report that: "values for jitter are correlated with F_0 floor, thus indicating that period to period F_0 variation tends to be larger with higher F_0. This tendency is absent for anxious and tense speech though, which is in agreement with previous findings of a reduction of jitter for speakers under stress". Happy speech presents significantly higher values of jitter than all other emotions. Also as expected, F_0 floor was found to be lowest for the emotions bored and depressed, and highest for happy and anxious speech.

Zovato et al. (2004) [5], used three basic simulated emotional styles (besides neutral, they had anger; happiness and sadness). An Italian female professional speaker recorded 25 sentences; the main point was to investigate the correlation between emotions and acoustic parameters (F_0 minimum, maximum, mean and range, plus RMS energy). They also applied a perceptual test to 10 volunteers to evaluate the corpus. It was shown that there was some confusion in discerning the pairs neutral/sad and happy/angry.

Recently, Toivanen et al. [6] investigated how well voice quality conveys emotional information to be perceived by humans and computers. They used nine professional actors to produce their data, simulating: neutral; sadness; joy; anger and tenderness states, in which they extract a vowel from the entire running speech (approximately one minute). They analyzed vowel [a]. The samples were presented to 50 listeners to recognize the emotion and classified using automatic methods. Human listeners were better than the machine at recognizing anger.

Drioli et al. [7] analyzed F_0, duration, intensity, jitter, shimmer, HNR and other voice quality indexes such as Hammarberg Index. The authors utilized Praat voice report. Regarding irregularities, and for stressed vowels, they report a high shimmer value for anger; higher jitter values for joy and surprise (with anger in third place). The HNR is lower for anger and joy.

Chung, in 2000, investigated acoustical properties of Korean emotional speech. The author measured: F_0 parameters (mean, maximum, minimum, mean of the 20% lowest values, range), jitter, shimmer, speaking rate and spectral distribution. The analysis showed that joy increases F_0 mean, whereas sadness enhances the decrease of F_0 minimum. The increase of F_0 maximum and of F_0 range was found to be "a good indicator of the general emotional arousal". "The jitter and the shimmer values seem to increase under the emotional tension (...). However, these variations (...) were not statistically significant in the case of Korean data" [8].

Voice quality aspects are very often described qualitatively. In quantitative investigations, the most studied parameters relate to F_0. More recently, the list of investigated parameters expanded to include jitter, shimmer, HNR, glottal source parameters, etc.

1.2 Crosslinguistic Information on Emotional Speech

Probably beginning with Darwin [1], it is known that facial expressions are more universal than prosody, even though studies that only contemplate prosody or non-verbal aspects revealed that anger is reasonably well perceived, but the same does not occur with joy. It is also well known that speakers are commonly better on perceiving emotions in their own language.

Many emotion theorists defend that emotions are mostly learned and affected by social environment. As a result, emotions are conjectured to vary considerably across cultures.

Scientific studies have been made crossing speakers and listeners of several origins. However, according to [9], the languages and cultures studied so far are not actually very diverse. Moreover, only a few specific emotions have been studied systematically, usually basic emotions [10, p. 244].

1.3 Related Work for European Portuguese

Not much research was conducted on the subject of emotional speech for European Portuguese. There is no corpus, big or small, of EP emotional speech available. Even the recent work on emotional speech synthesis of Portuguese [11,12] was based on information published for other languages, complemented by extraction of glottal parameters (such as open quotient) from a German database.

The corpus used in the present work was previously partially used in a recent Master thesis work [13]. The parameters investigated were related to F_0 and articulation. Parameters such as jitter were not contemplated.

1.4 Objectives and Paper Structure

Our main objectives were: to obtain values for the parameters commonly contemplated in acoustic analyses of emotional speech; to compare the obtained parameter values with the ones reported for other languages, in order to determine which follow general tendencies and which are characteristic of the EP language.

In the next section, corpus and information on the extraction of voice quality related parameters are presented. Section 3 presents results for the different parameters contemplated in our study, ending with the joint analysis of four of them. The paper ends with discussion, main conclusions and suggestions for future work.

2 Material and Method

The corpus that was recorded and analyzed is composed of two sentences - one simple, the other more complex - both extracted from the Portuguese version of the naturalistic dialogue 'The human voice' by the French writer Jean Cocteau. The simple was "O melhor será tomares conta deles" (You've better take care of them) and the complex "Não tenho com certeza a voz de uma pessoa que esconde qualquer coisa" (I don't really have the voice of a person who hides something). The chosen sentences do not present by themselves any emotional charge, so that the actor may interpret them according to the intended values: joy, despair, anger, fear, sadness and the neutral form.

The informant is a professional actor, male, 42 years old, with a wide experience in national television and theatre (actor of Portuguese National Theater Company). He has also performed several cartoon voices and has been the voice of well-known advertisements.

Recording sessions took place in a soundproof booth at the Municipal Theatre of Guarda (North Interior of Portugal), all in the same day, with a gap of one hour between the two recording sessions. A microphone AKG C 451 B and a DAT recorder Tascam DA-P1 were used, and an experienced sound technician was present. Between each utterance there was no more time than the needed for respiratory pauses. We consider that these circumstances allowed the production to get closer to spontaneous emotion, since the actor did not have much time to concentrate, and thus the register became less acted.

Annotation and Feature Extraction - The utterances were first annotated at word and phone levels, using SAMPA transcription in SFS (Speech Filing System). The limits of each segment were marked and a broad phonetic transcription was made, considering phenomena such as elision, sandhi and epenthesis. All data were processed in Praat software [14] which allowed the extraction of the needed elements using Praat Voice Report function. Statistical analyses were made in SPSS (v. 16) and R. As part of our parameters departs from a Normal distribution, non-parametric tests were employed.

3 Results

The following subsections present the results of our analyses regarding the most commonly studied parameters relevant to voice quality: F_0, jitter, shimmer, and harmonic noise ratio, complemented with autocorrelation.

3.1 F_0 Parameters

Four different F_0 related parameters were investigated. The results for F_0 minimum, F_0 max, F_0 mean and F_0 standard deviations are presented in Fig. 1 as function of emotion.

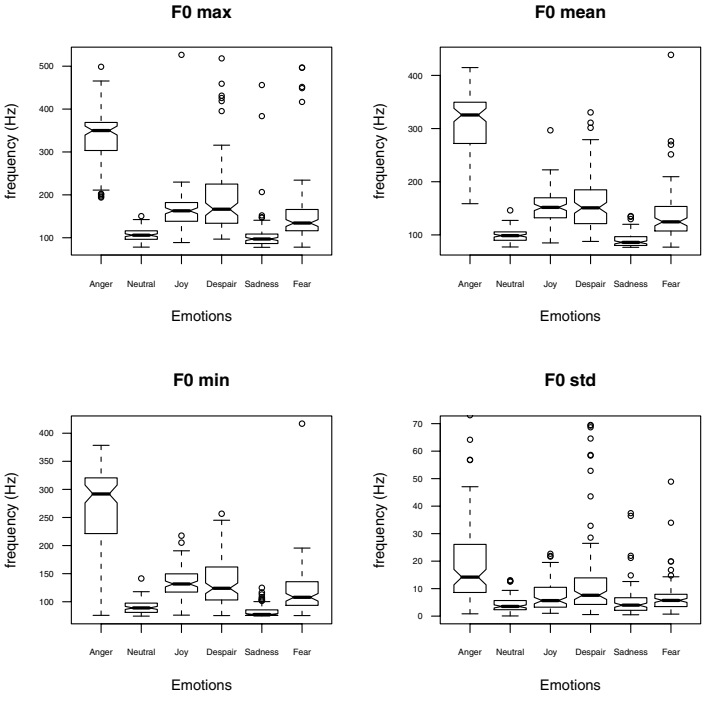

Fig. 1. Effect of emotions on four fundamental frequency (F_0) related parameters. From top left: maximum, minimum, mean and standard deviation.

The analysis of different F_0 parameters shows that anger is clearly differentiated, presenting an average value near 300 Hz and the highest standard deviation and range.

Joy and despair present similar values on the four F_0 parameters, with mean around 150 for F_0 mean and F_0 max. One difference between the two is the higher range of values for despair.

Fear has values of F_0 a little lower than the previous pair. Standard deviation is also lower. Sadness presents the lower values for the parameters, some similar to the neutral.

The Kruskal-Wallis (KW) test (p=0.01 corrected for multiple comparisons) confirms as significant the factor emotion for all four parameters: $[\chi^2(5) = 338.65, p < 0.001]$ for minimum; $[\chi^2(5) = 370.72, p < 0.001]$ for maximum, $[\chi^2(5) = 408.46, p < 0.001]$ for mean; and $[\chi^2(5) = 140.11, p < 0.001]$ for standard deviation.

For F_0 maximum, minimum and mean, post-hoc tests showed as significantly different all pairs except despair-fear, despair-joy, fear-joy, and neutral-sadness. For F_0 standard deviation, also the following pairs were not significantly different: fear-neutral, fear-sadness and joy-sadness. At least some pairs are difficult to differentiate based on F_0 parameters. The standard deviation presents the lowest discrimination power, the other three show similar power.

3.2 Irregularities: Jitter and Shimmer

To compensate for the intonation related variations of F_0, as we used speech from sentences, we only contemplated PPQ5 parameter for jitter. Influence of emotion in PPQ5 is presented in Fig. 2.

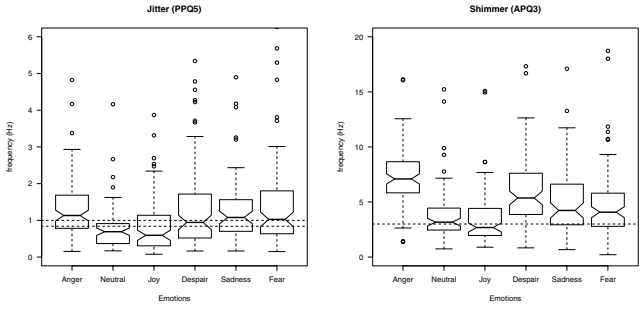

Fig. 2. Effect of emotions on jitter PPQ5 and shimmer APQ3. Horizontal lines represent two thresholds usually associated with normality.

Comparing with F_0 related parameters, differences in jitter are not so evident. Our results showed that higher jitter values are associated with despair, fear, anger and sadness, more negative emotions. The neutral speech and joy present lower or similar values.

Only joy presents smaller PPQ5 than the normality threshold indicated in MDVP . The other emotions don't present values significantly lower or higher than both thresholds.

The KW test confirms as significant the factor emotion $[\chi^2(5) = 53.9012, p < 0.001]$. Post-hoc multiple comparisons test after KW showed as different: joy-anger, joy-fear, joy-sadness and neutral-sadness. Concerning jitter values, joy appears clearly lower than three of the other emotions. Jitter seems a relevant factor to detect joy.

Analysing shimmer parameters, in Fig. 2 at the right, it was clear that they are particularly high for anger, followed by the group integrating despair, sadness and fear.

Looking at the threshold of normality, only neutral and joy are not significantly above. The shimmer values for anger and despair are clearly in the region usually considered as pathologic. The KW test confirms as significant the factor emotion $[\chi^2(5) = 123.99, p < 0.001]$. Multiple comparisons test after KW shows as significant the following differences: anger-fear, anger-joy, anger-neutral, despair-joy and despair-neutral. That is, anger only does not present significantly higher shimmer values than sadness and despair, related emotions; despair also has significantly higher shimmer values than joy and neutral; other emotions present no significant differences. Shimmer only differentiated anger and despair from all the remaining.

3.3 Harmonic Noise Ratio (HNR)

HNR values are presented in Fig. 3. A horizontal line represents a common threshold, 12 dB. While joy presents a value higher than 12 dB, most of the emotions present values around that value and some, like anger and despair, values significantly smaller.

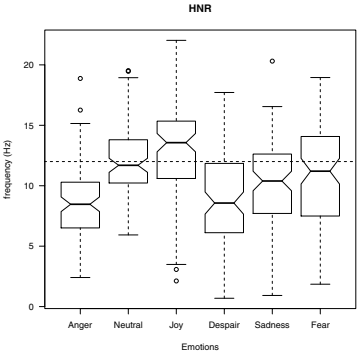

Fig. 3. Effect of emotions on Harmonic Noise Ration (HNR). Horizontal line represents a common normality threshold.

The KW non-parametric ANOVA confirms as significant the factor emotion $[\chi^2(5) = 84.23, p < 0.001]$. Multiple comparisons test after KW shows as significant the following differences: anger-fear, anger-joy, anger-neutral, despair-joy, despair-neutral and joy-sadness. The situation is similar to the one reported for shimmer, with the addition of a new significant difference between joy and sadness. In HNR besides the salient differences of anger and despair (now the lowest values) we also have differences for the positive emotion joy, with significantly higher values of HNR than all others. The values of HNR for anger and despair are in a region potentially classifiable as pathologic.

3.4 Parameters Combination

Inspired by multidimensional representations of voice parameters (ex: Kay Elemetrics MDVP and Hoarsness Diagram) in Fig. 4 we combine F_0 (only mean, as no significant differences were obtained by applying the other three parameters), jitter, shimmer and HNR in a plot. For normalization, values of each parameter were dividided by the mean.

Based on the figure and taking in consideration the results from statistical tests presented before, we have: (1) anger differing from neutral by all four parameters. Jitter, shimmer and F_0 are increased, HNR decreases; (2) despair also differing by the four parameters, but with smaller differences than for anger; (3) sadness is essentially only different by irregularities (Jit & Shim); (4) fear similar to despair, with HNR closest to neutral values and smaller differences; (5) joy, a positive emotion, with only distinct F_0 values.

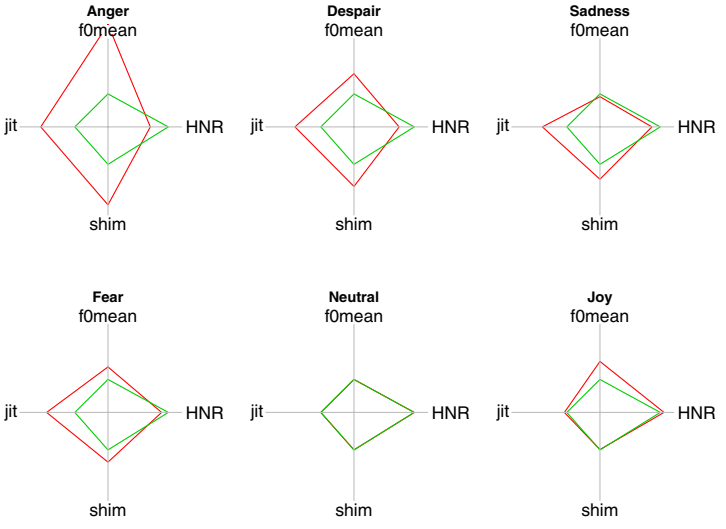

Fig. 4. Comparing the five emotions and neutral speech based on the four types of parameters used. For F_0, mean was chosen as representative of the four F_0 parameters included in the study.

4 Discussion

Parameters analyzed included several F_0 related measures, jitter, shimmer and HNR. Five emotions plus neutral were contemplated. As the languages and cultures studied so far are restricted, this research presents a small contribution to increasing the diversity in quantitative data regarding the complex relation between voice quality and emotions. In general, results were in accordance with the consulted literature. This is particularly true for F_0 related parameters and emotions such as anger and despair. Our results for joy disagree, at least in some parameters, from some previously reported results.

F_0 maximum and average differentiates anger, sadness and joy as reported in [2] and [15]. Anger presents the highest F_0, joy an equally high value, sadness the lowest value. Sadness, as reported in [15], has values close to neutral. Our measures don't confirm the increase of F_0 for fear. As Scherer (cited by [16]) suggested, our F_0 measures correlate with activation dimension; high activation relates with higher values of F_0.

The comparison of jitter and shimmer values with the literature is more difficult. Firstly, there are fewer studies reporting such parameters; secondly, there is some uncertainty on the exact parameter report (ex: local, PPQ5); thirdly, the process of parameter extraction is not necessarily equivalent. Our option was to compare essentially with [7], also using Praat in their analyzes.

Regarding shimmer, our results, showing that only anger does not present significantly higher shimmer values than sadness and despair, are in agreement with published measures such as [7] by Drioli et al. (2003) showing high shimmer values for anger.

Higher jitter values for joy and surprise (with anger in the third place) were reported [7]. [4] reported higher jitter for happy. In our results, joy appears clearly lower in jitter than three of the other emotions. Jitter seems a relevant factor to detect joy, but values are to the lower side, contrary to the reported higher values of the mentioned study. The similarity of jitter for joy and neutral observed for our EP data is in agreement with the results obtained by Monzo et al. (2007) [17] for Spanish happy and neutral.

In agreement with [7], HNR is lower for anger and for despair, also negative. Contrary to same work, we verified significantly higher values of HNR for joy, placing this emotion very far from fear relative to HNR.

The differences observed for joy can be related to the difficulty in identifying the emotion in perceptual tests conducted with the same material [18], in agreement with the observations of Darwin, presenting joy as more difficult to transmit by voice alone. Joy was often confused with neutral, by native and non-native EP speakers. The corpus is too reduced to generalize, but this question of possible differences in joy expression or relative unsuccessful of the actor in expressing this emotion, recommend follow-up studies.

For some of the emotions, parameters such as HNR and jitter present values in the "pathological" ranges usually considered in voice evaluation. This points to the necessity of controlling the emotional state of the subject to whom a voice evaluation procedure is applied. Being sad or happy has, as demonstrated by our results, very noticeable effects on the "normal" range of several parameters.

The main shortcoming of the study is the corpus size and the use of an actor. Extensions of the work regarding the number of subjects, the extent of speech material and more natural emotional speech are needed.

The parameters analyzed are the most common, providing a good starting picture of the effects of emotions on the voice quality, but do not cover all possibilities. The glottal source parameters, such as Open Quotient, and the spectral parameters should be added.

5 Conclusion

For the first time was investigated several voice quality related parameters for five different emotions in European Portuguese: sadness; happiness; fear; anger and despair. Analyses focused on F_0 related parameters; shimmer; jitter and Harmonic Noise Ratio (HNR).

From our analyses, and taking into consideration the reduced size of our corpus and the use of an actor as informant, no clear EP characteristic emerged. Nevertheless, differences on several parameters were observed for joy that, if confirmed, would constitute a cultural difference.

References

1. Darwin, C.: The Expression of Emotions in Man and Animals. Portuguese translation by Relógio D' Água (2000) (1872)
2. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70(3), 614–636 (1996)
3. Schirmer, A., Kotz, S.A.: Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. TRENDS in Cognitive Sciences 10(1) (2006)
4. Johnstone, T., Scherer, K.R.: The effects of emotion on voice quality. In: International Congress on Phonetic Sciences (ICPhS), San Francisco (1999)
5. Zovato, E., Pacchiotti, A., Quazza, S., Sandri, S.: Towards emotional speech synthesis: A rule based approach. In: ISCA SSW (2004)
6. Toivanen, J., Waaramaa, T., Alku, P., Laukkanen, A.M., Seppänen, T., Väyrynen, E., Airas, M.: Emotions in [a]: a perceptual and acoustic study. Logoped Phoniatr Vocol 31(1), 43–48 (2006)
7. Drioli, C., Tisato, G., Cosi, P., Tesser, F.: Emotions and voice quality: Experiments with sinusoidal modeling. In: VOQUAL (2003)
8. Chung, S.J.: Expression and Perception of emotion extracted from the Spontaneous Speech in Korean and in English. PhD thesis, Sorbonne Nouvelle University (2000)
9. Zinken, J., Knoll, M.A., Panksepp, J.: Universality and diversity in the vocalisation of emotion. In: Isdebski, K. (ed.) Emotions of the human voice. San Diego Plural Publishing (in press)
10. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)
11. Cabral, J.: Transforming Prosody and Voice Quality to Generate Emotions in Speech. Dissertação de mestrado, IST/UTL (2006)
12. Cabral, J., Oliveira, L.C.: EmoVoice: a system to generate emotions in speech. In: InterSpeech, pp. 1798–1801 (2006)
13. Rodrigues, A.: As Emoções na Fala (Emotions in Speech). Masters dissertation, Universidade de Aveiro, PT (2007)
14. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
15. Cowie, E.D., Cowie, R., Schroeder, M.: The description of naturally occurring emotional speech. In: ICPhS, pp. 2877–2880 (2003)
16. Airas, M., Alku, P.: Emotions in short vowel segments: Effects of the glottal flow as reflected by the normalized amplitude quotient. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 13–24. Springer, Heidelberg (2004)
17. Monzo, C., Alías, F., Ignasi, I., Gonzalvo, X., Planet, S.: Discriminating expressive speech styles by voice quality parameterization. In: XVIth International Conference on Phonetic Sciences (ICPhS), pp. 2081–2084 (2007)
18. Nunes, A.M.B., Roussel, N., Rodrigues, A., Coimbra, R.L., Teixeira, A.: Cross-linguistic effects on the perception of emotions. In: ICPLA (2008)

Prosodic Prediction in Brazilian Portuguese: A Contribution to Speech Synthesis

Cirineu Cecote Stein

Universidade Federal da Paraíba – CCHLA/DLCV. Conjunto Humanístico – Bloco IV – Campus I 58059-900 João Pessoa-PB, Brazil
cirineu.stein@cchla.ufpb.br

Abstract. The prosodic prediction phenomenon concerns the possibility of a listener to predict what will be said, following the melodic modulation used for the initial part of a full sentence. This prediction concerns the semantic value and/or the syntactic relations in the sentence. The existence of that phenomenon was confirmed for Brazilian Portuguese, considering the relations between main clauses and subordinate adverbial clauses. Using the speech resynthesis process, it was possible to identify the specific fundamental frequency and durational behavior responsible for the predictive value in the main clauses. The implementation of an algorithm which considers that predictive phenomenon in a speech synthesis system will produce more natural final auditory results.

Keywords: Prosodic prediction; syntactic relations; main and subordinate clauses; Brazilian Portuguese.

1 Introduction

The prosodic prediction theory used in this research refers to Fónagy (1974), who considered that the intonation is able to point out, besides the outline of the semantic and syntactic units, the sentence duration or its quality. It would be possible, using the intonation, to notice whether the sentences are coordinate or subordinate, or if it is the case of an intercalation or an enumeration, among other possibilities.

This study focuses specifically on the characterization of the prosodic prediction (also known as pre-indicative function of the intonation) in subordinate adverbial clauses in the Rio de Janeiro variety of Brazilian Portuguese (BP). Assuming the canonical order "main clause (MC) – (connective) subordinate adverbial clause (SAdvC)", it would be possible for the listener to predict the semantic characterization to be used in the SAdvC, considering the melody the speaker uses in the MC. Some work has been developed in the area, such as Blin & Edgnington (2000), considering the choice of prosody based on a tree-like structure sentence corpus, or Marsi (2004), who considers pitch accents and intonational boundaries prediction, but nothing, as far as we know, related to the specific topic of this research.

In this work, the evidences for the predictive phenomenon are presented, and also the two main pattern configurations for F0: ascending and descending. After the confirmation of that phenomenon in BP (Test 1), two MC's, in their neuter and

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 152–161, 2010.

monotonous versions, were resynthesized in a way to reproduce the original configuration observed in the MC's used in Test 1. Those two MC's preceded SAdvC's of cause, consequence, finality, concession, conformity, and temporality. That proceeding aimed at isolating and identifying the prosodic components responsible for the predictive value of each of those adverbial subcategories. After the selection of two critical ressyntheses for each subcategory, a perception test (Test 2) identified one of them as a better representative of those predictive values.

With the identification of the predictive prosodic contours' configuration for adverbial sentences, it is expected that their application to the speech synthesis systems allows a final auditory result with an even higher quality, compared to the one verified lately.

2 Methodology

This research's first hypothesis was focused on the prosodic predictive phenomenon in BP itself. To confirm it, two MC's were used: MC1 *Ficava infeliz* (fi˛kavĩfeˈliʃ) ("he/she got sad"), and MC2 *Mostrava-se cansado* (moʃ˛travasikãˈsadʊ) ("he/she seemed tired"). The second hypothesis considered that the auditory perception of the predictive contours can be lessened if the F0 main movement is not followed by a pause (House, 1995). In that sense, MC2 presents a post-stressed syllable, which prevents the stressed syllable from being immediately followed by a pause (see Test 1 discussion for details).

In order to confirm the predictive phenomenon, the two MC's were put together with SAdvC's, producing complete sentences. Appropriate melodic configurations should be applied to them. The SAdvC's referred to the nine adverbial subcategories established by BP traditional grammar (Bechara, 1992): cause, comparison, concession, condition, conformity, consequence, finality, proportion, and temporality. To avoid ambiguous values caused by sentence reduction, all of the SAdvC's were introduced by connectives, which made the semantic relations established between the MC's and them, explicit. For example, *Ficava infeliz porque o trabalho não produzia frutos* ("he/she got sad, because the job did not work out").

The 18 periods (2 MC's X 9 SAdvC's) were recorded three times by a Brazilian female speaker born in Rio de Janeiro, experimented in phonetics, and in the stylized use of the voice. All of the 18 periods were previously inserted in short texts, which evidenced the semantic values at goal for the adverbial relations represented by the connectives. The speaker was free to use the melody she judged more appropriate in each case.

In order to select which of the three tokens for each period presented the melodic modulation more appropriate to the semantic relations established between the SAdvC's and the MC's, the acoustic signals of the periods were isolated from their contexts, and presented to four Brazilian speakers, all of them experimented in phonetics. They used a judgment scale from 0 to 4, where "0" means "not appropriate" and "4", "totally appropriate". The tokens with the best scores were considered as a reference (MC's of reference) for Test 1 design (prosodic prediction phenomenon confirmation), and for the resyntheses to be used in Test 2.

Once the existence of predictive melodic contours for six adverbial sentence subcategories in BP was confirmed (see details in section 3), it was necessary to identify which specific configuration of the prosodic components should be responsible for the identity of each of those contours. The voice resynthesis process was considered the most appropriate for that case.

The identification of melodic contours implies their applicability to the voice of any speaker of the language. In that sense, considering that the MC's tokens were produced by a woman, the option, considering the resynthesis process, was to use a male voice. As a first step, the two MC's (separated from their SAdvC's) were recorded three times by a Brazilian male speaker, experimented in phonetics and in the stylized use of the voice. He used a neuter intonation, here understood as an intonation which did not reflect neither emotional and/or attitudinal values, nor adequateness to one of the adverbial values used in the research. In the same sense as for Test 1, four Brazilians, also experimented in phonetics, judged the neutrality level of those MC's intonations, using a scale from 0 to 4.

Assumed a neuter version for MC1 and MC2, it was decided to use also a monotonous version of those MC's, in a way to isolate any possible interference of the prosodic characteristics possibly present in the neutral version. That monotonous version was produced by means of the pitch points elimination along the fundamental frequency, at 132Hz for MC1, and at 127Hz for MC2. To resynthesize those four versions (MC's 1 and 2 neuter; MC's 1 and 2 monotonous), it was used the acoustic analysis program Praat, developed by Paul Boersma and David Weenink.

MC's 1 and 2 resynthesized versions were used in a second perception test (Test 2), making it possible to identify the specific characteristics of the prosodic configuration responsible for the predictive value of each of the six subcategories identified in Test 1.

2.1 Test 1 Layout

After the identification of the best tokens, the sound segments representing the MC's were isolated from the sound segments representing the SAdvC's. The first ones were used in Test 2, constituting independent sound files. They were crossed by binary combinatory analysis with all of the nine SAdvC's subcategories, producing a total of 72 crossings [MC1 X 9 predictions = 9 sound files X (9-1) adverbial subcategories = 72 combinations : 2 (binary crossing) = 36 crossings X 2 MC's = 72 total crossings].

In Test 1, MC's and SAdvC's were diagramed as shown in figure 1. The participants (40 undergraduate students in Letters, in the Universidade Federal do Rio de Janeiro) listened, for each item, to only the sound files concerning the MC's, in pairs (that is, for each item, two of the MC's of reference sound files were listened to, with melodies corresponding to the SAdvC's in the item). Simultaneously, they should read, in the last column, the SAdvC's. They were asked to choose which of the two SAdvC's was the appropriate complement for the melodies listened for each of the MC's. This procedure produced the crossing. In that case, consequently, it was produced a forced choice test: once an association was chosen, the other would be automatically established. In all of the crossings, each of the modulations of the MC's

pair corresponded necessarily to one of the sentences in the SAdvC's pair. In this sense, each listener was exposed 9 times to the same MC predictive melody. Test 1 was applied during two sessions in a same week.

1	A	Ficava infeliz	A B	1	porque o trabalho não produzia frutos.
	B	Ficava infeliz		2	como uma pessoa que não via resultados.
2	A	Mostrava-se cansado	A B	1	embora o trabalho o motivasse.
	B	Mostrava-se cansado		2	se os exercícios não fossem adequados.

Fig. 1. Test 1 lay-out. The participants listened to only the MC's in the third column, which should be connected to the SAdvC's in the last column. The crossings were indicated in the fourth column; for example, A2 B1.

2.2 The Resynthesis Process

The resynthesis process goal was to adapt the neuter and monotonous versions of the MC's to the melodies observed in MC's 1 of reference, by means of the durational and F0 values manipulation, isolated or joined. Once Test 1 confirmed the hypothesis that a post-stressed syllable interferes with the melodic movement's perception, it was decided to assume the prosodic configuration for MC's 1 of reference as well for MC's 2 manipulation, even if some details of MC's 2 configuration were slightly different from MC's 1's. For each subcategory involved, manipulations for both durational values and F0, in the segments considered critical, followed a gradual scale of percentage, under and over the values observed in MC's of reference. In that sense, F0 rising was established in 10%, in 20%, and so on. The same procedure was used for durational values. Whenever possible, the same steps were used for both MC's 1 and MC's 2, considering a same adverbial prediction subcategory (that is, the same variation for F0 and duration were used for MC's 1 and 2 of cause, for example). Intercategorially, on the other hand, it was not possible to keep that same uniformity, due to the peculiar characteristics of each subcategory.

Considering the three subcategories in which F0 presents an ascending movement from the final stressed syllable (FSS) on (cf. figure 4), it was necessary to establish two punctual referents for F0 manipulation. First, a base line, parallel to the imaginary ax of duration (absciss), touching F0 at the first moment of its visual manifestation on the final pre-stressed vowel (FPreSV) (in the monotonous MC's, the base line is the F0 line itself, established at 132Hz, for MC1, and at 127Hz, for MC2). Second, a reference value (RV), in Hertz, based on the descending variation of F0, established between the vowel attack in the final pre-stressed syllable (FPSS) and the minimum pitch value observed at the descending movement. Assuming the F0 valley observed in MC1 causal of reference as a referential, 1 RV corresponds to a variation of 20%, in Hertz. Both the base line and the RV are illustrated in figure 2.

From the total resyntheses produced, two were chosen – *a* and *b*, considering if F0 contour itself would be enough to confer a certain predictive value to the MC in question, or if it would be necessary to combine it with the durational changes. Those resyntheses *a* and *b*, selected for the six subcategories chosen in Test 1, were used in Test 2.

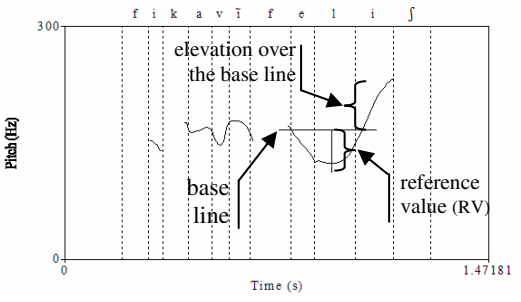

Fig. 2. Layout of the reference value (RV), of the base line, and of the F0 elevation over the base line, in the MC1 consecutive of reference. In this example, 2RV's.

2.3 Test 2 Layout

Each of the resyntheses *a* and *b*, produced for the four MC's, were crossed by simple combinatory analysis with the six SAdvC subcategories, two by two, producing 120 crossings for neutral MC's and 120 crossing for monotonous MC's [2 resyntheses X 6 predictive subcategories = 12 resyntheses X (6 – 1) adverbial subcategories = 60 crossings X 2 MC's = 120 total crossings].

The 40 participants of Test 2 (20 undergraduate students in Letters from Universidade Federal do Rio de Janeiro, and 20 from Universidade Federal do Espírito Santo) were divided in two groups: 20 (10 of each university) took part on the test directed to the neutral MC's resyntheses; 20 (10 of each university), to the monotonous MC's resyntheses. The participants listened to each of the sound files relative to each of the resyntheses four times, with an intervening time of 2300msec between each token. They were asked to choose which of the two SAdvC's was the best complement for the MC, according to the listened melody. Test 2 layout is illustrated in figure 3.

1	Ficava infeliz	A ( )	porque o trabalho não produzia frutos.
		B ( )	que dava pena ver seu rosto.
2	Mostrava-se cansado	A ( )	para que todos sentissem pena dele.
		B ( )	conforme convinha a um pessimista.

Fig. 3. Test 2 layout. The participants listened only to the MC's in the second column, which should be complemented by the SAdvC's in the last column. The associations should be indicated in the third column, with an X.

3 Results

It is possible that more than one melodic contour be used in a same communicative situation. It is expected to be possible to identify at least one contour, which can be recognized by the users of a language, as applicable to a specific context. The following results were evidenced within a group of undergraduate students of Letters of a same BP variety, used to the linguistic nuances of sentences and to a same kind of melody. Even though, the initial expectation was that there would not be a unanimous choice of the predictive contours. Because of that, it was established that the modulation used for the

MC's would be considered predictive of an adverbial subcategory if, among the eight possible crossings for that subcategory, the modulation would be significantly identified in five, at least (half plus one). A Chi-square test was used, with a significance level of 95%.

3.1 Test 1

Considering MC1 isolately, Test 1 evidenced four, among the nine total prediction possibilities, as significantly recognized: cause, concession, consequence, and temporality. Considering MC2, there was only one significant identification: concession (which confirms hypothesis 2).

Nevertheless, if the results for MC's 1 and 2 are superposed, six prediction values are identified: cause, concession, consequence, temporality, conformity, and finality.

3.2 Test 2

Test 2 showed a tendency to choose one of the resyntheses *a* or *b* as presenting the prosodic behavior responsible for the prediction in each of the six adverbial subcategories. But it was not possible to identify one of them as exclusive. Following, the characteristics of resyntheses *a* and *b* for MC1 monotonous (the same for MC1 neuter and MC's 2 neuter and monotonous) are presented, as well as the results for Test 2.

3.3 Causal Prediction

Resynthesis *a* was characterized by:

. F0 valley formation, with 1RV;
. centralization of F0 valley in the attack consonant (C) of FSS;
. F0 raising, 5RV's over the base line.

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. raising of F0 previous to FPSS 3% over the base line.

The results identified resynthesis *a*, illustrated by figure 4, as a better representative of the causal prediction.

3.4 Consequence Prediction

Resynthesis *a* was characterized by:

. F0 valley formation, with 2RV's;
. centralization of F0 valley in the attack consonant (C) of FSS;
. F0 raising, 5RV's over the base line.

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. duration of C in FSS 97% longer, equalizing it to the duration in MC1 causal of reference.

The results identified resynthesis *b*, illustrated by figure 4, as a better representative of the consequence prediction.

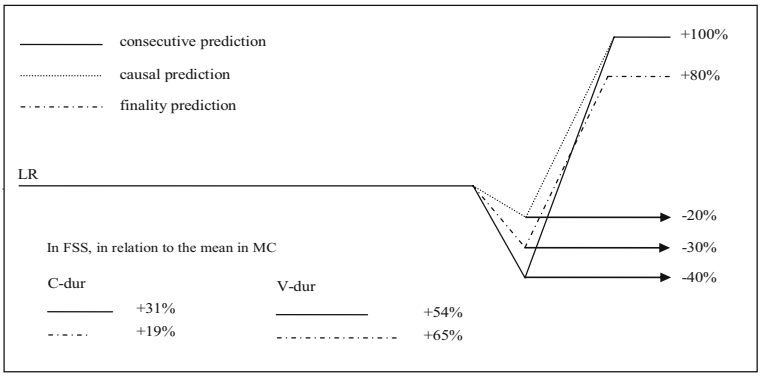

Fig. 4. Schematic representation of the predictive prosodic contours for consequence, cause, and finality, for BP. The descending movement of F0 begins on the final pre-stressed syllable. LR: reference line; FSS: final stressed syllable; MC: main clause; C-dur: consonant duration; V-dur: stressed vowel duration.

3.5 Finality Prediction

Resynthesis *a* was characterized by:

. F0 valley formation, with 1,5RV;
. centralization of F0 valley in the attack consonant (C) of FSS;
. F0 raising, 4RV's over the base line.

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. duration of C in FSS 54% longer, equalizing it to the duration in MC1 consecutive of reference.

The results identified resynthesis *b*, illustrated by figure 4, as a better representative of the finality prediction.

3.6 Concessive Prediction

Resynthesis *a* was characterized by:

. beginning of F0 descending movement 12% over the maximum pitch of the V in the foregoing syllable to the final pre-stressed syllable (FPSS); variation of -4% in the FPreSV; variation of -30% in the C of the FSS; variation of -50% in FSV; pitch rising of the whole sentence in 30% (to compensate for the total lowering);

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. creation of a concave shape for F0 in FSV, with a variation of 15%.

The results identified resynthesis *b*, illustrated by figure 5, as a better representative of the concessive prediction.

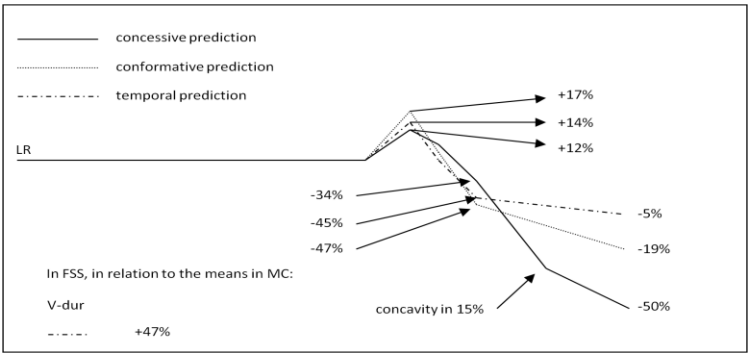

Fig. 5. Schematic representation of the predictive prosodic contours for concession, conformity, and temporality, for BP. The peak of F0 is located on the vowel of the syllable previous to the final pre-stressed syllable. LR: reference line; FSS: final stressed syllable; MC: main clause; C-dur: consonant duration; V-dur: stressed vowel duration.

3.7 Conformity Prediction

Resynthesis *a* was characterized by:

. beginning of F0 descending movement 17% over the maximum pitch of the V in the FPSS; variation of -10% in the FPreSV; variation of -37% in the C of the FSS; variation of -19% in FSV; pitch rising of the whole sentence in 30% (to compensate for the total lowering);

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. creation of a convex shape for F0 in FSV, with a variation of 20%.

The results identified resynthesis *a*, illustrated by figure 5, as a better representative of the conformity prediction.

3.8 Temporality Prediction

Resynthesis *a* was characterized by:

. beginning of F0 descending movement 14% over the maximum pitch of the V in the FPSS; variation of -11% in the FPreSV; variation of -34% in the C of the FSS; variation of -5% in FSV; pitch rising of the whole sentence in 30% (to compensate for the total lowering);

Resynthesis *b* was characterized by:

. the same procedure for resynthesis *a*;
. duration of V in FSS 61% smaller, equalizing it to the duration in MC1 temporal of reference.

The results identified resynthesis *b*, illustrated by figure 5, as a better representative of the temporality prediction.

4 Discussion

Although the prosodic predictive phenomenon existence has been confirmed by Test 1, considering the adverbial subordinate clauses in Rio de Janeiro variety of BP, the configuration of the components responsible for that predictive value cannot be easily specified. The attribution of a phonological status to the contours confirmed by Test 2 does not mean that variations of those contours cannot be noticed with the same predictive effect, considering a same adverbial value. In spite of the experiments having been developed only in Rio de Janeiro, it is probable that the results obtained would be similar for other varieties of BP, but further studies must be accomplished.

Considering the application of those melodic contours to vocal synthesis systems, it is important to notice that, although there are other potential contours, the ones identified in this research are recognized by BP speakers from Rio de Janeiro and Espírito Santo as predictive of the adverbial relations in question. More than the synthesis of primary prosodic parameters – such as assertive or interrogative modulation –, those contours application implies a perceptual sophistication, producing a final auditory result closer to the human language.

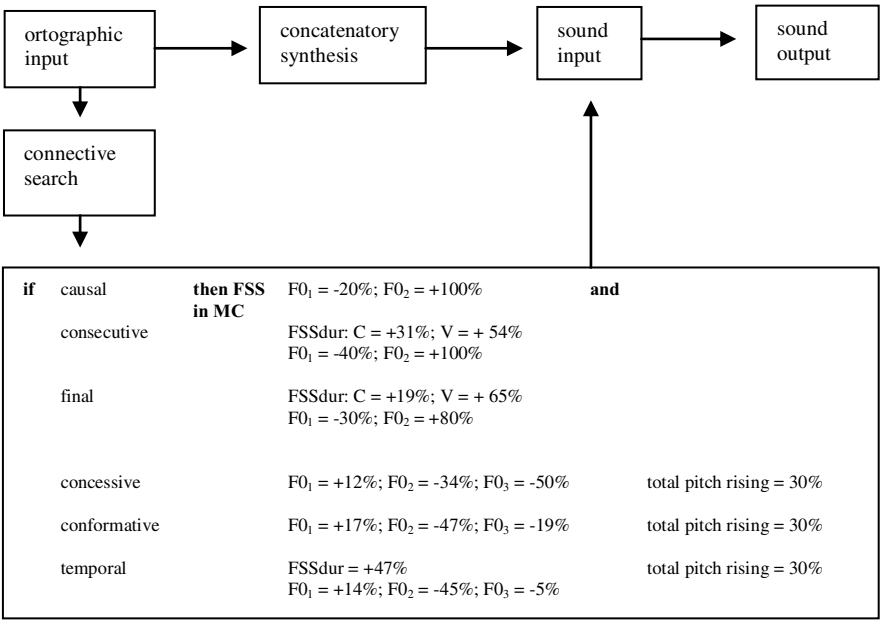

Fig. 6. Algorithm proposal for prosodic prediction modulation for adverbial clauses in Brazilian Portuguese. FSS: final stressed syllable; MC: main clause; F0: fundamental frequency; dur: duration.

It is possible that those configurations be applied to main clauses containing any durational extension. Prosodically, the significant melodic movements of a contour are active on the final stressed syllable. In that sense, it is enough, in order not to impair the effects of naturalness of the whole sentence, to make some adaptations on

the fundamental frequency, such as its initial rising, allowing a declination line with an angle open enough to cover all of the sentence extension, till the nucleus of the prosodic contour.

Considering the canonical order of a sentence in BP (MC – connective – SAdvC), the algorithm to be implemented (figure 6) should identify, as a first step, the connective. Then, it would apply the changes in the durational values and fundamental frequency retroactively, following the specifications in figures 4 and 5.

Acknowledgments. Special thanks are due to the judicious comments of three anonymous reviewers.

References

1. Bechara, E.: Moderna gramática portuguesa. São Paulo, Cia. Editora Nacional (1992)
2. Blin, L., Edgnington, M.: Prosody prediction from tree-like structure similarities. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2000. LNCS (LNAI), vol. 1902, pp. 369–374. Springer, Heidelberg (2000)
3. Boersma, P., Weenink, D.: Praat: doing phonetics by computer Version 4.3.01 (Computer program) (2005), http://www.praat.org/
4. Fónagy, I.: La fonction préindicative de l'intonation en français et en hongrois. In: Travaux de l'Institut d'Études Linguistiques et Phonétiques, vol. I, pp. 44–75 (1974)
5. House, D.: The influence of silence on perceiving the preceding tonal contour. In: Proc. Int. Congr. Phonetic Sciences 13, vol. 1, pp. 122–125 (1995)
6. Marsi, E.: Optionality in evaluating prosody prediction. In: 5th ISCA Speech Synthesis Workshop. Pittsburgh (2004)

The Role of Morphology in Generating High-Quality Pronunciation Lexica for Regional Variants of Portuguese

Simone Ashby and José Pedro Ferreira

Instituto de Linguística Teórica e Computacional (ILTEC)
Rua Conde de Redondo, 74-5º, 1150-109 Lisbon, Portugal
{simone,zpferreira}@iltec.pt

Abstract. Grapheme to phoneme (GTP) systems for languages such as English, German, and Korean have been shown to achieve better performance rates with the inclusion of a morpho-phonological preprocessing component. While semi-automatic and automatic GTP approaches for Portuguese continue to achieve steady gains, such algorithms do not take morphology into account, despite a growing need to do so, based in part on the recent spelling reform. This paper presents a pilot study in the development of the Portuguese Unisyn Lexicon (LUPo) for assessing the role of morphological information in the generation of high-quality pronunciation lexica for regional variants of Portuguese. Some problematic orthographic contexts are identified, along with the associated difficulties that arise when morphology is left out of the equation. Expanding from known issues that affect Portuguese GTP systems, new orthographic contexts stemming from the recent spelling reform are addressed.

Keywords: Morphology, pronunciation lexicon, grapheme to phoneme conversion, Portuguese accents, orthography, lexical database, speech synthesis.

1 Introduction

This paper presents a pilot study for assessing the role of morphology to generate high-quality pronunciation lexica for regional variants of Portuguese. The study marks the initial phase of a three-year project to develop an accent-independent pronunciation lexicon and rule system for generating accent-specific phonetic transcriptions for Portuguese, known as the Portuguese Unisyn Lexicon (LUPo). The project aims to model a multitude of Portuguese accents spanning Africa, Asia, Europe, and South America.

The motivation for this study stems from GTP studies such as [1], [2], [3], [4], and [5], which show that significant gains can be made by including morphological information in the transcription system for English, German, and Korean. Indeed, [6] demonstrates that some frequent errors in a state-of-the-art GTP system were in part the result of a *lack* of morphological preprocessing. Thus, the need for various levels of linguistic input in statistically derived speech processing algorithms continues to figure in the potential success of these technologies.

T.A.S. Pardo et al. (Eds.): PROPOR 2010, LNAI 6001, pp. 162–165, 2010.

Good performance rates, such as those reported in [7], [8], [9], and [10], are often cited for Portuguese GTP systems, with varying levels of analysis of the types of errors that occurred. In [7], the authors cite high rates of confusability in the classification of EP vowels, laterals, and phonetic realizations of the graphemes <x> and <s>. However, the authors go no further in analyzing the contexts in which such errors occurred. Rather, regarding the misclassification of laterals and some vowels, it is stated in [7] that such mismatched phones lack perceptual significance or serve as suitable alternate transcriptions. In [8], a more thoughtful discussion is given to understanding the distribution of errors in linguistically salient contexts for a knowledge-driven GTP algorithm for EP. Here, the authors cite homographs, verbal conjunctions, foreign loan words, proper names, and toponyms as the more problematic word classes. Yet, no specific attention is paid in [8] to analyzing classification errors in terms of the words' morphological constituents. The same is true of [9] for EP and [10] for Brazilian Portuguese (BP). Thus, it remains to be seen what impact morpho-phonological preprocessing, *or a lack thereof*, might have.

In the current study, we identify a set of problematic orthographic contexts, and assess the disadvantages that arise when morphological information is left out of Portuguese GTP algorithms. Some of the more well known issues affecting these systems are discussed, along with the implications of new orthographic contexts to be introduced as part of the new spelling accord.

2 LUPo

One of the distinguishing features that sets LUPo apart from other data-driven and knowledge-driven GTP systems – and, indeed, from the original Unisyn Lexicon for English [11] – is its inclusion within and immediate access to the multi-dimensional and lexicographically rich Portal da Língua Portuguesa online knowledge base [12], hereafter referred to as the 'Portal'. As one of many Portal modules, LUPo benefits from the ability to link entries to their inflected and derived forms, spelling variants, part of speech information, foreign loan word attributes, toponyms, and other types of lexicographic information. The Portal's morphological database, which is currently in development, ties entries to their morphological constituents, including base forms.

Unlike [7], which adopts a broad stance on sound substitutions deemed perceptually relevant for representing standard EP, and [13], whose authors sought to create a more abstract “neuter accent” for lack of a BP standard, LUPo is expressly devoted to creating pronunciation models for regional variants of Portuguese *as they are actually spoken* in different areas of the pan Lusophone world.

The end product will be a set of open-source tools for generating accent-specific output for individual lexical entries and multi-word texts. Inclusion of LUPo in the Portal will enhance the Portal's value as a pan Lusophone resource and the only one of its kind to provide detailed and varied phonetic output for a large number of Portuguese accents. Indeed, it will be the first freely available online resource to provide any manner of high-quality transcription data for Portuguese.

Given these objectives, the types of errors deemed acceptable in the GTP systems described in section 2 present LUPo with some of its biggest challenges. In the

following section, we attempt to come to a better understanding of these and other errors in terms of the contexts in which they occur, and the potential role of morphological information in generating higher quality transcription data.

3 The Costs of Ignoring Morphology

We examined orthographic and broad phonetic data at the word level for several EP and BP dialects. This was done as a means of more closely observing some of the contexts where a lack of morphological preprocessing might lead GTP systems astray.

As indicated in [7], vowel height presents one of the most challenging aspects of developing a good GTP system for Portuguese. The graphemes <e> and <o> in lexical roots are particularly troublesome, as the phones they represent are not predictable in many contexts without knowing the underlying phoneme. The latter is identifiable only via its realization in a stressed position. Yet, orthography is often opaque in such contexts (e.g. 's[o]pa' vs. 'r[O]da'; 'm[o]lho' vs. 'm[O]lho').

These phonemes behave consistently in words containing the same root, but the regularity they show across the different accents examined is not exploited in existing GTP systems for Portuguese. The notion that speakers access morphological information when producing vowels with different height attributes is well supported in phonological treatments, such as [14]. In the contexts mentioned, the correct phone can only be positively established by knowing the underlying phoneme of the root, the type of morphological boundary and the class of the base form in derived words (e.g., in EP accents, 's[u]pinha' vs. 'r[O]dinha', 'r[o]dinha' vs. 'r[u]dagem', 'm[u]lhada' vs. 'm[O]lhada'), along with stress position and syllable boundaries.

LUPo has the advantage of residing within a relational database that contains all of the above information. Thus, problematic word roots need only be hand-checked once to achieve consistent and accurate transcriptions for morphologically related inflected forms, derived words, and compounds.

The same principle applies to other problematic contexts, namely words with <x> in the root (e.g. 'e[z]ame' vs 'sinta[s]e'), and unstressed hiatuses that are not marked orthographically, as in the example 'cairão' (from the verb 'cair') or 'miudagem'.

New hyphenation rules in the recent spelling reform gave rise to a large number of problematic contexts for hiatus identification. Preposed formation elements ending in a vowel no longer take a hyphen when the second element starts with a different vowel, as in the examples 'c[oi]gual', 'c[ou]tente', 'ret[Ou]retral' (cf. 'c[ow]tada', and 'ret[ow]çar'). These vowel combinations usually represent diphtongs. Add this to the existing problem of neoclassical compounds with preposed combining elements ending in <o> (e.g. 'retr[O]carga' vs. 'retr[u]car'), where the linking element is pronounced as [+low] in several accents. In both these cases, having access to morphological boundaries would enable GTP systems to tell these sequences apart from non-compounds, correctly transcribe them, and save the costs of checking all the words with problematic orthographic sequences manually.

4 Conclusion

Including morphology in Portuguese GTP systems poses significant benefits for the generation of high-quality pronunciation data. Most of the types of errors mentioned in the literature can be addressed with such an approach, and greater consistency can be achieved by treating morphologically related words in one pass. The recent spelling reform further strengthens these claims.

References

1. Fitt, S.: Morphological Approaches for an English Pronunciation Lexicon. In: Proceedings of Eurospeech, pp. 1069–1072 (2001)
2. Wothke, K.: Morphologically Based Automatic Phonetic Transcription. IBM Systems Journal 32, 486–511 (1993)
3. Reichel, U.D., Schiel, F.: Using Morphology and Phoneme History to Improve Grapheme-to-Phoneme Conversion. In: Proceedings of Eurospeech, pp. 1937–1940 (2005)
4. Kim, B., Lee, G., Lee, J.: Morpheme-Based Grapheme to Phoneme Conversion Using Phonetic Patterns and Morpho-phonemic Connectivity Information. ACM Transactions on Asian Language Information Processing 1(1), 65–82 (2002)
5. Yoon, K., Brew, C.: A Linguistically Motivated Approach to Grapheme-to-Phoneme Conversion for Korean. Computer Speech and Language 20(4), 357–381 (2006)
6. Taylor, P.: Hidden Markov Models for Grapheme to Phoneme Conversion. In: Proceedings of Interspeech, pp. 1973–1976 (2005)
7. Barros, M.J., Weiss, C.: Maximum Entropy Motivated Grapheme-to-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech. In: Proceedings of the 4th Biennial Workshop on Speech Technology, pp. 177–182 (2006)
8. Braga, D., Coelho, L., Resende Jr., F.: A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese. In: Proceedings of the 6th International Telecommunications Symposium, pp. 328–333 (2006)
9. Caseiro, D., Trancoso, I., Oliveira, L., Viana, C.: Grapheme-to-Phone Using Finite-State Transducers. In: Proceedings of the IEEE Workshop on Speech Synthesis, pp. 1349–1360 (2002)
10. Silva, D., Lima, A., Maia, R., Braga, D., Moraes, J.F., Moraes, J.A., Resende Jr., F.: A Rule-Based Grapheme-Phone Converter and Stress Determination for Brazilian Portuguese Natural Language Processing. In: Proceedings of the 6th International Telecommunications Symposium, pp. 992–996 (2006)
11. Fitt, S.: Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules. Online Technical Report, Centre for Speech Technology Research, University of Edinburgh (2000), `http://www.cstr.ed.ac.uk`
12. Portal da Língua Portuguesa, `http://www.portaldalinguaportuguesa.org`
13. Barbosa, F., Pinto, G., Resende, F.G., Gonçalves, C.A., Monserrat, R., Rosa, M.C.: Grapheme-Phone Transcription Algorithm for a Brazilian Portuguese TTS. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.d.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 23–30. Springer, Heidelberg (2003)
14. Mateus, M.H., d'Andrade, E.: The Phonology of Portuguese. Oxford University Press, New York (2002)

Author Index

Adami, André Gustavo 132
Aluísio, Sandra M. 40
Ashby, Simone 162
Atkinson, Martin 21

Baptista, Jorge 110, 120
Barone, Dante Augusto Couto 132
Batista, Pedro 128
Branco, António 75, 86

Cabral, Maria 120
Carvalho, Gracinda 1
Caseli, Helena de Medeiros 65
Castro, Sérgio 75
Coimbra, Rosa Lídia 142
Costa, Francisco 86
Costa, Neuza 120

Fernandes, Eraldo R. 55
Ferreira, José Pedro 162
Finatto, Maria José 65
Flores, Felipe N. 11
Franco, Wellington 90
Furtado, Vasco 90

Gasperin, Caroline 40
Gomes, Fernando 110
Grant, Tim 51
Guerra, Joaquim 120

Heuser, Carlos A. 11

Jorge, Maria Lucía Castro 25

Klautau, Aldebaro 128

Linge, Jens 21
Lopes, Carla 124

Machado, André 65
Maia, Belinda 51
Mamede, Nuno 110, 120
Matos, David Martins de 1
Maziero, Erick 40
Milidiú, Ruy L. 55
Moreira, Viviane P. 11

Neto, Nelson 128
Nunes, Ana 142

Oliveira, Eugénio 51, 100

Pardo, Thiago Alexandre Salgueiro 25
Pequeno, Tarcisio 90
Perdigão, Fernando 124
Pinheiro, Vladia 90
Piskorski, Jakub 21

Ramisch, Carlos 65
Reis, Ruben 75
Rocio, Vitor 1

Santos, Cícero N. dos 55
Santos, Fabiano Weimar dos 132
Sarmento, Luís 51, 100
Silva, João 75
Silva, Patrick 128
Sousa-Silva, Rui 51
Specia, Lucia 30
Steinberger, Ralf 21
Stein, Cirineu Cecote 152

Tanev, Hristo 21
Teixeira, António 142
Teixeira, Jorge 100

Veiga, Arlindo 124
Villavicencio, Aline 65

Zampieri, Marcos 120
Zavarella, Vanni 21