

# Portuguese Pronoun Resolution: Resources and Evaluation

Ramon Ré Moya Cuevas, Willian Yukio Honda, Diego Jesus de Lucena,  
Ivandré Paraboni and Patrícia Rufino Oliveira<sup>1</sup>

<sup>1</sup> Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH / USP)  
Av. Arlindo Bettio, 1000 - 03828-000, São Paulo, Brazil  
{fusion, kio, diego.si, ivandre, proliveira}@usp.br

**Abstract.** Despite being one of the most widely-spoken languages in the world, Portuguese remains a relatively resource-poor language, for which only in recently years NLP tools such as parsers, taggers and (fairly) large corpora have become available. In this work we describe the task of pronominal co-reference annotation and resolution in Portuguese texts, in which we take advantage of information provided by a tagged corpus and a simple annotation tool that has been developed for this purpose. Besides developing some of these basic resources from scratch, our ultimate goal is to investigate the multilingual resolution of Portuguese personal pronouns to improve the accuracy of their translations to both Spanish and English in an underlying MT project.

## 1 Introduction

Pronoun resolution – the task of identifying the antecedent of a pronoun in discourse - is often crucial to a variety of NLP applications, ranging from Text Summarization to Machine Translation (MT), and it has long been recognised as a challenging computational problem for which existing approaches - either making use of learning techniques or otherwise - often need to resort to large amounts of knowledge produced by standard NLP tools such as parsers and POS taggers applied on annotated corpora [1].

The challenge is however considerably increased if we speak of languages for which some of these basic resources are still under development, or may have only recently become available. This is the case, for instance, of Portuguese, one of the most widely-spoken languages in the world, and which still lacks somewhat behind as a relatively resource-poor language in NLP.

Our own work focuses on pronoun resolution as required by a Portuguese-Spanish-English MT project under development. Our present choice - Portuguese third person plural pronouns (“Eles/Elas”) - is based on the assumption that these (as well as their Spanish counterparts) are less prone to ambiguity, and arguably easier to resolve than the English equivalent (“They”), which may suggest an interesting multilingual approach to anaphora resolution not unlike [7].