UNIVERSIDADE DE ÉVORA

Mestrado em Engenharia Informática

Resolução de Anáforas Pronominais em Documentos em Língua Portuguesa

Ana Margarida Pereira dos Santos Aires

<aaires@di.uevora.pt>

Orientador: Prof. Doutor Paulo Quaresma

<pq@di.uevora.pt>

Outubro de 2006

UNIVERSIDADE DE ÉVORA

Mestrado em Engenharia Informática

Resolução de Anáforas Pronominais em Documentos em Língua Portuguesa

Ana Margarida Pereira dos Santos Aires

<aaires@di.uevora.pt>

Orientador: Prof. Doutor Paulo Quaresma

<pq@di.uevora.pt>

Outubro de 2006

Prefácio

Este documento contém uma dissertação intitulada *Resolução de Anáforas Pronominais em Documentos em Língua Portuguesa*, um trabalho da aluna Ana Margarida Pereira dos Santos Aires[1], estudante de Mestrado em Engenharia Informática na Universidade de Évora.

O orientador deste trabalho é o Professor Doutor Paulo Quaresma[2], do Departamento de Informática da Universidade de Évora.

A autora deste trabalho é licenciada em Engenharia Informática, pela Universidade de Évora. A presente dissertação foi entregue em Outubro de 2006.

[1] aaires@di.uevora.pt
[2] pq@di.uevora.pt

Resumo

O processo de resolução de anáforas é fundamental para compreender um texto, enquanto que o ser o humano o faz com facilidade, simulá-lo computacionalmente não é tarefa fácil. O grande objectivo deste trabalho, está em construir um sistema que dê ao computador a capacidade de inferir para anáforas pronominais, quais os seus antecedentes.

O sistema desenvolvido é baseado na metodologia do *centering*, não só pelos seus princípios, mas também pela possível adequação à língua portuguesa. A avaliação dos resultados obtidos, reflectiu algumas limitações, comuns a este tipo de sistemas, pelo que foi proposta e implementada, uma alteração ao algoritmo inicial, com acréscimo de três extensões que permitem preferir uma solução às restantes, em caso de empate. Pela nova avaliação, conclui-se uma melhoria de eficiência na segunda versão do algoritmo que tem em média uma taxa de sucesso crítica de 54% que se entende bastante positiva, uma vez que não se dispunham de corpora isentos de erros de pré-processamento.

Abstract

Pronominal Anaphora Resolution in Portuguese Language Documents

The process of anaphora resolution is fundamental for the understanding of a text and although a human can do it easily, simulate it on the computer isn't a trivial task. The main goal of this work is to develop a system capable of mining the computer with the capacity to associate pronoun anaphor with the expression they refer to. The developed system is based on the methodology known as centering, not only due to its core ideas, but also because of it's adaptability to the Portuguese language. The evaluation of the results obtained showed some limitations, common to these type of systems which lead to a proposal and implementation of improvements, over the first approach, with three extensions that overcome draw situations. The new evaluation shows a improvement over the second version of the algorithm, and has a critical success rate of 54% on average, which is believed to be quite positive considering that no corpora, free of pre-processing errors, was available.

Agradecimentos

O trabalho aqui desenvolvido, não teria sido possível sem o apoio de várias pessoas, e é com grande prazer, que aproveito esta oportunidade para agradecer a todas elas.

Agradeço ao meu orientador, o Prof. Doutor Paulo Quaresma, não só pelo apoio dado neste trabalho em particular, mas também por me ter inciado na investigação na área de processamento de língua natural. Agradeço aos colegas do Brasil, da Universidade do Vale Do Rio dos Sinos, por todos incentivos e esclarecimentos, dados durante a cooperação com o Departamento de Informática da Universidade de Évora, com especial atenção à Prof. Doutora Renata Vieira, à Mestre Sandra Collovini Abreu, e o licenciado César Coelho. Outro agradecimento muito especial, por todo carinho, apoio e incentivo, é dedicado ao companheiro da minha vida, Nuno Morgadinho e às grandes mulheres da minha família, a minha avó Maria Claudina, a minha mãe Odália Aires e a minha irmã Isabel Aires.

Ana Aires

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Esta dissertação resulta de um estudo feito no domínio de processamento de língua natural, particularmente para a língua portuguesa. Descreve o desenvolvimento e teoria de uma aplicação, capaz de resolver a ocorrência de anáforas, especificamente pronomes pessoais, com os seus respectivos antecedentes, baseada na metodologia do *Centering*.

1.1 Motivações

O trabalho desenvolvido está inserido no "mundo" extraordinário da **inteligência artificial**. Uma vez que se dedica, a permitir que o computador realize operações, que exigem comportamento inteligente, nomeadamente inferir num discurso, que entidades são equivalentes, e quais os seus referentes. Ao falar-se em discurso, particulariza-se a área da inteligência artificial em que o estudo está inserido. A sub-área, **processamento de língua natural (PLN)**, que se entenda por língua natural, a linguagem utilizada por humanos, nesta sub-área podem distinguir-se especialmente dois tipos de sistemas:

- Sistemas que convertem informação contida em bases de dados, em linguagem natural.
- Sistemas que convertem a linguagem natural numa representação formal, que permite manipulação computacional

As principais tarefas associadas a esta área são o desenvolvimento de sistemas de: tradução, pergunta/resposta, reconhecimento de discurso, sumarização automática, extracção de informação, entre outros. Uma das grandes limitações de todos eles, prende-se com a necessidade de conhecimento do mundo real, para se interpretar diferentes situações, que ocorrem na língua natural. O texto em (1.2) é disso exemplo.

(1.2)*As meninas comeram* ***as sopas***, ***elas*** *estavam a arrefecer.*
As meninas *comeram as sopas,* ***elas*** *tinham muita fome.*

Neste caso o pronome *elas* refere-se a duas entidades diferentes, em cada situação. A inferência não pode ser feita correctamente, sem informação sobre características das *meninas*, assim como das *sopas*.
O que para uns é visto como uma tarefa quase impossível, para os investigadores nesta área é interpretado como um desafio.

Para se compreender um discurso, a resolução de anáforas é fundamental. Ela é feita intuitivamente pelos leitores/ouvintes, e é alvo de vários estudos que tentam tornar o computadores capazes de reproduzir este comportamento. O que só por si é motivação para o trabalho aqui apresentado. Antes de proseguir, veja-se um exemplo de uma anáfora pronominal, às quais se dedica especial importância neste estudo. A descrição em detalhe deste fenómeno é deixada para o capítulo seguinte.

No exemplo ,

> *"Acabar com as contratações cíclicas e passar para as escolar a responsabilidade de contratar os seus professores está nos planos do Ministério da Educação. Os sindicatos desconhecem* ***a proposta.*** *Receberam-****na*** *apenas ontem ao final da tarde e acusam a tutela de não ter negociado a medida..."*[3]

o pronome *na*, corresponde a uma anáfora pronominal, que refere uma entidade já introduzida no discurso. O seu antecedente é facilmente identificado com *a proposta.* É essa associação, que se pretende modelar o computador a realizar, essa tarefa é referente aos sistemas de resolução de anáforas, e é fundamental para o desempenho dos sistemas de processamento de língua natural acima identificados. De seguida, para alguns deles, explica-se essa "dependência".

Os primeiros sistemas de **tradução** desenvolvidos entre os anos 70 e 80, não dedicavam especial importância, à resolução de anáforas, como sua parte integrante. Como consequência, apenas um número limitado obteve resultados satisfatórios. A justificação para esta situação não está só relacionada com a dificuldade da tarefa de resolução de

[3]fonte: jornal Público 28 Outubro 2006

anáforas, mas também com o custo adicional que o problema de tradução impõe. Isto é, ao traduzir um texto, não basta que a máquina "perceba"qual a entidade de determinado referente, como se exige que no texto de output (traduzido) essa ligação continue a existir. Outro aspecto, que marca a real importância da identificação, da associação entidade-referente, para o caso particular da tradução, é a discrepância entre géneros de pronomes para diferentes línguas, ou a diferença entre línguas que atribuem género e as que não. Por exemplo, ao traduzir a seguinte frase em inglês: *"**The chair** was broken, **it** was very old."* obtém-se *"**A cadeira** estava partida, **ela** era muito velha."*. O pronome *it* não tem género atribuído enquanto *ela* tem. A preocupação em incorporar, resolução de anáforas para sistemas de tradução aumentou pelos anos 90, em consequência dos vários trabalhos promissores, desenvolvidos nesta época, como por exemplo o de [H.Wada(1990].

O objectivo fundamental da tarefa de **extracção de informação**, é a partir de um texto sem *metadata*, extrair informação sobre grupo de entidades, relações ou eventos. Pode ser usada por exemplo, para popular uma base de dados, com a informação que se extrai de textos *on-line*. É perfeitamente visível a importância não só de resolução de anáforas nestes sistemas, mas principalmente de resolução de co-referência, portanto encontrar entidades, descritas em várias posições no discurso, que são equivalentes, mais do que isso, referem o mesmo elemento no mundo real. Em 1999, foi apresentado um sistema de extracção de informação, com incorporação de resolução de anáforas,que se destina a identificar e analisar depoimentos em tribunal. [Al-Kofani *et al.*(1999].

Os estudos de **sumarização**, dedicam especial importância ao processo de resolução de anáforas. Estes são mais eficientes, sempre que a determinação das frases mais importantes de um texto, é feita com apoio da identificação de referentes anafóricos, dos conceitos indicativos de importância. Supondo que existem um conjunto de documentos sobre o Presidente da República, dos quais se pretende fazer uma sumarização , esta será mais eficiente, se se fizer uso, de um sistema de resolução de co-referência, que permita identificar/extrair todas as frases em que a entidade *Presidente da República* seja referida.

Em sistemas de **pergunta/resposta** a identificação de cadeias de co-referência[4], assim como resolução de anáforas, pode permitir, por exemplo, atribuir uma classificação às várias frases, conforme elas dizem respeito, ou não (há ou não alguma relação anafórica, ou de co-referência) a determina entidade, para a qual foi feita alguma questão. O resultado será o conjunto de frases melhor classificadas.

Estes são alguns exemplos onde a resolução de anáforas, pode marcar a diferença, na eficiência de aplicações de processamento de língua natural, independentemente da sua natureza.

Sendo a língua portuguesa, a oitava língua mais falada em todo o mundo, onde pessoas entre os 170 e os 210 milhões falam português, parece natural o desenvolvimento de mais e melhores aplicações de processamento de língua natural, para este idioma. Se o desenvolvimento, e eficácia desses sistemas está intimamente ligado à resolução de anáforas, não há outra maior motivação , para este trabalho aqui desenvolvido, que é um dos poucos, a dedicado à resolução de anáforas para o português.

1.2 Objectivos

O principal objectivo deste trabalho é simular o processo de inferência, característico dos humanos, que dado um qualquer discurso bem estruturado, facilmente se associa um referente à sua entidade já introduzida. Por exemplo no excerto,

> "*...Os homens estavam encapuçados e lançaram fogo, ao* ***veículo*** *depois de terem evacuado os quinze passageiros e o condutor que* ***nele*** *seguiam...*"[5]

o leitor contextualizado, facilmente associa o pronome *nele* ao seu antecedente *veículo*. Considera-se que quanto menor o custo desta tarefa, mais coerente é o discurso, ou seja, um discurso é tanto coerente, quanto melhor permite, em qualquer momento, associar com o menor esforço, entidades e seus referentes. Este objectivo, faz com que na prática,

[4]conjunto de elementos no texto, que se referem à mesma entidade no mundo real

[5]fonte:Jornal Público, 28 de Outubro de 2006.

se pretenda construir um sistema de resolução de referência anafórica , que se irá dedicar, ao encontro das entidades antecedentes para cada pronome pessoal(anáforas pronominais) da terceira pessoa tanto no singular como no plural, encontrado num texto. Será desenvolvido em *Prolog*, linguagem de programação de excelência para problemas de processamento de língua natural. Terá como base a teoria do *Centering*, uma vez que esta, se baseia principalmente, na noção de coerência, que é um princípio comum a qualquer língua. Da construção deste sistema, surgem outros objectivos, nomeadamente a sua análise fiável que obtenha uma medida de sucesso igualmente exacta.

Obtidos os primeiros resultados, da implementação baseada na teoria de [Brennan *et al.*(1987], pretende-se elaborar uma análise dos mesmos, da qual deve resultar propostas de alterações, e /ou extensões, que se entendam vantajosas e aproximem a solução à língua portuguesa, com finalidade principal, a melhoria do desempenho. Estas propostas devem ser implementadas, devem sofrer avaliação e por fim os seus resultados devem ser igualmente interpretados.

Entre os vários objectivos, destaca-se a contribuição do estudo no domínio de processamento de língua natural para o português, onde pouco tem sido feito. Pretende-se perceber, não só até que ponto esta metodologia em particular, pode ser aplicada ao português, mas também, qual a influência do género literário nos resultados obtidos.

É uma tarefa ambiciosa, ensinar o computadores a resolver problemas de linguagem, mas igualmente estimulante.

1.3 Contribuições

Todo o trabalho, representa em si, uma contribuição na área do processamento de língua natural, em particular, para a língua portuguesa. Especificamente, estudos e aplicações com sistemas de resolução de anáforas, não existem em abundância, e os que existem, dedicam-se a resolver diferentes tipos, por exemplo anáforas que resultam da ocorrência de pronomes possessivos. Daqui resulta que não há apenas o contributo da aplicação propriamente dita, mas também a possibilidade de comparação, entre outros

sistemas, baseados noutras metodologias que não, a do *centering*.

Além da aplicação em si, se se pensar na sua integração com outros sistemas de processamento de língua natural, por exemplo um sistema pergunta/resposta, os bons resultados aqui obtidos, poderão influenciar positivamente também esses sistemas.

A secção dedicada à avaliação e principalmente, a secção de conclusões, enumeram algumas das limitações da aplicação, que resultam de um estudo em detalhe dos resultados obtidos. Estas limitações servem de ponte ao trabalho futuro. Não serão observações úteis, apenas para a extensão da metodologia aqui desenvolvida, como podem servir de apoio a outras metodologias em estudo.

De uma forma mais objectiva, as principais contribuições deste trabalho são:

- A apresentação de um estudo comparativo de diversas metodologias de resolução de anáforas
- Implementação de um sistema de anáforas pronominais, baseado na metodologia do *centering*
- Aplicação da metodologia desenvolvida, a um conjunto de documentos em língua portuguesa.
- Definição das medidas de avaliação com vista a separar a avaliação do algoritmo da avaliação geral do sistema.
- Avaliação dos resultados obtidos.
- Proposta e avaliação de extenções, à implementação inicial.

1.4 Estrutura

A estrutura desta tese, está dividida em sete grandes capítulos, onde naturalmente o primeiro diz respeito à introdução que descreve principalmente motivações e objectivos. Os restantes serão aqui resumidos, indicando assim o que de mais importante tratam.

Capítulo 2 é reservado à descrição do estado de arte da temática resolução de anáforas. Começa por definir o que é uma anáfora, descreve o processo de resolução de anáforas e conhecimentos em que estes se podem basear. Introduz quatro diferentes abordagens, *Lappin* e *Leass*, *Hobbs' Algorithm*, *Mitkov's Robust knowledge-poor Algorithm* incluindo a teoria do *Centering* por Brennan, Friedman e Pollard. Esta última é referida sucintamente, o seu detalhe corresponde ao capítulo seguinte. Por fim, faz-se um apanhado da evolução dos processos de resolução de anáforas desde os anos 80.

Capítulo 3
Este capítulo descreve detalhadamente a metodologia do *Centering* que vai servir de base ao estudo aqui apresentado. Mostra a sua origem, define os princípios básicos em que se baseia e por fim enuncia algumas derivações da teoria original.

Capítulo 4 é reservado à descrição do sistema de resolução de anáforas pronominais desenvolvido. Engloba uma descrição das várias fases do algoritmo, desde o pré-processamento até ao candidato resultante para cada pronome. Mostra um exemplo da aplicação do algoritmo, ao que se segue a descrição de uma outra versão, em alternativa à primeira implementação, o *centering* por orações. As páginas finais, dedicam-se à descrição de três extensões propostas à segunda versão do algoritmo implementada.

Capítulo 5 Este capítulo dedica-se à descrição dos corpora utilizados. Não só os "analisa"qualitativamente, de acordo com os seus géneros literários, como quantitativamente, indicando o número de frases, de palavras, de palavras distintas, e anáforas que cada texto, e consequentemente cada corpus, contempla.

Capítulo 6 Esta, é uma secção de extrema importância, porque faz a avaliação das implementações e suas extensões desenvolvidas. Começa por descrever os métodos usados na avaliação, em seguida apresenta os resultados tabelados por corpus e por fim interpreta os resultados.

Capítulo 7 É naturalmente a conclusão. Sintetiza o objectivo primário deste trabalho, indica o que foi feito que limitações se encontraram, e por fim faz as propostas de trabalho futuro.

2 Estado da Arte

2.1 Anáfora

A definição de anáfora está associada a noção de coesão, e por isso é a partir desta que se vai chegar à primeira.

Sempre que existe comunicação, seja ela escrita ou falada, observa-se um fenómeno a que se chama **coesão**. Em linguística, diz-se que coesão, é a característica da comunicação, que em determinado momento, centra o tópico num aspecto em particular, faz o discurso centrado nesse tópico e só depois altera o assunto. Pode-se afirmar que um discurso, é constituído por uma sequência de frases, relacionadas entre si pelo seu conteúdo, e não uma sequência de frases soltas sem qualquer espécie de ligação entre elas. A coesão, observa-se sempre que se recorre a entidades equivalentes, para referir outras já introduzidas no texto, e que o leitor ou ouvinte, associa facilmente, como representantes da mesma entidade . Vejamos um exemplo:

> (2.1)*"Mas* ***a minha mãe*** *sonhava para mim com um casamento acima do nosso nível social e o Nicolas era o presente que* ***ela*** *pedira aos céus. Foi com medo de o perder que* ***ela*** *concordou em deixar-me ir morar com ele..."*[6]

Neste caso assume-se que há relação entre a primeira e segunda frase, e faz-se automaticamente a equivalência entre os dois elas marcados a negrito com *a minha mãe*. É precisamente esta equivalência que mantém a coesão do texto.

Introduzida a noção de coesão é altura de definir **anáfora**. Estamos presente uma expressão anafórica, quando esta refere uma entidade previamente introduzida no texto. Portanto anáfora é a entidade que refere outra já conhecida, a que chamamos **antecedente**. Quando estes dois elementos representam a mesma entidade no mundo real, eles são **co-referentes**, como está ilustrado em (2.2). O processo de **resolução de anáforas** não é

[6]Rosa Lobato Faria,*O prenuncio das águas*,p.40, ASA - 2003

mais do que, para cada uma, encontrar o seu antecedente. Sempre que existem, anáforas co-referentes entre si, estamos perante uma cadeia coreferencial, que tem grande utilidade no processo de resolução das expressões anafóricas.

(2.2) *José Saramago* ganhou um Nobel da literatura. *O escritor* é um marco na escrita portuguesa, com muitas provas dadas, *ele* será sempre um nome a ficar na história.

Neste exemplo *José Saramago*, *O escritor* e *ele* representam a mesma entidade no mundo real, sendo por isso co-referentes. Tanto *O escritor*, como o pronome pessoal *ele*, são anáforas, entidades que referem *José Saramago* isto é, têm-no como seu antecedente.

2.2 Tipos de Anáforas

De acordo com [Mitkov(2002], as anáforas podem ser classificadas segundo a sua forma, ou pela sua localização comparativamente ao antecedente. De acordo com esta última categoria, podemos dividi-las em anáforas cujo **antecedente está na mesma frase**, e anáforas cujo **antecedente se situa em frases anteriores**.

Relativamente à forma, o leque de alternativas alarga-se.

2.2.1 Anáfora Pronominal

Como o próprio nome indica, anáfora **pronominal** diz respeito a todas as anáforas na forma de pronomes. Quanto ao número, ocorrem tanto no singular como no plural, mas relativamente ao género são mais frequente na 3ª pessoa. Existem na forma de pronomes pessoais (2.3), possessivos (2.4), relativos (2.5), e demonstrativos (2.6).

(2.3) Adoro a minha mãe, *ela* é a melhor mãe do mundo.

(2.4) A Isabel é linda, com o *seu* cabelo castanho aos caracóis.

(2.5) O gato *cujo* pelo é branco chama-se João.

(2.6) A avó contou-me uma história para dormir, *aquela* que eu sempre queria ouvir.

2.2.2 Anáforas Definidas

Este tipo de anáfora ocorre sempre que uma descrição definida[7] é antecedida de i) uma expressão com o mesmo núcleo e refere-se à mesma entidade no discurso; ii)com núcleo diferente mas que se refere à mesma entidade; iii) um elemento não co-referente. (2.7) exemplifica a situação ii).

(2.7) *O Público* noticiou a maior apreensão de cocaína de sempre. *O jornal* obteve fontes directamente da polícia judiciária.

2.2.3 Substantivo Anafórico

(2.8) Não vou comer um *crepe* com amêndoas, apenas *um* simples.

Neste exemplo *um* refere-se a crepe, e não a *crepe com amêndoas*. Esta distinção é muito importante, pois este, é o caso em que o referente diz respeito ao núcleo do sintagma nominal e não ao sintagma completo.

2.2.4 Anáforas Verbais e Adverbiais

Ocorre uma anáfora verbal ou adverbial, sempre que a interpretação de um verbo ou um adverbio é dependente da sua relação anafórica com o seu antecedente. O exemplo (2.9) ilustra esta situação.

(2.9) A Jacinta *assinou o Público* por um ano, assim *fez*, o António.

[7]Grupo de palavras começado por um artigo definido e que tem um nome como núcleo [de Abreu(2005]

2.2.5 Anáforas Vazias

Anáforas **Vazias** (), também chamadas de **elipses** são anáforas "invisíveis"pois não ocorrem no texto na forma de alguma palavra ou frase. Este é uma representação sofisticada das anáforas, que visa reduzir a quantidade de informação sob forma abreviada. As formas mais comuns são, redução de pronomes (2.10), nomes (2.11) e verbos (2.12).

(2.10) *A Ana* tem um carro novo. () Foi ontem experimentá-lo.

(2.11) O Nuno ofereceu-me *chocolates*, não resisti e poucos () ficaram para o outro dia.

(2.12) *Preferes* sopa de legumes ou () canja?

Para concluir a referência aos vários tipos de anáforas, há que referir duas situações. Primeiro a existência de **anáforas indirectas** (2.13), casos em que a referência ao antecedente é feita pelo leitor ou ouvinte indirectamente, sem que esteja explícito no discurso. Exige, por vezes, conhecimento adicional por parte do leitor/ouvinte. Em segundo, referir que sempre que uma entidade referir outra, que só aparece em seguida no texto, está-se perante uma **catáfora**(2.14) e não de uma anáfora.

(2.13) *Nirvana* é uma banda fantástica. *Kurt Cobain* era um elemento essencial.

(2.14) *Ele* vive em Espanha mas é Português, já ganhou um prémio Nobel e é um dos grandes nome da literatura portuguesa, todos conhecem *José Saramago*.

Para o estudo em questão, o interesse recai sobre as anáforas pronominais.

2.3 Processo de Resolução de Anáforas

> "A real-world anaphora resolution system vitally depends on the efficiency of the pre-processing tools which analyse the input before feeding it to the resolution algorithm."[8]

O processo de resolução de anáforas implica para o leitor, ou ouvinte, uma associação entre o referente e o antecedente. Esta identificação é feita usando o conhecimento do discurso que o leitor ou ouvinte adquiriu até ao momento, assim como o conhecimento, que ele possui, do mundo envolvente. Para a resolução automática de uma anáfora muitas estratégias, individualmente ou em conjunto podem ser utilizadas, fazem uso da informação linguística e cognitiva. A informação linguística é providenciada, principalmente pela análise sintáctica e semântica,enquanto que a cognitiva está incorporada nos modelos computacionais do discurso. Daqui resulta, que a existência de um pré-processamento, que adicione ao input do algoritmo o maior número possível de informação, nestes domínios, é de grande utilidade para o sucesso do processo de resolução de anáforas. É certo que há custo temporal e possivelmente a introdução de alguns erros, contudo se esta informação for o mais exacta possível, os resultados também o serão. Recentemente tem-se demonstrado que métodos *knowledge-poor* têm sido igualmente eficazes, na resolução de certas formas anafóricas.[9]

Alguns métodos computacionais que fizeram uso de informação linguística e cognitiva foram [Hobbs(1978], [Lappin & Leass(1994],[Webber(1988] e [Brennan *et al.*(1987], como se poderá ver mais adiante. Uma abordagem *knowledge-poor* pode ser vista com mais detalhe em [Mitkov(1998].

[8] A vitalidade de um sistema de resolução de anáforas do mundo-real depende da eficiência das ferramentas de pré-processamento, que analisam o input antes que este seja passado ao algoritmo de resolução.[Mitkov(2001]

[9] [Mitkov(1998].

2.3.1 Conhecimentos Fundamentais para a Resolução de Anáforas

2.3.1.1 Conhecimento Morfológico e Lexical

Exigir a concordância do género e do número entre a anáfora e o antecedente, em muitos casos é suficiente para encontrar a solução (2.15), noutros permite-nos descartar hipóteses que morfologicamente não fazem sentido(2.15).

> (2.15) *Os vizinhos* do *João* foram aos toiros. *Eles* insistiram para que o João e a *Diana* também fossem, mas *ele* preferiu ficar em casa a estudar e *ela* preferiu ir até à praia.

Com este exemplo pode-se observar ambas as situações descritas acima. O pronome pessoal *ela* apenas concorda em género e número com *a Diana* pelo que, *a Diana* e o antecedente desta anáfora, assim como o nome próprio *João* é o antecedente do pronome *ele*.
Relativamente ao pronome *Eles*, descartam-se os candidatos *João* e *a Diana* pois não concordam em número, ficando como possíveis candidatos os elementos *Os vizinhos* e *toiros*.

Este tipo de concordância é extremamente útil na resolução de anáforas pronominais, contudo há excepções, veja-se:

> (2.16) O João comprou uma casa em conjunto com a Diana. Eles vão casar no próximo mês.

neste exemplo *eles*, de acordo com análise morfológica, não tem antecedente possível. Num algoritmo que exija concordância entre género e número, o antecedente para este pronome não será encontrado.

2.3.1.2 Conhecimento Sintáctico

Este é um conhecimento indispensável à resolução de anáforas, não só fornece informação do que é por exemplo, um sintagma nominal, um pronome um verbo, como faz a divisão da frase em orações e consequentemente possibilita estabelecer regras. Regras estas, que quando impostas permitem determinar se um elemento é sintácticamente compatível com a anáfora e assim encontrar mais um candidato, ou eleminar um possível antecedente. Por exemplo, sintácticamente um pronome que ocorre numa oração, nunca pode ser resolvido com um sintagma nominal que ocorre numa oração subordinada à primeira (2.17). Este tipo de análise exige um Parser capaz de extrair esta informação de um texto.

(2.17) Ele sempre pensou em voltar para Portugal, porque Diogo tinha ali sua família.[10]

2.3.1.3 Conhecimento Semântico

A componente semântica pode ser muito útil no processo de resolução de anáforas, especialmente nas situações em que a análise morfológica, lexical e sintáctica não permite chegar directamente à solução. Em (2.18) e um (2.19) ilustra-se esse cenário.

(2.18) As irmãs comeram as uvas. Elas eram deliciosas.

(2.19) As irmãs comeram as uvas. Elas ficaram deliciadas.

No exemplo anterior, pode verificar-se que existe concordância em género e número, e sintácticamente tanto *as irmãs*, como *as uvas* são candidatos a antecedente de *Elas*. Ao associar-se informação semântica, através de um dicionário ou ontologia, poder-se-ia concluir que *deliciosas* caracteriza *as uvas* e *ficar deliciado* é próprio da entidade, aqui representada por *as irmãs*.

[10]pela regra descrita acima *Ele* não pode ser identificado com *Diogo*

2.3.1.4 Conhecimento de Discurso

Os vários tipos de conhecimento apresentados até aqui servem principalmente como forma de descartar hipóteses impossíveis e não como método de escolha preferencial de algum candidato. Entende-se que sempre que existe mais do que uma alternativa para solução de uma anáfora, e que nenhuma outra estratégia a resolveu, escolhe-se a hipótese que tem o candidato mais saliente, a quem, em linguística computacional chamamos **foco**. O conhecimento de discurso, é um conhecimento que nos permite determinar este foco e preferir a hipótese que o tem como candidato, em detrimento das restantes. Esta noção é baseada na definição de coesão, previamente introduzida e necessita de mecanismos capazes de determinar os vários focos de um discurso.

2.3.1.5 Conhecimento Mundo Real (senso comum)

O exemplo (2.16) demonstra claramente a necessidade deste tipo de conhecimento para conseguir resolver com sucesso a anáfora nele representada. Outros casos como (2.20) e (2.21) necessitam do mesmo tipo de informação.

(2.20) O Diogo vende batatas, ao sr. João. Ele vendeu tudo.

(2.21) O Diogo vende batatas ao sr. João. Ele comprou tudo.

A resolução de anáforas que depende deste conhecimento, tem a menor probabilidade de ser bem sucedida, dada a escassez, a dificuldade de produzir, e o domínio incrivelmente extenso de representações como as seguintes:

Se X vende a Y e se Z (Z é X ou Y) vende, é mais provável que Z seja X.

Se X vende a Y e se Z (Z é X ou Y) compra, é mais provável que Z seja Y.

Os vários conhecimentos vistos até aqui, podem estar embebidos no processo de resolução de anáforas. Estes processos englobam obviamente, regras de selecção, baseadas em diferentes conhecimentos,e permitem fazer a melhor escolha do antecedente, entre os vários candidatos. Estas regras, ou preferências, são geralmente referidas como **factores de resolução de anáforas**. Alguns deles são apresentados de seguida. As três primeiras representam restrições e as restantes, factores ppreferenciais

- A concordância entre género e número

  Este factor exige concordância entre o género e o número do pronome e seu antecedente. Na maioria dos casos, basta que a concordância se verifique, entre o pronome e o *header* do sintagma nominal.

- Restrições semânticas

  Esta exigência, indica que as restrições aplicadas às anáforas também devem ser aplicadas aos seus antecedentes. Considere-se o seguinte exemplo.

  (2.22) *O joão retirou o CD, do* ***leitor*** *e desligou-****o***.

  (2.23) *O João retirou o* ***CD****, do leitor e copiou-****o***.

  Neste exemplo, na frase 2.22 o antecedente de *o* tem que ser um objecto que se possa desligar, e por isso, o pronome é resolvido com *leitor*, em 2.23, o referente tem que estar associado a *CD*, pois é o único objecto, ali presente, que pode ser copiado.

- Restrições *c-command*

  Pensando na estrutura sintáctica de uma frase como uma árvore, considerando dois nós, que representam duas palavras, aí presentes diz-se que:

  Um nó A c-command um nó B se e só se se verificam três condições:

  i) A não domina B;

  ii) B não domina A;

iii) O primeiro ramo a dominar A também domina B

Esta restrição exige que só pode haver correspondência entre um pronome e um antecedente, se este não c-command o pronome, e ambos não estão,no mesmo domínio local. [Mitkov(2002]. É esta regra, que impede no exemplo, 1.23 que *Joana* seja antecedente de *ela*.

(1.23)*A Joana foi com a* ***Isabel*** *à praia. A Joana saiu com* ***ela****, porque tinha saudades.*

- Paralelismos sintáctico e semântico

  Factor de preferência que privilegia os sintagmas nominais com a mesma função sintáctica e semântica que a anáfora. Pode ser útil em caso de empate, no entanto facilmente se encontram exemplos em que os elementos anáfora e antecedente não partilham estas funções.

- Saliência

  Ao aplicar este factor, dá-se preferência ao elemento mais saliente, pois é ele o mais susceptível de ser pronominalizado. É esta observação que atribui a este factor, grande importância, no âmbito de resolução de anáforas pronominais. Entende-se que num texto coerente, o elemento mais saliente, é o foco do discurso, e é a ele, que as anáforas em frases seguintes, geralmente se referem.

- Proximidade

  O factores de proximidade, como o próprio nome indica, preferem o candidato mais próximo à ocorrência do pronome. Pode ser uma preferência aplicada, favoravelmente, em algoritmos que prevêem que os antecedentes ocorram na mesma frase que a anáfora.

Estes factores, aqui descritos, podem funcionar como filtro, descartando hipóteses impossíveis ou como factores preferenciais, atribuindo valores a cada hipótese para que se possa escolher a preferida.

2.3.2 Diferentes Abordagens

Qualquer que seja a abordagem, esta deve dividir-se em etapas. Primeiro, a identificação das anáforas, de seguida identificação dos candidatos a antecedentes para cada uma delas e por último, de cada grupo de candidatos, eleger o antecedente correcto. Seguem-se algumas abordagens, já aqui referidas.

2.3.2.1 Lappin And Leass

Lappin e Leass propuseram um algoritmo para resolução de anáforas pronominais, na língua Inglesa, a que chamaram RAP[11]. Tanto o algoritmo como o analisador sintáctico McCord's, usados, estão implementados em Prolog. O algoritmo tem como base um sistema de medida de saliência, baseado na estrutura sintáctica de cada frase, e uma representação simples do modelo de discurso. De uma forma simples pode-se dividir este algoritmo em 7 componentes:

- Um filtro sintáctico, que actua dentro de uma frase, descartando dependências sintácticas entre pronomes e sintagmas nominais[Lappin & McCord(1990b].
- Um filtro morfológico a fim descartar hipóteses que não coincidam em pessoa, género e número.
- Um procedimento para encontrar pronomes pleonásticos[12].Este procedimento é dedicado unicamente às ocorrências do pronome *it*.
- Um algoritmo de ligação para identificar candidatos a antecedente de pronomes reflexivos[13], pode ver-se com mais detalhe em [Lappin & McCord(1990a].
- Um procedimento para atribuir factores de saliência aos sintagmas nominais, por paralelismos sintáctico, de sujeito, entre outros.

[11] Resolution of Anaphora Procedure
[12] pronomes semanticamente vazios
[13] representam a mesma pessoa que o sujeito

- Um procedimento para identificar sintagmas nominais com ligações entre si, que formam uma classe de equivalência, que terá como valor de saliência a soma dos valores de cada elemento que constitui essa mesma classe.

- Um procedimento de decisão, que de entre os candidatos escolhe o "melhor classificado"como antecedente. Em caso de empate prefere o candidato mais próximo da anáfora.

O algoritmo de ligação, associa os candidatos que ocorrem na mesma frase que o pronome, aos antecedentes de pronomes reflexivos. Por sua vez, o filtro sintáctico descarta os candidatos dentro da mesma frase, como antecedentes de pronomes na terceira pessoa não reflexivos, são as imposições sintácticas que eliminam esta possibilidade de co-referência. O filtro morfológico elimina incompatiblidades de pessoa, género e número. Aos restantes candidatos é aplicado o procedimento de saliência, que atribui a cada candidato uma pontuação. O procedimento de decisão, é o último a ser aplicado.

Resta perceber como é atribuida a pontuação de acordo com o grau de saliência. Para cada candidato, isto é, cada sintagma nominal não eliminado pelos filtros previamente aplicados, é atribuido um valor de saliência de acordo com o seu papel gramatical. Esse valor resulta da soma de todos os factores de saliência que se podem aplicar a um sintagma nominal, é o que ilustra a tabela 2.1.

Esta é uma abordagem que pode ser vista com mais detalhe em [Lappin & Leass(1994] e que na sua versão original, os testes apresentados pelos autores têm uma taxa de sucesso de 86%.

Thiago Coelho [Coelho(2005], adaptou este algoritmo à língua portuguesa, passando a usar o analisador sintáctico PALAVRAS[Bick(2000]. Susbtituiu o filtrosintácticoo

Tabela 2.1: Valores dos factores de saliência [Lappin & Leass(1994]

Factores de Saliência	Valores	Exemplo
Frase actual	100	**A Ana** joga futebol, ela sempre gostou de desporto.
Sujeito	80	**O João** casou ontem.
Construção existencial	70	Havia **uma caixa de chocolates** em cima da mesa.
Objecto directo	50	Joana vai pagar **a conta** do seu jantar de anos.
Objecto indirecto	40	Capuchinho vermelho leva uma cesta de fruta **à sua avó.**
Ênfase não adverbial	50	Não jantámos ao ar livre por causa **da chuva**
Sintagma nominal não contido	80	**Ana** comprou **um livro** de que ouvira falar.
Paralelismo sintáctico	35	**A Vera** é gorda. Ela anda a fazer dieta.
Catáfora	-175	Ela comprou mais do que devia, **a Vera** não tem limites.

e o algoritmo de ligação pelas restrições de co-referência de [Reinhart(1983], não implementou o procedimento para identificação de pronomes pleonásticos (apenas faz sentido quando aplicado em textos em Inglés) nem fez tratamento de catáforas. Quanto ao grau de saliência, foi usado um quadro idêntico à tabela 2.1. Os melhores resultados obtidos durante a avaliação, resultam do algoritmo aplicado a um corpus jurídico e têm uma taxa de sucesso de 43,56%.

2.3.2.2 Hobbs' Algorithm

Este algoritmo baseia-se principalmente no conhecimento sintáctico. O processo implica uma pesquisa numa árvore sintáctica, que ao terminar, obtém o sintagma nominal (NPi) mais provável de ser o antecedente. Esse resultado é dado pelo caminho (T) mais curto, que satisfaz as seguintes restrições:

- T é o caminho desde o NP que domina o pronome P, até ao primeiro NP ou S, que domina este NP.
- T contém um nó N, constituído por um NP ou S, que por sua vez, contém o NP que domina P.
- Nenhum nó contém Npi

Este processo é aplicado à árvore sintáctica da frase onde ocorre o pronome, caso não seja encontrado nenhum candidato que satisfaça estas condições o processo é repetido em frases anteriores, preferindo sempre as mais próximas. Dentro da mesma frase, a pesquisa começa no pronome, e sobe na árvore até à raiz da frase. Durante este percurso, sempre que se encontra um NP ou S, faz-se uma pesquisa em largura esquerda-direita, aos NPi, primeiro na sub-arvore à esquerda do caminho que leva até ao pronome (anáfora), e em seguida na sub-árvore à direita (catáfora).

Durante a execução é ainda garantida a concordância de género e número entre os candidatos e anáfora, pois a árvore sintáctica dispõe dessa informação morfológica. Pode-se ainda observar, que dada a forma como é aplicado o algoritmo, há preferência por soluções mais próximas do pronome e/ou na mesma frase.

Ao avaliar os resultados, Hobbs[14] obteve uma taxa de sucesso de 88% que melhorou para 92% ao adicionar umas simples restrições de selecção. Este algoritmo, já foi considerado inadequado, contudo o valor percentual dos seus resultados assim como a sua vantagem computacional, comparativamente a métodos de cariz semântico, fazem dele um algoritmo capaz de competir com estudos mais recentes.

2.3.2.3 Mitkov's *Robust, Knowledge-poor Algorithm*

Segundo Ruslan Mitkov, foi a necessidade, por parte dos sistemas de processamento de língua natural, de ter um algoritmo de resolução de anáforas robusto, que actuasse eficazmente em ambientes "knowledge-poor", que motivou o desenvolvimento desta abordagem aqui apresentada. O seu algoritmo de resolução de pronomes, dispensa conhecimentos complexos de sintaxe, semântica e análise de discurso. A estratégia baseia-se no uso de *indicadores de antecedente*, que serão explicados mais adiante, em síntese, o seu algoritmo após um pré-processamento, evolui em três etapas.

Em primeiro lugar, examina a frase onde ocorre o pronome, e as duas anteriores, se houver, anteriores a esta, o ponto de partida é a esquerda desse pronome. O objectivo é

[14][Hobbs(1978]

encontrar os vários sintagmas nominais que ocorrem neste domínio. A segunda etapa, passa por seleccionar, os sintagmas nominais que concordam em género e número com o pronome, formando um grupo de candidatos. Por último, aplica-se a cada um dos candidatos os indicadores de antecedente[15], atribuindo a cada um deles uma pontuação, que no final permitirá escolher o candidato melhor classificado, como antecedente.

Relativamente ao pré-processamento, é usado um ***POS tagger*** para a divisão das frases, e regras gramaticais, para encontrar os vários sintagmas nominais durante a pesquisa dos mesmos. Resta esclarecer o que são os indicadores de antecedente. Quando se diz que, a cada antecedente se aplicam os vários indicadores, é como se, cada antecedente passasse por uma série de "testes", que lhe atribuem pontuação, essa pontuação pode ser positiva ou negativa, conforme se verifica uma característica que favorece um candidato ou o desfavorece na hipótese de ser o antecedente. Os indicadores têm um carácter preferencial e não devem funcionar como regras obrigatórias para descartar hipóteses inválidas. Os vários indicadores são[16]:

- ***First noun phrases*** O primeiro sintagma nominal (NP) da frase tem pontuação +1.
- ***Indicating verbs*** É atribuido +1 ponto aos NP's que ocorrem imediatamente depois, dos verbos que constituem um grupo previamente definido. Por exemplo, os verbos, analisar, considerar, cobrir, entre outros, fazem parte desse grupo.
- ***Lexical reiteration*** Aos pronomes que aparecem repetidos duas ou mais vezes, num parágrafo, atribui-se +2 pontos, aos que apenas se repetem uma vez atribui-se +1 ponto.
- ***Section heading preference*** A pontuação +1 é atribuida aos NP's que aparecem no cabeçalho da secção onde ocorre o pronome.
- ***Collocation match*** +2 é a pontuação dada aos NP's que têm um padrão de colocação

[15] antecedent indicators [Mitkov(2002]
[16] Serão mantidas as nomenclaturas originais para evitar imprecisão na tradução.

idêntico ao do pronome, por exemplo a sequência nome/pronome, verbo ou verbo, nome/pronome, etc.

- ***Immediate reference*** Aos NP's que ocorrem numa construção do tipo "...(You) V1 NP ... *con* (you) V2 it (con (you) V3 it)"onde *con* é um dos elementos (and/or/before/after/util....), atribui-se +2.

- ***Sequential instructions*** Os NP's que ocorrem na posição do Np1,numa construção do tipo "...To V1 Np1, V2 Np2. (Sentence). To V3 it, V4 NP4". são os antecedentes mais prováveis do pronome it.

- ***Term preference*** Atribui-se +1 aos sintagmas nominais, identificados como termos representantes do género do texto.

- ***Indefiniteness*** Aos sintagmas nominais indefinidos atribui-se -1.

- ***Prepositional noun phrases*** Os NP's que ocorrem em frases preposicionais têm pontuação -1.

- ***Referencial distance*** Este indicador pode atribuir pontuação positiva ou negativa, dependendo da localização do NP relativamente ao pronome. Isto é, um Np na oração anterior à do pronome, mas ainda na mesma frase tem pontuação +2, se estiver na frase anterior, tem pontuação +1 , na segunda frase acima tem +0 e em qualquer outra, tem pontuação -1.

Após a execução das três fases, descritas mais acima, pode acontecer que não tenha resultado apenas um candidato (a solução), mas sim alternativas com a mesma pontuação. Nesse caso deve-se preferir o candidato com melhor pontuação de acordo com os seguintes indicadores apresentados na ordem decrescente. Assim que se destacar um, a solução está encontrada. A ordem será: *immediate reference*, *collocational pattern*, *indicating verbs* e finalmente *candidato mais recente.*

Encontrados os resultados, são feitas as avaliações e de acordo com os estudos apresentados o algoritmo knowledge-poor de Mitkov tem uma taxa de sucesso de 89.7%. Esta

taxa é conseguida com a particularidade, que após o pré-processamento, o input do algoritmo, foi manualmente "corrigido"a fim de garantir a inexistência de erros. Um exemplo detalhado da execução da abordagem de Mitkov's, assim como qualquer outro esclarecimento adicional, pode ser consultado em [Mitkov(2002].

Mais recentemente este algoritmo foi melhorado e reimplementado no que resultou uma versão automática denominada MARS. Comparativamente com o algoritmo original as grandes diferenças são a introdução de mais três indicadores de antecedente, e algumas alterações na implementação resultantes das novas ferramentas de pré-processamento, então disponíveis. Passa-se a usar um FDG[17]-parser e o seu output é automaticamente o input do MARS, sem qualquer edição manual. Quanto às alterações de implementação não se entrará em detalhe, e por isso segue-se a enumeração dos três novos indicadores:

- ***Boost prounoun*** Tal como os sintagmas nominais, os pronomes também passam a ser candidatos de outros pronomes.
- ***Syntactic Parallelism*** NP com a mesma função sintáctica do pronome tem pontuação +1.
- ***Frequent candidates*** A pontuação +1 é atribuída a NP que ocorrem mais frequentemente como candidatos de todos os pronomes no texto.

Os resultados desta nova versão são mais baixos que os da versão original 59,35%, o que é justificado pela aplicação do algoritmo, ao output não editado, que resulta do pré-processamento. Os detalhes podem ser consultados em[Mitkov(2002].

2.3.2.4 *Centering*

A teoria do *Centering* surge primeiro por mão de [B.J. Grosz(1986] e é depois extendida por [Brennan *et al.*(1987]. Estes últimos autores, propõem um conjunto de regras e restrições que representam a estrutura do discurso, para escolher um antecedente de entre

[17]Functional Dependency Grammar

um conjunto de candidatos já filtrados morfologicamente. Fazem uso, principalmente, de duas estruturas fundamentais o **forward-looking center** e o **backward-looking center**, que representam as entidades do discurso e o foco/assunto em questão num determinado momento. A teoria é baseada na noção de coerência de discurso, este é tanto mais coerente, quanto menor for o "grau"de inferência necessário para resolver um pronome. O *centering* prevê que num discurso coerente, os seus intervenientes vão preferir **continuar** a falar/escreve sobre a entidade em foco, e que esta, é a mais provável de sofrer pronominalização nesse segmento do discurso. Esta teoria será descrita em detalhe na secção três deste trabalho, uma vez que é ela a base de implementação desenvolvida. Aí, estas e outras definições serão detalhadamente apresentadas.

2.3.3 Evolução dos Processos de Resolução de Anáforas

A temática, **resolução de anáforas** foi alvo de muito estudo nos anos 80, passou depois, por um período menos activo e voltou novamente, ao longo destes últimos dez anos, a ser alvo de interesse. As primeiras abordagens, [Carbonell & Brown(1988], [Rich & LuperFoy(1988], [Sidner(1979] baseavam-se no conhecimento da linguagem, o que era difícil de representar e de processar. Face a este cenário, e perante a necessidade de criar sistemas de processamento de língua natural usáveis, surgiram estratégias knowledge-poor, como é o caso da abordagem de [Baldwin(1997] ou mesmo do algoritmo de Mitkov, já aqui apresentado. O emergir de novos analisadores sintácticos, *POS taggers* para manuseamento de novos Corpus, e o desenvolvimento de léxicos semânticos como *WordNet*, foram fundamentais para o aparecimento destes estudos. É facto, a necessidade de pré-processamento, por mais reduzido que seja, assim como são necessários Corpus anotados para aplicar os vários algoritmos e proceder à avaliação dos mesmos. Actualmente, o pré-processamento, influência negativamente os resultados obtidos. Viu-se, por exemplo na abordagem de Mitkov, a taxa de sucesso a baixar incrivelmente quando o algoritmo foi testado sem verificação manual do input. Neste momento, esta é uma área em desenvolvimento que depende doutras que a influenciam directamente.

"The accuracy of today's pre-processing is still unsatisfactory from the point of view of anaphora resolution."[18]

Nos últimos anos registou-se um aumento significativo de estudos nesta área, uns na tentativa de minimizar esta dependência, procurando alternativas ao pré-processamento, por exemplo recorrendo à *Web* [Market *et al.*(2003], outros tentando outras abordagens. Dois exemplos são dados por [Soon *et al.*(2001] e por [Kheler(1997]. O primeiro é na área de *Machine Learning* onde o método de escolha é baseado em árvores de decisão, o segundo usa um modelo de *Entropia Máxima* para atribuir valores de probabilidade às relações de co-referência.

Os sistemas de processamento de língua natural, têm exigido novos desenvolvimentos. Hoje em dia, os sistemas de resolução de anáforas sejam elas pronominais ou não, são usados em muitos domínios, por exemplo, em interfaces de língua natural, geração automática de resumos, extracção de informação e sistemas de tradução automática.

A preocupação com as formas de avaliação dos sistemas desenvolvidos deve ser igualmente objecto de estudo. Nota-se uma falta de distinção entre a avaliação ao algoritmo de resolução de anáforas e ao sistema da resolução de anáforas [Mitkov(2000]. Esta distinção é extremamente importante e precisa de ser clarificada, para que se perceba, por exemplo, onde residem as dificuldades, se no algoritmo em si, se no pré-processamento, ou se na anotação manual (por isso susceptível de erros), dos Corpus.

Comparativamente com estudos para a língua inglesa, francesa ou espanhola, os estudos para o português são muito escassos, alguns exemplos do que tem sido feito são [de Abreu(2005] [Coelho(2005] e [Paraboni & de Lima(1998], onde apenas um deles se dedica a anáforas pronominais. É clara, a necessidade de evoluir nesta área especificamente para a língua portuguesa, um exemplo disso é a participação no CLEF[19] onde são testados sistemas de extracção de informação para línguas europeias, inclusive o português. É neste âmbito que surge este trabalho que tem como base a metodologia do

[18]Do ponto de vista da resolução de anáforas, a actual precisão do pre-processamento continua insatisfatória.[Mitkov(2001]

[19]Cross Language Evaluation Forum; `http://www.clef-campaign.org`

Centering, por esta se basear principalmente, no princípio de coesão do discurso, já aqui falado.

3 *Centering*

3.1 Origem e Definição

A metodologia de *Centering* aqui apresentada foi introduzida por [Brennan *et al.*(1987] em *A Centering Approach to Pronouns*, e resulta de uma extensão à já apresentada em [B.J. Grosz(1986]. Esta última descreve, que do processo de centrar atenção nas entidades do discurso, resultam três estados de transição entre frases, *continuing, retaining* e *shifting* . A proposta de [Brennan *et al.*(1987] é uma extensão a estes estados, que permite contemplar casos ambíguos que escapavam à primeira abordagem. O principal objectivo está em encontrar para cada pronome num discurso o seu antecedente de entre um conjunto de possíveis candidatos, após aplicação de um filtro morfo-sintáctico convencional.

A teoria do *Centering* é uma teoria de coerência e saliência [Poesio *et al.*(2000]. Coerência porque se baseia na noção de discurso coerente ,dada pela forma como as entidades são introduzidas num discurso e se relacionam entre si. Saliência, porque tem igualmente como base, a ideia de que há uma entidade mais saliente num dado momento do discurso. Pode ser vista como um conjunto de regras e restrições que regem a relação entre a entidade central do discurso, e as escolhas linguísticas dos seus intervenientes. De acordo com o *Centering* há entidades mais centrais que outras, e o uso de referentes, particularmente de pronomes, está condicionado por este grau de saliência.

A aplicação do *Centering* aos pronomes é natural, deriva da noção de saliência e do propósito, desta metodologia, em determinar o tema de assunto (ou foco) do discurso. As entidades que representam o foco, são as mais usuais de sofrer pronominalização [J.K. Gundel(1993], determiná-las é vantajoso, por exemplo, para geradores de língua natural.

De acordo com esta metodologia um **discurso** é constituído por **segmentos de discurso** contínuos, um segmento de discurso **D** é, por sua vez, constituído por ***Utterances*** $\mathbf{U}_1, \mathbf{U}_3 ... \mathbf{U}_n$[20]. Cada *utterance* **U** tem associado um conjunto de entidades possíveis de ser

[20]de grosso modo pode ser entendido como uma oração finita ou frase.

o próximo centro, a que se chama **Cf(U)**- ***Forward-looking centers***- que é constituído por todos os sintagmas nominais referidos nessa *utterance*.

Em cada *uterrance*, excepto a primeira, existe um único elemento central denominado ***backward-looking center*, Cb(U)**. Este elemento corresponde a uma das entidades presente no Cf da *uterrance* anterior e referida na actual, ou *nill*. Portanto faz a ponte para entidades previamente introduzidas no texto, é o foco do discurso no momento dado pela *utterance* U .

A estrutura CF está ordenada de acordo com um *ranking* específico, isto é, está ordenada por grau de saliência dos seus elementos. O primeiro é denominado ***Preferred Center*, Cp(U)** e representa o elemento mais provável de ser referido na próxima *utterance*, ou seja é preferencialmente o próximo Cb.

Existem quatro tipos de transição entre utterances. A sua tipologia é baseada em dois factores:

- o centro Cb, é ou não, igual para a *utterance* U_n e para a anterior, U_{n-1}.
- na mesma *utterance* o Cb é ou não igual, ao Cp.

Os estados de transição são:

1. **Continuing**

   Acontece quando $Cb(U_{n-1}) = Cb(U_n) = Cp(U_n)$. O foco da *utterance* anterior mantém-se na *utterance* U_n e é, o elemento mais provável de ser foco na próxima.

2. **Retaining**

   Acontece quando $Cb(U_{n-1}) = Cb(U_n) \neq Cp(U_n)$. Mantém-se o foco na transição da U_{n-1} para U_n, e nesta última *utterance*, o centro deixa de ser o elemento melhor "classificado"para ser foco da próxima.

3. **Shifting-1**

   Acontece quando $Cb(U_{n-1}) \neq Cb(U_n) = Cp(U_n)$. O centro alterou-se de U_{n-1} para

U_n, e o novo centro é o elemento mais provável de ser foco na próxima *utterance*. É este estado faz parte da extensão feita por [Brennan *et al.*(1987], à primeira versão que não o contemplava.

4. **Shifting**

   Acontece quando $Cb(U_{n-1}) \neq Cb(U_n) \neq Cp(U_n)$. O centro alterou-se de U_{n-1} para U_n, e o foco actual não corresponde ao elemento mais provável de ser o próximo foco.

Estas transições aqui apresentados demonstram como as *utterances* se devem relacionar num discurso coerente. Este é um discurso em que os intervenientes centram-se num foco, falam/escrevem sobre ele e mantêm-no, a fim de fazer todas as "observações"sobre o mesmo (continuing), antes de introduzirem novas entidades no discurso (retaining), só depois fazem um *shift* alterando o foco, o elemento central ou assunto, para uma destas novas entidades.

Como ilustração considerem-se os seguintes exemplos que diferem apenas na última frase:

Discurso A:

(3.1) O João gosta de jogar às cartas.

(3.2) Ele joga com a Sofia.

(3.3) O joão ensinou-a a jogar.

(3.4) Ele é um óptimo professor.

Discurso B:

(3.1) O João gosta de jogar às cartas.

(3.2) Ele joga com a Sofia.

(3.3) O joão ensinou a Sofia a jogar.

(3.5) A Sofia conhece-o há muitos anos.

Relativamente ao Cf, Cb e Cp têm-se:

(3.1) Cf= { João, cartas } , Cb= *nill* [21] e Cp= João.

(3.2) Cf = { João, Sofia }, Cb=João, Cp = João.

(3.3) Cf = { João, Sofia }, Cb=João, Cp= João.

(3.4) Cf={João }, Cb="João",Cp="João"

(3.5) Cf={ Sofia, João }, Cb=Sofia, Cp=Sofia.

De acordo com as classificações já aqui apresentadas, no discurso A, a transição de (3.3) para (3.4) corresponde ao estado *Continuing* e no discurso B, de (3.3) para (3.5) observa-se o estado *Shifting-1*. Imagine-se que ao discurso B, se acrescentava a frase (3.4), neste caso, comparativamente com o discurso A, este seria menos coerente, manteria o foco nas três primeiras frases faziam um *shift* na quarta, para depois voltar a falar do João, o primeiro foco.

As exigências desta teoria são feitas sobre a forma de restrições e regras sobre o Cf e sobre o Cb.

Restrições

- Para cada *utterance* existe apenas um Cb.
- Qualquer elemento em $Cf(U_n)$ é constituinte de U_n.
- O $Cb(U_n)$ é o elemento melhor classificado de $Cf(U_{n-1})$ mencionado em U_n.

Regras

- Se algum elemento no $Cf(U_n)$ está pronominalizado[22], então o $Cb(U_n)$ também o será.
- Após classificação a escolha da solução deve preferir umas hipóteses em detrimento de outras. A preferência, é aqui apresentada por ordem decrescente.

[21] A primeira frase de um discurso nunca tem Cb, o que se compreende, pois o Cb é a ligação com as frase anteriores.

[22] aparece sob forma de pronome

Continuing > *Retaining* > *Shifting-1* > *Shifting*

A primeira restrição e a segunda regra reforçam o princípio decoerênciaa do discurso fundamental para esta teoria.

Particularmente, as extensões do algoritmo de [Brennan *et al.*(1987] são a adição de um novo estado de transição *shifting-1* já indicado, e adição de restrições de contra-indexação. Diz-se que o elemento A é contra-indexado com o pronome B se, A *c-command* B. Este filtro de contra-indexação, garante que um pronome nunca se refere a um sintagma nominal que o *c-command* no mesmo domínio local. A noção de *c-command* foi introduzida no final da secção 2.3.1 e pode ser melhor esclarecida numa secção mais adiante (3.2) ou em [Mitkov(2002]. De uma forma muito resumida, o algoritmo destes autores está divido em três fases. Na primeira fase, geram todas as combinações Cf-Cb possíveis. Na segunda etapa, filtram as diferentes opções, aplicam três filtros distintos, a cada uma das opções previamente geradas, descartam hipóteses impossíveis e deixam as restantes. Por último classificam e escolham a melhor hipótese como solução. A segunda etapa é sem dúvida, a mais interessante,o primeiro filtro verifica a contra-indexação, se o mesmo antecedente para dois pronomes contra-indexados, ou se o pronome e o seu possível antecedente, estão contra-indexados, descarta-se essa alternativa. Os restantes dois filtros, confirmam a 3ª restrição e a 1ª regra.

Anos mais tarde, ao aparecimento desta extensão, [Walker(1989] concluiu, num estudo dedicado a avaliação e comparação do *Centering* com o algoritmo de Hobbs, onde exigiu dados[23] num "ambiente ideal"livres de erros de pré-processamento, que a *performance* dos dois algoritmos, naquele conjunto de dados específicos, não tinha diferenças significativas, como ilustra a tabela 3.1. Descobriu ainda, que as situações em que o algoritmo de Hobbs acertava e o de *Centering* não, estavam relacionadas com a preferência que Hobbs dada às hipóteses, dentro da mesma frase onde ocorre o pronome. Daqui surge uma tentativa de melhoramento do *Centering* em que ela propõe a preferência das hipóteses na mesma frase, apenas nas situações em que não se encontrou um Cb, ou o

[23] oriundos do primeiro capítulo do romance *Wheels* de Arthur Hailey's e de uma edição de 7 de Julho de 1975 do jornal *Newsweek*

mesmo foi descartado.

Tabela 3.1: Anáforas resolvidas com sucesso

Texto Fonte	Total de Anáforas	[Brennan *et al.*(1987]	Hobbs
Wheels	100	90	88
Newsweek	100	79	99
Diálogos	81	49	51

A aproximação de Brennan, Friedman e Pollard é largamente referida na documentação encontrada no âmbito, resolução de anáforas. Tem sido usada para comparações entre algoritmos assim como base a outros estudos.

3.2 Derivações do *Centering*

Centering é uma metodologia que ambiciona conhecer, ao longo do texto, o foco do mesmo, e resolver os vários pronomes com a entidade central, ou foco, aquando a ocorrência de cada um. Como qualquer outra metodologia tem as suas limitações, o que impulsiona novos estudos baseados neste. Na secção anterior, onde se apresentou a extensão de [Brennan *et al.*(1987] foram propositadamente usadas as palavras *utterance*, assim como *ranking* sem tradução. Isto porque, nas fontes consultadas para compreensão desta metodologia, estas definições não foram especificadas, provavelmente com intuito de reservar essa decisão para a fase de implementação, em que se adequaria as suas definições, à língua para a qual a implementação se desenvolve. No fundo o principio base é comum às diferentes línguas. Esse facto é também relatado por [Poesio *et al.*(2000] um estudo recente que se propõe a verificar a validade das exigências de que *para cada utterance apenas existe um Cb*, e *se algum elemento de Cf é pronominalizado então o Cb, também será*. Os resultados apresentados por estes investigadores, reportam que ambas as exigências estão sujeitas a frequentes violações, concluem depois de algumas variações ao algoritmo, que um texto pode ser coerente mesmo que viole alguma destas exigências, uma vez que existem outros mecanismos para alcançar a coerência, tais como relações

retóricas.

Outros estudos, propõem variações/extensões à metodologia *centering*, as próximas referências são exemplo disso.

A versão original do *centering* defende que a procura dos antecedentes, deve ser limitada às *utterances* do segmento do discurso, onde ocorre a anáfora, em particular a versão [B.J. Grosz(1986] entendia que a procura devia ser feita apenas na própria frase e na anterior. O estudo apresentado em [Walker(1997] defende o desaparecimento dessa limitação e favorece a interacção do *centering* com a estrutura global do discurso, propondo o uso de um *attentional state model* denominado *cache model*.

Por sua vez [Strube(1998] propôs uma alternativa, em que o Cb e os estados de transição são substituídos por uma *S-list* constituída pelas entidades do discurso, ordenadas por saliência. O critério de ordenação é baseado, principalmente, na distinção entre entidades *hearer-old*[24] e *hearer-new*[25].

[Kibble(2001] propõe uma reformulação da segunda regra da metodologia *Centering*. Na original, a preferência de escolha entre transições, resultava dos princípios de coesão (manter o mesmo foco) e saliência (considerar o foco o sujeito), a opção aqui apresentada mantém estes dois princípios e adiciona a noção de *cheapness*[26] introduzida por [Strube & Hahn(1999]][27]. O autor defende que esta reformulação, permite tratar situações de alteração de foco, que violavam a ordem de preferência dada pela teoria do *Centering*.

Outras referências a estudos baseados nas noções centro/foco podem ser consultadas em [Dahl & Ball(1990], [Stys & Zemke(1995] ou [Tetreault(1999],entre outros.

[24] já conhecidas no discurso

[25] novas no discurso

[26] *menor custo*

[27] *Cheapness*: A transição entre duas utterances, tem *menor custo* se o Cb da *utterance* actual é correctamente previsto pelo Cp da utterance anterior. Esse tipo de transição deve ser preferido às restantes.

4 Abordagem implementada: *Centering* para a Língua Portuguesa

4.1 Descrição

O estudo aqui apresentado, é dedicado à resolução de anáforas pronominais, particularmente às representadas pelo seguinte subconjunto de pronomes pessoais : { ele; ela; o; a; no; na; lo; la; lhe } e suas respectivas formas no plural. Foram escolhidos estes pronomes em particular, pois segundo uma análise a corpora, são estes que apresentam a tradicional "função pronome", isto é, a função anafórica por excelência, aquela que necessita do antecedente para interpretação do enunciado [Coelho(2004].

À semelhança de outros algoritmos, o *centering* necessita de um pré-processamento. A estrutura sintáctica das frases, é utilizada durante o processo de resolução de anáforas, sendo fundamental para as restrições morfossintácticas. O analisador sintáctico utilizado, para efectuar essa análise, foi o PALAVRAS, desenvolvido no VISL[28] que é um projecto de investigação e desenvolvimento do *Institute of Language and Communication* da *University of Southern Denmark*, o seu output pode ser visto com o exemplo (3.7) dado pela frase (3.6), o seu desenvolvimento está detalhado em [Bick(2000]. A formatação em árvore, do output, dificulta a consulta da análise sintáctica do texto, pelo que se usou uma ferramenta desenvolvida no Departamento de Informática da Universidade de Évora, o Xtractor, para transformar este output noutro mais amigável, e em formato Prolog, ideal para esta aplicação, também ela desenvolvida nesta linguagem. Após o pré-processamento, o input do algoritmo é constituído por um ficheiro, em que cada frase é representada por um facto com a sua respectiva análise sintáctica, sem perda de nenhuma informação dada previamente pelo analisador sintáctico, como ilustra o exemplo 4.3.

(4.1) O Diogo adora brincar, ele tem 8 anos.

[28] Visual Interactive Syntax Learning

Exemplo (4.2)

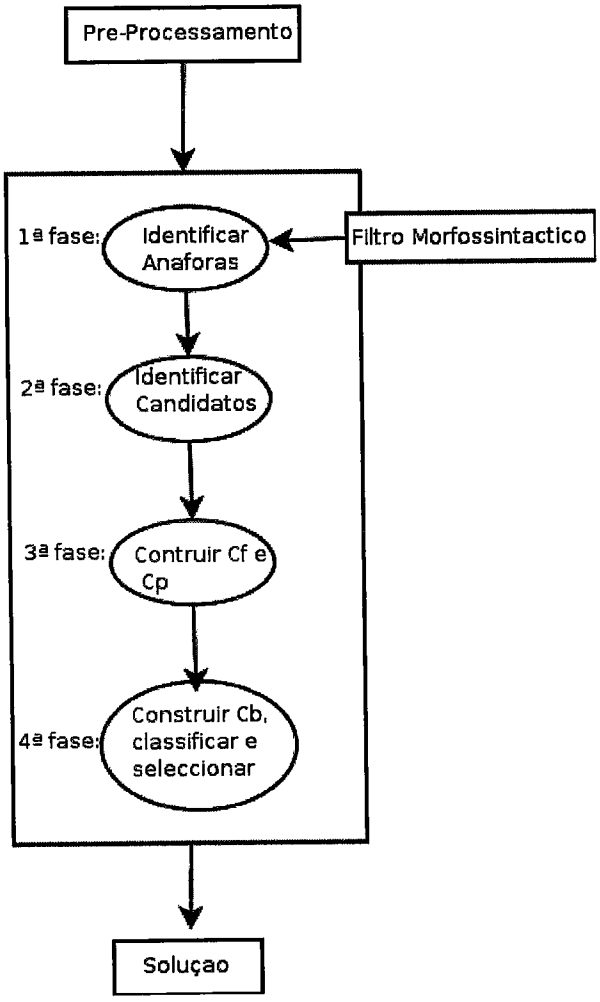
```
STA:cu
=CJT:fcl
==SUBJ:np
===>N:art('o' <artd> M S)        O
===H:prop('Diogo' <hum> M S)     Diogo
==P:v-fin('adorar' PR 3S IND)    adora
==ACC:v-inf('brincar' 0/1/3S)    brincar
=,
=CJT:fcl
==SUBJ:pron-pers('ele' M 3S NOM/PIV)    ele
==P:v-fin('ter' PR 3S IND)       tem
==ACC:np
===>N:num('8' <card> M/F P)      8
===H:n('ano' <per> <dur> M P)    anos
=.
```

Exemplo (4.3)

```
sentence(syn(
            sta(cu,
                cjt(fcl,
                    subj(np,n(art(o,'<artd>','M','S'),'O',1),
  h(prop('Diogo','<hum>','M','S'),'Diogo',2)),
                    p(v_fin(adorar,'PR','3S','IND'),adora,3),
                    acc(v_inf(brincar,'0/1/3S'),brincar,',',4)),
                cjt(fcl,
                    subj(pron_pers(ele,'M','3S','NOM/PIV'),ele,6),
                    p(v_fin(ter,'PR','3S','IND'),tem,7),
                    acc(np,n(num('8','<card>','M/F','P'),'8',8),
  h(n(ano,'<per>','<dur>','M','P'),anos,'.',9))))))).
```

Exceptuando o pré-processamento, a implementação do algoritmo, pode, de grosso modo, dividir-se em quatro fases, como ilustra a figura 4.1. Cada uma delas será de seguida, descrita em detalhe.

Figura 4.1: Esquema do Algoritmo



1ª Fase: Identificação das Anáforas

Naturalmente que a primeira fase do algoritmo é dedicada à identificação das anáforas. A identificação dos pronomes anafóricos, exige uma procura exaustiva em todas as frases de elementos identificados sintacticamente como pronomes pessoais, o exemplo 4.4 ilustra um desses elementos marcado a negrito, que representa o pronome pessoal "ela", onde se

poder verificar o seu género (feminino), número (singular) e pessoa (3ª).

(4.4) subj(**pron_pers(ela,'F','3S','NOM')** ,'Ela',7)

Além dos elementos já referidos, verifica-se que o pronome pessoal está englobado noutro facto, que permite saber a sua função gramatical, neste caso, sujeito (subj)[29] e a posição em que a palavra aparece no sexto (7ª palavra), fundamental para perceber que elementos se encontram antes e depois do pronome.

Paralelamente a esta pesquisa, procuram-se todos os sintagmas nominais representados em cada frase, uma das formas em que estes ocorrem é dada pelo exemplo (4.5). O intuito é facilitar a próxima etapa, evitando, mais uma procura exaustiva destas estruturas, nos factos que representam as frases dados pelo input do algoritmo.

(4.5) "A Isabel"

```
subj(np,n(art(o,'F','S','<artd>'),'A',12),
     h(prop('Isabel','F','S'),'Isabel',13))
```

Neste exemplo, percebe-se que a representação de um sintagma nominal (np), é mais complexa que de um pronome. Interessa principalmente o seu *"header"*[30], aqui representado pelo termo Prolog h, do qual se extrai a informação de género, número e pessoa do sintagma nominal. Associado a esta procura, existe um predicado que determina o número das palavras que constituem o sintagma, indicando a primeira e a última, neste caso 12-13.

Já na fase dos testes percebeu-se que havia nomes próprios externos a um sintagma nominal, extendeu-se o predicado em questão e também eles foram aqui seleccionados. No final desta fase, para cada frase que ocorre no texto, temos a sua identificação associada a uma lista dos sintagmas nominais, nomes próprios e/ou pronomes pessoais que nela ocorrem, como ilustra o exemplo (4.6).

[29] a descrição do que significa cada um destes elementos pode ser encontrada em: `http://visl.sdu.dk/visl/pt/symbolset-floresta.html`

[30] elemento central do sintagma nominal

(4.6) "A Isabel concorre com ela aos fins de semana."

```
st(3)-[np(12-13,'F','S','Isabel',subj),
       pp(16,'F','3S',p),
       np(18-19,'M','P',fim_de_semana,p)]
```

Cada sintagma nominal ou nome próprio é representado por um facto Prolog de termo np, com cinco argumentos, eles são: número das palavras que o constituem, género, pessoa/número, palavra dada pelo *header* e a função gramatical que representam. Por sua vez os pronomes, são representados por um facto Prolog de termo pp, com quatro argumentos, o número da palavra, género, pessoa/número e finalmente a função gramatical.

Feito isto, passa-se à identificação dos candidatos para cada um dos pronomes encontrados.

2ª Fase: Identificação dos Candidatos

Esta é uma fase muito simples, consiste em, para cada frase verificar se existem pronomes, se sim, para cada uma deles, comparar todos os sintagmas nominais anteriores ao pronome em questão, seja na própria frase ou até três frases atrás inclusive. Esta comparação consiste no filtro morfossintático, portanto verifica a concordância entre género, número e pessoa entre o pronome e os vários sintagmas nominais. Sempre que se verifica a concordância, é criado um novo candidato como o do exemplo (4.7) com a informação seguinte: a frase onde ocorre o pronome; o número da palavra do pronome; a função gramatical do pronome; o número da frase onde está o sintagma nominal candidato; o número das palavras que o constituem; o seu *header* e a função gramatical do sintagma.

(4.7) "A Isabel"é candidato do pronome 16.

```
cand(3,16,p,3,12-13,'Isabel',subj)
```

A restrição de procurar até três frases inclusive, é discutível. [Hobbs(1978] mostra, no estudo que fez, que 98% dos antecedentes estavam na mesma frase, ou na anterior, que

a do pronome. [A. McEnery(1997] num estudo baseado em 4681 anáforas, indica que em 85.64% dos casos o antecedente está entre as três frases anteriores e em 94.91% nas cinco frases anteriores. Outros estudos como os de [Ariel(1990] indicam que é mais comum o antecedente estar a uma distância considerável para anáforas demonstrativas do que para as pronominais. [Mitkov(2002] concluí que a maioria dos estudos, restrigem a procura às 2/3 frases. Como resultado destas leituras optou-se aqui também por restrigir a procura a três frases.

3ª Fase: Construção Cf e Cp

Neste momento o algoritmo além das várias frases com os seus constituintes, que interessam para a implementação[31], já representou, uma listagem de todos os candidatos para cada um dos pronomes. Pode então construir a *forward-looking-center*, ou abreviadamente Cf.

Já foi introduzido que o Cf de uma frase, é o conjunto de entidades[32], que aí ocorrem, em que cada pronome é substituído por um possível candidato. Como tal, o objectivo final desta fase é ter uma estrutura Cf que represente estas entidades. Para as frases onde não ocorrem pronomes há apenas uma hipótese para esta estrutura, nas outras situações, existem tantos Cf's quanto as combinações de candidatos possíveis. Isto é, imagine que numa frase ocorre um pronome, este tem três candidatos possíveis, logo, no final desta etapa, a frase em questão terá três hipóteses de Cf. Se existir mais do que um pronome, na mesma frase, as hipóteses crescem, por exemplo, numa frase com dois pronomes, cada qual com dois candidatos, há quatro Cf's possíveis. O exemplo (4.8) ilustra uma situação em que existe apenas um pronome com dois possíveis antecedentes.

[31] sintagmas nominais, nomes próprios e pronomes
[32] pronomes ou sintagmas nominais

(4.8) Para o pronome 16 na frase 3 há dois candidatos possíveis.

```
st(3)-[cf([np(12-13,'F','S','Isabel',subj),
   cand(3,16,p,1,1-2,'Ana',subj),
   np(18-19,'M','P',fim_de_semana,p)]),

      cf([np(12-13,'F','S','Isabel',subj),
   cand(3,16,p,3,12-13,'Isabel',subj),
   np(18-19,'M','P',fim_de_semana,p)])]
```

De uma forma simples, isto significa que há duas hipótese para a frase já apresentada em (4.6). Esta será equivalente a dizer "A Ana concorre com a Ana aos fins de semana", ou com "A Ana concorre com a Isabel aos fins de semana". A decisão é deixada para as fases seguintes.

Após construção das várias hipóteses do Cf, cada uma delas é organizada por saliência, a fim de possibilitar uma posterior identificação do Cp. Como se viu, tanto os sintagmas nominais como os pronomes tinham associada a sua função gramatical, a partir desses dados determina-se que elementos têm função de sujeito, de complemento directo, de complemento indirecto, de adjunto ou qualquer outra. A ordem pela qual, as funções foram enumeradas representa a ordem decrescente de preferência. Intuitivamente se percebe que o sujeito de uma frase é o seu elemento mais saliente, em termos de implementação, a sua identificação é directa, aliás já foi aqui exemplificada no exemplo (4.5) em que o termo é *subj*. As restantes podem não ser tão evidentes, contudo a consulta ao site `http://visl.sdu.dk/visl/pt/symbolset-floresta.html` pode elucidar neste domínio. O quadro seguinte mostra como identificar as funções dos vários elementos, nas "estruturas"que constituem o input.

Estando o Cf ordenado, por ordem de saliência, o seu primeiro elemento é naturalmente, o mais saliente e consequentemente o Cp.

Tabela 4.1: Correspondência entre factores de saliência e função gramatical dada pelo PALAVRAS

Factores de Saliência	Função G. PALAVRAS
Sujeito	subj
Objecto Directo	acc
Objecto Indirecto	dat
Adjuntos	{ advl; pred; pass; vok; top}

4ª Fase: Construir Cb, classificar e seleccionar

É claro, que o primeiro passo, ao analisar um Cf nesta fase, é determinar o Cb, só depois temos todos os dados para classificar e seleccionar. De acordo com [Brennan *et al.*(1987] o Cb ou é, o elemento mais saliente da frase anterior (Cp da frase anterior), ou *Nill*. Para garantir esta exigência, basta comparar o Cp anterior, com os elementos do Cf que se está a analisar. Se houver alguma coincidência seja por sintagma nominal ou por outro candidato, o Cb está encontrado, caso contrário é *Nill*. Após determinação do Cb, é necessário validar o Cf a que pertence, para garantir a primeira regra do *Centering*[33]. Caso o Cf tenha um ou mais pronomes, o Cb encontrado, deve ser equivalente a um dos candidatos aí presentes, se isso não se verificar o Cf em questão é descartado do leque de hipóteses. Na posse do $Cp(U_{n-1})$, do $Cp(U_{ni})$ e do $Cb(U_{ni})$, o algoritmo classifica a opção $Cf(U_{ni})$ não eliminada. Todas as hipóteses para a mesma frase são agora argumento de entrada, de um predicado que as escolhe, pela ordem de preferência, já apresentada, *Continuing* > *Retaining* > *Shifting-1* > *Shifting*. Em caso de empate não está estabelecido que opção tomar, pelo que é escolhido arbitrariamentee, o primeiro Cf da listagem.

Todo esta 4ª fase é aplicada na integra por frases. Portanto primeiro encontra-se uma solução para a primeira frase no texto, depois para a segunda, só depois para a terceira e assim sucessivamente. O objectivo é que a influência da solução da frase x, não seja negativa, para solução da frase seguinte. Considere-se por exemplo que a frase x tem como Cp um pronome. Ao determinar o Cb da frase x+1 a informação do Cp anterior é necessária, assim como é necessária, para classificar as hipóteses da frase x+1. Logo

[33] Se algum elemento no $Cf(U_n)$ está pronominalizado então o $Cb(U_n)$ também o será.

a solução da frase x tem que se encontrada antes de determinar o Cb da frase x+1, e de classificar as suas hipóteses.

4.2 Exemplo da aplicação do Algoritmo.

Considere-se o pequeno texto:

O joão abraçou a sua amiga Joana, nos jardins de sua casa.
Ele é muito amigo dela e do Rui.
Ele sempre foi fiel aos seus amigos, e o João tinha muitos.
Ele a Joana e o Rui são pessoas especiais.

Após a primeira fase, temos a identificação por frase, dos pronomes e sintagmas nominais que ocorre em cada uma, como ilustra o exemplo (4.9)

(4.9)

```
st(1).
[np(1-2,'M','S','João',subj),
       np(7-7,'F','S','Joana',n),
       np(4-7,'F','S',amiga,acc),
       np(10-14,'M','P',jardim,p)]
st(2).
[pp(16,'M','3S',subj),
       pp(21,'F','3S',p),
       np(24-25,'M','S','Rui',p)]
st(3).
[pp(27,'M','3S',subj),
       np(32-34,'M','P',amigo,p),
       np(37-38,'M','S','João',subj)]
st(4).
```

```
[pp(42,'M','3S',subj),
        np(43-44,'F','S','Joana',cjt),
        np(46-47,'M','S','Rui',cjt),
        np(49-50,'F','P',pessoa,sc)]
```

Com este resultado, procede-se à procura dos candidatos, e obtém-se a seguinte listagem:

(4.10)

```
cand(2,21,p,1,7-7,'Joana',n)

cand(2,21,p,1,4-7,amiga,acc)

cand(2,16,subj,1,1-2,'João',subj)

cand(3,27,subj,1,1-2,'João',subj)

cand(3,27,subj,2,24-25,'Rui',p)

cand(4,42,subj,2,24-25,'Rui',p)

cand(4,42,subj,3,37-38,'João',subj)
```

De notar, que os dois pronomes da segunda frase, embora antecedidos por ele, não têm como candidato, o sintagma nominal *"jardins de sua casa"* , uma vez que entre este último e os dois referentes, não há concordância de género e número.

Aplica-se em seguida a 3ª fase, da qual resulta o Cf organizado e respectivo Cp.

(4.11)

```
st(1).
[cf([np(1-2,'M','S','João',subj),
   np(7-7,'F','S','Joana',n),
   np(4-7,'F','S',amiga,acc),
   np(10-14,'M','P',jardim,p)],[cp(1-2,'João')])]

st(2).
[cf([cand(2,16,subj,1,1-2,'João',subj),
   cand(2,21,p,1,4-7,amiga,acc),np(24-25,'M','S','Rui',p)],
  [cp(1-2,'João')]),
       cf([cand(2,16,subj,1,1-2,'João',subj),
   cand(2,21,p,1,7-7,'Joana',n),
   np(24-25,'M','S','Rui',p)],[cp(1-2,'João')])]

st(3).
[cf([cand(3,27,subj,1,1-2,'João',subj),
   np(37-38,'M','S','João',subj),
   np(32-34,'M','P',amigo,p)],
  [cp(1-2,'João')]),
       cf([np(37-38,'M','S','João',subj),
   cand(3,27,subj,2,24-25,'Rui',p),
   np(32-34,'M','P',amigo,p)],
  [cp(37-38,'João')])]

st(4).
[cf([cand(4,42,subj,2,24-25,'Rui',p),
   np(43-44,'F','S','Joana',cjt),
   np(46-47,'M','S','Rui',cjt),
```

```
 np(49-50,'F','P',pessoa,sc)],
[cp(24-25,'Rui')]),
     cf([cand(4,42,subj,3,37-38,'João',subj),
 np(43-44,'F','S','Joana',cjt),
 np(46-47,'M','S','Rui',cjt),
 np(49-50,'F','P',pessoa,sc)],
[cp(37-38,'João')])]
```

Chama-se a atenção para a representação do Cf como um tuplo, em que o primeiro elemento, é uma lista com as entidades da frase, e o segundo é uma outra lista com o valor do Cp. Observa-se que para a frase quatro, o Cp varia, dependendo do Cf a que está associado, como seria de esperar pois ele depende desta "estrutura".

Na última etapa, determina-se o Cb, das hipóteses estas são classificadas e obtém-se o seguinte resultado:

(4.12)

```
st(1).
cf([np(1-2,'M','S','João',subj),
    np(7-7,'F','S','Joana',n),
    np(4-7,'F','S',amiga,acc),
    np(10-14,'M','P',jardim,p)],
   [cb(_4586,null)],
   [cp(1-2,'João')],_4580)

st(2).
cf([cand(2,16,subj,1,1-2,'João',subj),
    cand(2,21,p,1,7-7,'Joana',n),
    np(24-25,'M','S','Rui',p)],
   [cb(1-2,'João')],
```

```
    [cp(1-2,'João')],shifting_1)

st(3).
cf([cand(3,27,subj,1,1-2,'João',subj),
    np(37-38,'M','S','João',subj),
    np(32-34,'M','P',amigo,p)],
   [cb(37-38,'João')],
   [cp(1-2,'João')],continue)

st(4).
cf([cand(4,42,subj,3,37-38,'João',subj),
    np(43-44,'F','S','Joana',cjt),
    np(46-47,'M','S','Rui',cjt),
    np(49-50,'F','P',pessoa,sc)],
   [cb(37-38,'João')],
   [cp(37-38,'João')],continue)
```

Nesta última fase, realça-se o desaparecimento da hipótese que contemplava o candidato *Rui*, tanto na frase três como quatro. Para a terceira, embora o Cb neste caso tenha sido *João*, a exigência de que havendo pronomes, o Cb será um deles, não se verificava. Na quarta frase, o Cb para a hipótese descartada era *null*, que em "competição"com um Cb "real"nunca é escolhida. Destaca-se também, que para a frase um, não está atribuída classificação. Esta situação representa o caso em que apenas há um possível candidato, logo a solução está naturalmente encontrada, evita-se assim, processar as sub-etapas de classificação e selecção.

Está assim concluída a introdução à implementação deste algoritmo. Procedeu-se à respectiva avaliação e da sua observação em conjunto com os conhecimentos adquiridos, durante a recolha bibliográfica, resulta a implementação de algumas alterações e/ou extensões, na tentativa de melhorar o desempenho desta metodologia aplicada, à língua

Portuguesa. Em seguida são apresentados esses desenvolvimentos, os seus resultados constam na secção dedicada à avaliação.

4.3 *Centering* por orações

Até aqui, a implementação do algoritmo sempre considerou que uma *utterance* era equivalente a uma frase. Surge a questão : *Será esta interpretação correcta?* Em 3.2 tinha sido deixada a ideia, de que a definição de *utterance*, ficava ao critério de quem implementasse o algoritmo, com objectivo, de aproximar essa definição, à língua para a qual a aplicação se destinava. A extensão aqui apresentada considera que cada *utterance* representa uma oração. Isto é, a menor unidade do discurso cujos elementos essenciais são o sujeito e o predicado[34], a sua identificação a partir do output do analisador sintáctico PALAVRAS está indicada no quadro 4.2.

Tabela 4.2: Identificação de orações após aplicar o *parser* PALAVRAS

Oração	Função G. PALAVRAS
Oração finita	fcl
Oração infinita	icl
Oração adverbial	acl
Oração composta	cu

Esta alteração, à versão inicial, visa tirar proveito, principalmente de duas situações:

1. Os constituintes de uma frase estão organizados hierarquicamente, na forma de árvore. As orações numa frase podem ou não estar todas ao mesmo nível. Uma oração subordinada, por exemplo, está num nível inferior, à oração principal, como ilustram os exemplo (4.13) e (4.14), neste último observa-se que a segunda oração (segundo fcl) pertence ao termo **acc**, que estando ao mesmo nível da primeira oração, deixa a segunda oração, num nível inferior. Esta distinção hierárquica possibilita uma maior restrição aquando da escolha dos candidatos.

[34] `http://www.priberam.pt/dlpo/gramatica/gramatica.aspx`

2. Os vários sujeitos de uma frase (quando existem), estão distribuídos pelas orações, que a constituem. Estes elementos são fundamentais, como já se viu, para determinar foco do discurso. Simplificando a frase em orações, determinar o **foco real** é uma tarefa simplificada, pois passa a ser feita por oração e não por frase, onde as alternativas são mais reduzidas, quando não, únicas.

(4.13)

Oração principal: O rapaz gostava

Oração subordinada: que todos olhassem para ele.

(4.14) Após aplicar o *parser* obtém-se:

```
sta(fcl,
    subj(np,n(art(o,'<artd>','M','S'),'O',1),
         h(n(rapaz,'<Hattr>','M','S'),rapaz,2)),
    p(v_fin(gostar,'IMPF','1/3S','IND'),gostava,3),
    acc(fcl,
        sub(conj_s(que),que,4),
        p(v_fin(olhar,'IMPF','3P','SUBJ'),olhassem,5),
        piv(pp,h(prp(para),para,6),
            p(pron_pers(ele,'M','3S','NOM/PIV'),ele,'.',7))))
```

Para proceder a esta alteração, à versão inicial , o primeiro passo é pegar no input dado pelo *parser*, identificar as diferentes orações e os níveis em que está cada uma. Considerando o exemplo (4.14) após divisão em orações e identificação da sua estrutura hierárquica obtém-se a seguinte representação:

(4.15)

```
st(1).
[prof(0,[np(1-2,'M','S',rapaz,subj)]),
 prof(1,[pp(7,'M','3S',p)])]
```

Observa-se que a frase é constituída por duas orações, cada termo **prof** representa uma delas. **prof** deriva de profundidade e indica, considerando a estrutura da frase em árvore, o nível em que a oração está. A primeira é a principal, e está a um nível 0 (1º argumento do termo prof) , a segunda, por ser subordinada, está num nível imediatamente inferior.

Concluída esta representação, procuram-se os candidatos para os vários pronomes. Entende-se que não é possível referir numa frase, entidades de nível não superior (a profundidades diferentes de zero) das frases anteriores, uma vez que, a estrutura hierárquica da organização, das orações, leva a que nem todas as entidades possam ser referenciadas numa dada altura. Daqui resultam duas regras, para a pesquisa de candidatos:

- A procura de candidatos, na mesma frase onde ocorre o pronome, deve ser feita em todas as orações, ao mesmo nível hierárquico ou num nível superior que o nível correspondente ao pronome.
- A procura de candidatos, nas frases que antecedem a frase do pronome, deve ser feita apenas no topo da hierarquia, isto é, à profundidade 0.

Encontrados os candidatos temos uma representação, semelhante à do exemplo (4.16), em que estes, estão associados às orações, onde ocorrem os pronomes, e não à frase no seu todo.

(4.16)

```
st(1)-[prof(0,[np(1-2,'M','S',rapaz,subj)]),
       prof(1,[pp(7,'M','3S',p)],
    [cand(1,7,p,1,1-2,rapaz,subj)])
```

Daqui constroiem-se as "estruturas"Cf, como ilustra o exemplo seguinte:

(4.17)

```
st(1)-[cf([np(1-2,'M','S',rapaz,subj)])]

st(2)-[cf([cand(1,7,p,1,1-2,rapaz,subj)])]
```

Nesta representação, estão contempladas duas orações (st(1) e st(2)), que correspondem à frase dada em (4.13), mantém-se a notação **st(x)** para denominar uma oração, a fim de permitir a integração com a implementação já desenvolvida, que a partir deste momento é exactamente a mesma, à excepção de um pormenor. OS Cfs estão organizados por ordem decrescente de acordo com a proximidade do antecedente ao pronome, isto é o primeiro elemento da lista de Cfs possíveis, corresponde ao Cf com candidato mais próximo da anáfora. O objectivo, é que, em caso de empate a escolha deixe de ser arbitrária e passe a ser de acordo com a distância entre referente e antecedente. Este procedimento está, mais detalhado na secção dedicada às extensões.

Outro aspecto extremamente importante, desta alternativa à primeira versão, prende-se com a noção de contra-indexação. A inexistência da representação hierárquica da frase, na versão anterior, não permitia aplicar ao algoritmo, o filtro acrescentado por [Brennan *et al.*(1987] à versão original do *Centering*. Contudo, nesta nova versão já é possível considerá-lo. Para que se perceba exactamente porque, retome-se a descrição do mesmo. Este filtro, ou regra, é usado como forma de descartar candidatos impossíveis, e enuncia que:

> Um pronome não pode ter como antecedente um sintagma nominal que o *c-command* no mesmo domínio local.

Na frase *"A Sílvia gosta dela."*, *dela* não pode ter *A Sílvia* como referente, assim como na frase *"Ela gosta dela"*, os dois pronomes não podem partilhar o mesmo antecedente.

Entende-se por, *mesmo domínio local*, a mesma oração, isto é, A e B têm o mesmo domínio local se estão na mesma oração. Da noção de *c-command* dada em 3.2.1 percebe-se que a representação reaqueça, é fundamental para perceber se existe ou não contra-indexação, entre um pronome e um possível sintagma antecedente, ou entre pronomes. Considerem-se os exemplos dados pelas figuras 4.2 e 4.3. onde se pode ver a representação gráfica, em árvore, que o algoritmo produz, após receber um termo com uma frase. É a partir desta representação, que se divide a frase em orações, e se "filtram"os candidatos de acordo com a restrição aqui discutida.

Figura 4.2: *Ela gosta dela.*

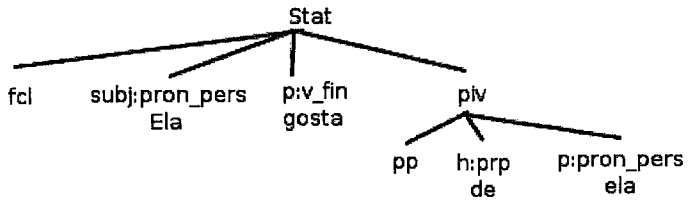
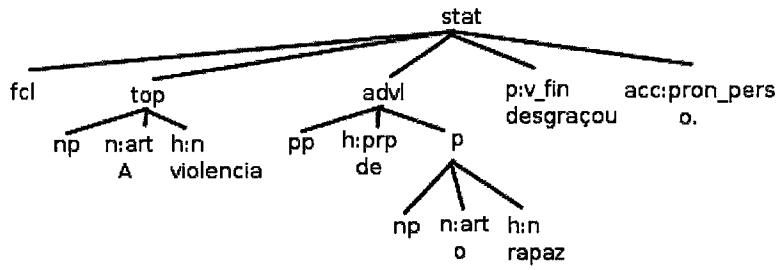


Figura 4.3: *A violência do rapaz desgraçou-o.*



Construir esta restrição, implicou implementar um predicado, que verificasse, se dois nós da árvore (N_1 e N_2) eramcontra-indexadoss, isto é se N_1 *c-comand* N_2. Para simplificar esta descrição, considera-se a simbologia da árvore. Dois nós são *irmãos*, quando são "dominados"pelo mesmo nó (têm o mesmo *pai*), e por definição, dois nós irmãos, *c-command* entre si, e aos seus *sobrinhos*. Em 4.2 pretende-se verificar se os dois pronomes podem ter o mesmo antecedente, pelo que o primeiro passo será testar se algum deles *c-comand* o outro. O sujeito (subj) tem o objecto preposicional (piv) como

irmão, pelo que é *tio* do segundo pronome pessoal *ela*, logo, por definição o primeiro pronome *c-comand* o segundo. Uma vez que estão no mesmo domínio local, não podem ter o mesmo antecedente. Em 3.3, o nó que domina *rapaz* não tem entre os seus irmãos, o nó do pronome pessoal, nem nenhum que domine o pronome, pelo que *rapaz* não o *c-comand*, e consequetemente pode ser dele, antecedente.

No algoritmo, este "teste"apenas é feito quando queremos confirmar se existem antecedentes na mesma oração que o pronome, ou mais que um pronome na mesma oração. Nesse caso o domínio local é o mesmo e é necessário garantir, que não ocorre contra-indexação. Nos restantes casos evita-se esta verificação e algum processamento desnecessário. Note-se agora um exemplo em que esta segunda versão do algoritmo permite contemplar situações que antes eram erradamente resolvidas.

Considere o pequeno texto:

(4.18) A Ana conduz um pólo.
(4.19) Ela conduz muito depressa.
(4.20) A Isabel concorre com ela aos fins de semana.
(4.21) Ela muitas vezes vence-a.

Com a primeira versão do algoritmo, para a frase (4.21) foram encontradas as seguintes opções:

```
cf([cand(4,21,subj,1,1-2,'Ana',subj),
    cand(4,24,acc,3,12-13,'Isabel',subj)],
   [cb(12-13,'Isabel')],
   [cp(1-2,'Ana')],shifting)

cf([cand(4,21,subj,1,1-2,'Ana',subj),
    cand(4,24,acc,1,1-2,'Ana',subj)],
   [cb(_2583,null)],
```

```
   [cp(1-2,'Ana')],shifting_1)

cf([cand(4,21,subj,3,12-13,'Isabel',subj),
    cand(4,24,acc,3,12-13,'Isabel',subj)],
   [cb(12-13,'Isabel')],
   [cp(12-13,'Isabel')],shifting_1)

cf([cand(4,21,subj,3,12-13,'Isabel',subj),
    cand(4,24,acc,1,1-2,'Ana',subj)],
   [cb(12-13,'Isabel')],
   [cp(12-13,'Isabel')],shifting_1)
```

A terceira opção foi eliminada, por ter o cb a *null*. A decisão recaiu sobre as últimas duas alternativas, que têm igual classificação, e a "menor"entre as restantes. Como em caso de empate, não há critério especificado, o algoritmo escolhe a primeira. Logo o antecedente de *Ela*, assim como, do pronome *a* foi *Isabel*. Naturalmente que este resultado está errado, pois é impossível que estes dois pronomes se refiram à mesma entidade.

Para a segunda versão, as hipóteses encontradas, foram:

```
cf([cand(4,21,subj,1,1-2,'Ana',subj),
    cand(4,24,acc,3,12-13,'Isabel',subj)],
   [cb(12-13,'Isabel')],
   [cp(1-2,'Ana')],
   shifting)

cf([cand(4,21,subj,3,12-13,'Isabel',subj),
    cand(4,24,acc,1,1-2,'Ana',subj)],
   [cb(12-13,'Isabel')],
   [cp(12-13,'Isabel')],
   shifting_1)
```

Dada a classificação de cada uma, conclui-se que o pronome *Ela* refere-se à *Isabel*, e o pronome *a* refere a *Ana*, como intuitivamente já se tinha entendido, pela leitura do texto. Este é um exemplo onde a versão do algoritmo que defende as iterações por oração e não por frase, encontra a solução correcta, quando o primeiro não o conseguia.

Os dados percentuais, com os quais se poderá comparar estas versões, são apresentados na secção dedicada à avaliação.

4.4 Extensões

Qualquer extensão que será aqui apresentada, é precedida da aplicação de um predicado, que de entre todas as possibilidades escolhe a (ou as), que tem a melhor classificação de acordo com a ordem já apresentada. Feito isto, ou se encontrou a solução, ou verifica-se um empate, isto é, existe mais do que um Cf candidato a solução, com idêntica classificação. Até aqui, sempre que isso acontecia, era arbitrariamente escolhida, a primeira alternativa que constava na lista. Estas extensões re-classificam as hipóteses, que de acordo com certos critérios, dão preferência a umas alternativas em detrimento de outras. Atenção que são extensões, que aplicam uma preferência e não funcionam como filtro para descartar hipóteses.

4.4.1 Preferência Hipótese mais próxima

Como o próprio nome indica, esta extenção, de entre hipóteses igualmente classificadas, devolve aquela cujo o antecedente ocorre mais perto do pronome. No caso em que o Cf tem mais do que uma anáfora, é preferida a hipótese que no conjunto está mais próxima. Embora [Mitkov(2002] tenha classificado esta preferência como fraca, no seu próprio algoritmo ele aplica-a, como último critério de preferência, [35] Hobbs já aqui descrito, em caso de empate prefere igualmente a opção mais recente. Outros autores também o fizeram, além disso, já aqui se relatou que a maior percentagem de antecedentes está, geralmente, na própria frase ou na imediatamente anterior, é destas observações

[35] ver secção 2.3.2.3

que surge esta extensão já incorporada na versão do algoritmo que resolve anáforas, por oração. O exemplo que se verá em seguida, corresponde a uma situação, em que a versão original do algoritmo (por frase) não resolve correctamente a anáfora, mas a versão por orações sim. Contudo isso acontece apenas pela organização da lista de Cf, (as restantes imposições não foram suficientes para eliminar todas as hipóteses incorrectas), que permite a aplicação directa desta preferência, isto é o primeiro elemento da lista de elementos com igual classificação, é aquele, em que o candidato a antecedente, está mais próximo do pronome.

(4.22) *"Atendendo à especificidade das questões suscitadas para a selecção dos equipamentos, o grupo de telecomunicações também depende da cooperação Política Europeia, procedeu a diversos estudos e realizou várias reuniões, tendo elaborado* ***uma proposta*** *que apresentou ao Grupo "Chefes de Comunicações", que* ***a*** *viria a aprovar na sua reunião de 23 de Maio de 1991 32, tendo-****a*** *recomendado ao Comité político, o qual por sua vez,* ***a*** *aprovou... "*

A lista de Cf's encontrados para o pronome *a* cuja palavra é a 5460 é:

```
cf([cand(139,5460,acc,139,5451-5452,proposta,acc),
    np(5465-5473,'F','S',reunião,p),
    np(5470-5473,'M','S','Maio',p)],
   [cb(_1456346,null)],
   [cp(5451-5452,proposta)],shifting)

cf([cand(139,5460,acc,139,5437-5438,'Cooperação_Política_Europeia',p),
    np(5465-5473,'F','S',reunião,p),
    np(5470-5473,'M','S','Maio',p)],
   [cb(_1456388,null)],
   [cp(5437-5438,'Cooperação_Política_Europeia')],shifting)
```

```
cf([cand(139,5460,acc,139,5425-5426,selecção,p),
    np(5465-5473,'F','S',reunião,p),
    np(5470-5473,'M','S','Maio',p)],
   [cb(_1456358,null)],
   [cp(5425-5426,selecção)],shifting)

cf([cand(139,5460,acc,139,5418-5419,especificidade,p),
    np(5465-5473,'F','S',reunião,p),
    np(5470-5473,'M','S','Maio',p)],
   [cb(_1456372,null)],
   [cp(5418-5419,especificidade)],shifting)


cf([cand(139,5460,acc,138,5409-5397,matéria,p),
    np(5465-5473,'F','S',reunião,p),
    np(5470-5473,'M','S','Maio',p)],
   [cb(_1456406,null)],
   [cp(5397-5409,matéria)],shifting)
```

O primeiro elemento deste conjunto tem como candidato o sintagma nominal *a proposta* dado pelos números 5451-5452, que naturalmente são os mais próximos do pronome 5460(*a*), e que resolve correctamente esta anáfora.

4.4.2 Preferência Gramatical

A preferência gramatical, atribui uma melhor classificação, às hipóteses, cujos candidatos a antecedente, têm a mesma função gramatical que o pronome. Percebe-se agora a notação dada aos candidatos, onde está inserida esta informação, particularmente pelo

terceiro e sétimo argumento, como se pode observar no exemplo (4.7) . Esta preferência é equivalente ao "fenómeno"de paralelismo sintáctico, já referido tanto na abordagem de [Lappin & Leass(1994], como na extensão à versão original da aproximação de Mitkov. Só é possível de ser aplicada, em algoritmos, que façam uso, na fase de pré-processamento, de um analisador sintáctico, como é o caso do estudo em questão. Na tentativa de tirar proveito desta situação, implementou-se esta extensão, a fim de melhorar o desempenho da versão original. Considere-se o exemplo dado pelo excerto (4.23)

(4.23)

... Os médicos discutiram ***a eutanásia passiva*** *em vários momentos da vida de Victor. No entanto, ninguém ousou executá-****la****.*

Na versão por orações obtêm-se as seguintes alternativas:

```
cf([cand(59,1168,acc,58,1161-1161,'Victor',p)],
   [cb(_153195,null)],
   [cp(1161-1161,'Victor')],shifting)

cf([cand(59,1168,acc,58,1158-1161,vida,p)],
   [cb(_153166,null)],
   [cp(1158-1161,vida)],shifting)

cf([cand(59,1168,acc,58,1151-1153,eutanásia,acc)],
   [cb(_153137,null)],
   [cp(1151-1153,eutanásia)],shifting)

cf([cand(59,1168,acc,56,1114-1114,'Victor',subj)],
   [cb(_153108,null)],
   [cp(1114-1114,'Victor')],shifting)
```

A única alternativa, onde o candidato tem a mesma função gramatical, de objecto directo (*acc*), que o pronome é a terceira. Portanto, de acordo com esta preferência , escolher-

se-ia *a eutanásia passiva* como antecedente, que é a opção correcta. Tanto na versão por orações, como na primeira (por frase), a escolha do candidato é errada.

4.4.3 Preferência de Centro

Esta preferência privilegia as hipóteses, em que, sendo o Cb *null*, o candidato tem como função gramatical, a mais saliente possível. Isto é, se não existir ligação com a frase anterior, em caso de empate, a solução será, se possível, equivalente ao Cp da *utterance* que se está a analisar.

Esta alternativa surge na tentativa de minorar a "indecisão"e consequentemente os erros de resolução, quando não são encontrados valores para os *backward-looking centers*. Simultaneamente, sendo o Cp o elemento mais saliente de uma *utterance*, parece lógico que em situação de dúvida seja ele, o antecedente do pronome. Veja-se um exemplo, dado pelo excerto do texto do exemplo (4.24), de uma situação em que a aplicação sem esta extensão, não resolve a anáfora, e com a extensão já resolve.

(4.24)

*"... **Cátia Sirley Moreira**, 23 filha de um mecânico,resolveu tentar a vida em um clube nocturno do Principiado das Astúrias, na Espanha. Depois de cinco anos de trabalho, ela voltou para Uruçu e virou dona de duas empresas de mototáxi, construiu uma casa, compro dois terrenos e, em breve, vai abrir uma loja de roupas. "Os homens de lá são horríveis, pinguços, drogados e fedorentos, mas você ganha bom dinheiro", conta **ela**, que acaba de embarcar para Espanha..."*

De acordo com a implementação por anáforas, obtêm-se as seguintes hipóteses para solução ao segundo *ela*, apresentado neste excerto.

```
cf([cand(15,333,subj,13,285-285,'Uruaçu',p),
    np(341,'F','S','Espanha',p)],
   [cb(_97775,null)],
   [cp(285-285,'Uruaçu')],shifting)

cf([cand(15,333,subj,12,254-257,filha,npred),
    np(341,'F','S','Espanha',p)],
   [cb(_97789,null)],
   [cp(254-257,filha)],shifting)

cf([cand(15,333,subj,12,249-257,'Cátia_Sirley_Moreira',subj),
    np(341,'F','S','Espanha',p)],
   [cb(_97805,null)],
   [cp(249-257,'Cátia_Sirley_Moreira')],shifting)
```

Portanto a anáfora 333 ***ela***, marcada a negrito no excerto em cima, tem três possíveis antecedentes com igual classificação. O algoritmo sem extensão, não aplica nenhum critério de preferência, e como tal escolheria para antecedente o candidato 285, *Uruaçu*. A anotação manual do Corpus, indica que esta anáfora tem como antecedente o sintagma nominal *Cátia Sirley Moreira*, pelo que a escolha estaria errada. Também na versão por frase, (1ª implementação do algoritmo) não seria encontrada a hipótese correcta, que tem, nesse caso, oito alternativas. Aplicando a preferência aqui referida, a situação em que o Cp é igual ao candidato é dada pelo último Cf. Esse seria o resultado escolhido, e uma vez que o antecedente é *Cátia Sirley Moreira*, conclui-se que com esta preferência é possível chegar a solução correcta.

A primeira extensão encontra um resultado, as outras duas, atribuem uma classificação numérica às hipóteses, para se sinalisar a preferência. Esta distinção foi feita, para permitir, caso se entendesse útil, aplicar as duas últimas extensões simultaneamente, também

por este motivo as pontuações dadas às situação preferenciais são idênticas (um ponto), de forma a que uma extensão, não prevaleça a sobre outra.

A fim de seleccionar a hipótese melhor classificada, de acordo com as duas últimas extensões, o predicado responsável, por encontrar a solução também foi extendido. Antes comparava as classificações *continuing*, *retaining*, *shifting_1* e *shifting* agora permite, caso a classificação não seja qualitativa, mas sim quantitativa de acordo com as preferências, escolher a que melhor foi classificada, isto é, a que teve uma pontuação mais elevada. No caso de prevalecer um empate, volta-se a escolher a primeira hipótese da listagem, sem nenhum critério em especial.

Antes de se proceder à avaliação das diferentes versões, introduzem-se os corpora usados neste trabalho.

5 Corpora

A existência de corpora anotados, mais para o Inglês do que para o Português, é uma das razões da evolução no estudo de resolução de anáforas. Os corpora são peça fundamental, nas tarefas e aplicações de processamento de língua natural, a sua anotação permite, particularmente para o processo de resolução de anáforas, o desenvolvimento de novas aplicações, a avaliação e melhoramento de outras já existentes. O corpus anotado, não só deve fornecer informação dos pares de co-referência antecedente-anáfora, mas também das cadeias anafóricas, que representam um fio condutor de uma entidade no discurso. Considerar que se determinou correctamente o antecedente de uma anáfora, não deve ser feito apenas quando esse elemento é o primeiro da cadeia entre a anáfora e os seus referentes, mas também, quando a associação é feita com qualquer elemento da cadeia. A anotação de corpora é uma tarefa morosa, exige tempo, estratégias e esquemas de anotação específicos, e por isso, embora a existência de corpora anotados seja tão fundamental, existem poucos e de tamanho relativamente reduzido.

Felizmente, este estudo beneficiou de Corpora anotados manualmente, que foram fundamentais para a próxima secção, a avaliação. Vão ser aqui apresentados quatro Corpora, como ilustra a tabela 5.1. A tabela mostra um número total de textos de cada corpus, assim como, e um número total de anáforas a resolver, que representam apenas, aquelas que este estudo se propõe resolver.

Tabela 5.1: Corpora e suas características

Corpora	**Nº de textos**	**Nº total de anáforas**
Corpus Jurídico	7	117
Corpus Jornalístico	9	102
Corpus Literário	3	424
Corpus Infantil	5	57
Total	24	700

O primeiro corpus (C_1) deriva do copora de textos jurídicos, e é composto por Pare-

ceres da Procuradoria Geral da República de Portugal. O segundo (C_2), é constituído por nove textos, jornalísticos. O terceiro Corpus (C_3) resulta de excertos do livro *O Alienista*, de Machado de Assis[36], formando assim um corpus literário. O último C_4, é constituído por um conjunto de textos escritos por crianças entre os 8 e 11 anos de idade[37]. Os corpora 1 e 4 estão escritos em português europeu, os 2 e 3, estão escritos em português do Brasil, todos eles foram anotados manualmente, com informação de co-referência.

Optou-se por se apresentar quatro corpora diferentes, na tentativa de perceber, se o género literário poderia ou não, influenciar os resultados. Todos eles têm um tipo de escrita, que os distingue naturalmente. C_1, o corpus jurídico tem uma escrita muito particular, pouco comum por exemplo, numa conversa falada, veja-se o exemplo dado pelo excerto (5.1), retirado do texto, apresentado no anexo A.

(5.1) *"...Artigo 3º.*
Designação de Conciliadores.
Cada Estado Parte na presente Convenção designará, num prazo de dois meses a contar da data de entrega em vigor da Convenção, ***dois conciliadores*** *sendo um* ***deles****, pelo menos nacional desse Estado..."*

O C_2 representa um estilo jornalístico como se pode observar pelo exemplo (5.2), retirado do texto apresentado no anexo B.

(5.2) *...Os integrantes do Conselho de Ética devem pedir a* ***Jefferson*** *provas de que parlamentares recebiam o mensalão. Até ao momento* ***ele*** *tem afirmado que não há provas. Alguns julgam que* ***Jefferson*** *não seria tão ingénuo e estaria blefando para apresentar as provas apenas na hora certa ...*

O estilo literário é dado em C_3, por frases como:

(5.3) *...O primeiro, um Falcão,* ***rapaz*** *de vinte e cinco anos, supunha-se estrela-d'alva, abria* ***os braços*** *e alargava* ***as pernas****, para dar-****lhes*** *certa

[36] http://bibvirt.futuro.usp.br/textos/autores/machadodeassis/alienista/alienista.html
[37] http://historiasinfantis.blogs.sapo.pt/

feição de raios, e ficava assim horas esquecidas a perguntar se o sol já tinha saído para ***ele*** *recolhe-se.*

que se podem observar no texto, exposto no anexo C.

Por último os textos dados por C_4 têm como objectivo representar estruturas sintácticas mais simples possível, daí que tenham sido procurados discursos escritos de, ou para, crianças. Um excerto que ilustra bem esta simplicidade é o dado pelo exemplo (5.4) e pode ser encontrado por completo no anexo D.

(5.4) ... *Certo dia, os três meninos foram à procura do* ***vento*** *e encontraram o* ***vento*** *num castelo. Os três ralharam muito com o* ***vento*** *mas não tinha sido* ***ele*** *o culpado*

Procedeu-se ainda, a uma análise quantitativa, relativamente ao número de frases, número de palavras, número de palavras distintas e ainda, número de palavras por frase em média. São esses dados que podem ser consultados nas próximas quatro tabelas.

Tabela 5.2: Frases e palavras corpus jurídico

Texto	Nº Frases	Nº Palavras	Nº Palavras distintas	Palavras/Frase
2	266	9745	1878	37
3	252	8203	1722	33
7	241	8534	1426	35
10	151	2111	973	14
12	334	10992	2042	33
13	382	9314	1896	24
14	129	3642	1013	28
Totais	**1755**	**52541**	**10950**	**29**

Tabela 5.3: Frases e palavras corpus jornalístico

Texto	Nº Frases	Nº Palavras	Nº Palavras distintas	Palavras/Frase
1	19	321	167	17
2	63	1607	615	26
4	45	1152	478	26
6	24	533	251	22
9	26	448	221	17
10	36	729	374	20
11	59	1168	481	20
12	38	782	365	21
13	33	672	305	20
Totais	**343**	**7412**	**3257**	**21**

Tabela 5.4: Frases e palavras corpus literário

Texto	Nº Frases	Nº Palavras	Nº Palavras distintas	Palavras/Frase
1	197	3759	1249	19
3	215	3212	1098	15
4	448	7358	2042	16
Totais	**860**	**14329**	**4389**	**17**

Tabela 5.5: Frases e palavras corpus infantil

Texto	Nº Frases	Nº Palavras	Nº Palavras distintas	Palavras/Frase
1	14	187	78	13
2	37	513	160	14
3	13	175	82	13
4	17	271	117	16
5	15	170	81	11
Totais	**96**	**1316**	**518**	**13**

Dos valores aqui tabelados, deve-se salientar, que embora o corpus jurídico tenha uma dimensão superior à do literário, na ordem das 38212 palavras, é o segundo corpus

que apresenta um maior número de anáforas, mais 307 que o jurídico. Esta informação, mostra as diferenças que os géneros literários impõem aos corpora. Sendo o literário mais reduzido, mas com maior número de anáforas, percebe-se que, nele é mais comum a ocorrência de expressões anafóricas, o que é relativamente compreensível, uma vez que nestes textos há tendência, para se manter entidades ao longo de todo o discurso (são geralmente textos centrados num número limitado de entidades), sendo por isso necessário, uso de referentes, de forma a tornar a leitura dos textos mais agradáveis. Um estilo completamente diferente do jurídico, que trata de várias matérias num só texto, sendo sucinto relativamente a cada uma delas. Por estas mesmas razões, os textos jornalísticos, embora com menos palavras que o jurídico têm, comparativamente com eles, um número de anáforas aproximado.

Estão assim introduzidos os corpora, utilizados neste estudo, resta saber os resultados conseguidos.

6 Avaliação

Nas duas secções precedentes, foram apresentadas não só a versão original da implementação, como a versão do *centering* por orações, as várias extensões e os corpora sobre os quais iriam ser feitos os testes. É altura de apresentar os resultados obtidos, avaliar o número de anáforas pronominais, correctamente resolvidas, e fazer algumas considerações sobre esses dados.

6.1 Método de Avaliação

Avaliar sistemas de resolução de anáforas, é uma tarefa extremamente difícil, que sofre interferência de muitos factores. É dependente de corpura com anotação manual, que sendo humana está igualmente sujeita a falhas. Seria ideal testar o sistema, num ambiente, também ele ideal, sem erros introduzidos pela fase de pré-processamento. Quanto maior o número de anáforas em teste, mais fiáveis os resultados. Os géneros literários podem e geralmente interferem, no tipo de anáforas que contemplam, estas podem ser, mais directas, ambíguas a ponto de exigir conhecimento do mundo real, com antecedentes muito afastados, ou com muitos candidatos, enfim as hipóteses são inúmeras. Este conjunto de situações, entre outras, dificulta a fiabilidade dos resultados. [Mitkov(2002] afirma:

> "In fact, it is possible that an anaphora resolution system that performs poorly is still based on a very effective algorithm."[38]

Daqui resulta a importância da distinção entre avaliação do sistema de resolução de anáforas (pré-processamento e algoritmo), e avaliação do algoritmo em si.

[38]De facto, é possível que um sistema de resolução de anáforas que tenha uma *performance* pobre, seja baseado num algoritmo bastante eficiente.

As primeiras medidas de avaliação usadas, são:

- Precisão (**P**)

  $P = \frac{X}{Y}$.

  Onde **X** representa o número de anáforas correctamente resolvidas e **Y** o número de anáforas resolvidas.

- Abrangência (**A**)

  $A = \frac{X}{Z}$.

  Onde **Z** representa o número de anáforas marcadas manualmente.

- F-Measure (**F-M**)

  $F\text{-}M = \frac{2*P*A}{P+A}$.

  Esta é uma medida que conjuga das duas anteriores, representando o desempenho geral do sistema.

O sistema inclui o pré-processamento e inevitavelmente, os erros que daí advêm, que na verdade não devem interferir com a medida de desempenho do algoritmo. Embora não se tenha testado o algoritmo no ambiente ideal, pois isso implicaria uma revisão(verificação de identificação de pronomes, sintagmas nominais, géneros e números dos mesmos...) ao corpora aqui usado, uma vez que é ele, que tem anotação manual, tentou fazer-se a distinção, sempre que possível, entre avaliação do sistema e do algoritmo. Para isso, comparou-se o número de anáforas encontradas manualmente, com as identificadas pelo analisador sintáctico. Com o resultado, percebe-se a diferença entre as duas situações, e a influência directa na avaliação do desempenho, caso as mesmas não sejam separadas. Na tentativa de apresentar os resultados mais próximos da realidade possíveis, sem interferências externas, foram consideradas, além das já apresentadas, as seguintes medidas de avaliação.

Taxa de sucesso = $\frac{A}{B}$.

Onde A, representa o número total de anáforas correctamente resolvidas. **B** pode ter diferentes interpretações, dependendo do corpus. Isto é, na situação em que o analisador sintáctico, encontra maior número de anáforas que as marcadas manualmente, **B** corresponde ao total das últimas, isto porque as anáforas bem resolvidas pertencem, necessariamente ao grupo das marcadas manualmente, as restantes não interessam e só produzem ruído. Para os casos em que o número de anáforas marcadas pelo analisador sintáctico é inferior às marcadas manualmente, **B** é dado pelas primeiras, pois só se pode avaliar o algoritmo tendo em conta o número de anáforas que ele realmente tentou resolver, contabilizar todas as marcadas manualmente seria, deixar interferir o ruído do pré-processamento. Se ao valor de B, retirarmos o número de anáforas para as quais não existe antecedente no campo de pesquisa (3 frases inclusive), ficamos com uma Taxa de sucesso$_x$, ainda mais "apurada", pois ignora as situações que o núcleo do algoritmo não tentou resolver. Permite-se fazer esta aproximação, pois a eficiência diz respeito principalmente à escolha da solução entre alternativas, comparativamente à fase inicial do filtro morfosintáctico e do *scope* onde se procuram os antecedentes.

Além da taxa de sucesso existem ainda, dois cálculos que permitem avaliar o algoritmo sob outra perspectiva e têm como base a taxa de sucesso$_x$.

Taxa sucesso trivial = $\frac{C}{D}$.

C = Número anáforas resolvidas com sucesso, que tinham apenas um candidato.

D = Número de anáforas com apenas um candidato.

Esta taxa representa essencialmente as anáforas com solução directa, dada pela concordância de género e número.

Taxa sucesso crítico = $\frac{E}{F}$.

E = Número anáforas resolvidas com sucesso, que tinham mais que um candidato.

F = Número de anáforas com mais que um candidato.

Tanto D como F são um subconjunto das anáforas a que o algoritmo se propõe resolver, são portanto um subconjunto de B.

A taxa de sucesso crítico, é de grande importância, pois permite perceber o sucesso dos factores de escolha, do algoritmo de resolução de anáforas, onde as restrições de concordância de género e número não foram suficientes para determinar o antecedente. É aqui que se percebe a real importância/eficiência da metodologia *centering*, para a resolução das anáforas pronominais.

6.2 Resultados

Os resultados são apresentados por corpus, onde figuram os dados de cada versão e extensão. Antes da apresentação dos resultados , a tabela 6.1 faz a correspondência da forma abreviada como serão denominadas as várias versões testadas.

Nos vários Corpus são apresentadas quatro tabelas. A primeira, faz a identificação das anáforas encontradas pela anotação manual, das encontradas pelo analisador sintáctico, das que o algoritmo não encontrou candidato no espaço da pesquisa e finalmente, o total de anáforas que devem ser contabilizadas como as que o algoritmo tentou resolver. Estes dados estão apresentados por texto e pela ordem em que foram enumerados. A segunda tabela, apresenta o número de anáforas total e correctamente resolvidas para duas situações distintas, a primeira diz respeito aos casos em que há apenas um candidato para uma anáfora e a segunda reporta vários candidatos para cada anáfora. Em terceiro, surge uma tabela dedicada às medidas precisão, abrangência e f-measure, apresentas por versão. Por último é apresentada uma tabela que, para cada versão, ilustra a taxa de sucesso$_x$, a taxa de sucesso trivial e a crítica.

Tabela 6.1: Versões e sua abreviatura

Abreviatura	Descrição da versão
v0	Implementação por frases
v1.0	Implementação por orações
v1.1	v1 + Preferência Gramatical
v1.2	v1 + Preferência de Centro
v1.3	v1 + Preferência de Centro + Preferência Gramatical

6.2.1 Corpus Jurídico

Tabela 6.2: Análise das Anáforas dos textos

Texto	Manual	*parser*	sem candidato	a Resolver
2	16	34	3	13
3	21	22	0	21
7	13	17	1	12
10	5	8	0	5
12	21	42	4	16
13	31	46	9	22
14	10	15	1	9
Totais	**117**	**184**	**18**	**99**

Neste caso o total de anáforas a resolver, é dado pelo número de anáforas encontrado manualmente menos aquelas para as quais, não se encontrou nenhum candidato.

Tabela 6.3: Valores obtidos de acordo com número de candidatos

Tipo	v0	v1.0	v1.1	v1.2	v1.3
A. com um candidato	16	24	24	24	24
A. correctamente resolvidas, com apenas um candidato	10	13	13	13	13
A. com mais do que um candidato	82	74	74	74	74
A. correctamente resolvidas, com mais do que um candidato	51	49	34	40	38

Tabela 6.4: Precisão, Abrangência e F-Measure

Versão	A. resolvidas	A. bem resolvidas	Precisão	Abrangência	F-Measure
v0	181	61	0,34	0,52	**0,41**
v1.0	161	62	0,39	0,53	**0,45**
v1.1	161	47	0,29	0,40	**0,37**
v1.2	161	53	0,33	0,45	**0,38**
v1.3	161	51	0,32	0,44	**0,37**

Tabela 6.5: Taxas de sucesso

Versão	Taxa sucesso$_x$	Taxa s. trivial	Taxa s. crítica
v0	**62,24%**	**62,5%**	**62,20%**
v1.0	**63,27%**	**54,17%**	**66,22%**
v1.1	47,96%	54,17%	45,95%
v1.2	54,08%	54,17%	54,05%
v1.3	52,04%	54,17%	51,35%

Nos vários resultados apresentados, observa-se pela tabela 6.7 o número de anáforas, com apenas um candidato é bastante inferior, às situações em que existem vários candidatos para o mesmo pronome. Essa situação é mais acentuada para as versões v1.$_x$. Relativamente à precisão, abrangência e *f-measure* observa-se uma melhoria entre as versões v0 e v1.0, para as três as medidas. Essa melhoria observa-se também na tabela 6.9, com excepção da taxa de sucesso trivial, que sofre alguma descida.

6.2.2 Corpus Jornalístico

Tabela 6.6: Análise das Anáforas dos textos

Texto	Manual	*parser*	sem candidato	a Resolver
1	7	5	0	5
2	19	17	1	16
4	12	5	0	5
6	5	3	1	2
9	9	6	1	5
10	17	10	2	8
11	22	13	0	13
12	8	5	0	5
13	3	1	0	1
Totais	**102**	**65**	**5**	**60**

Neste caso o total de Anáforas a resolver é dado pelo número de anáforas encontrado pelo analisador sintáctico menos aquelas para as quais, não se encontrou nenhum candidato.

Tabela 6.7: Valores obtidos de acordo com número de candidatos

Tipo	v0	v1.0	v1.1	v1.2	v1.3
A. com um candidato	10	13	13	13	13
A. correctamente resolvidas, com apenas um candidato	7	7	7	7	7
A. com mais do que um candidato	50	47	47	47	47
A. correctamente resolvidas, com mais do que um candidato	20	25	21	17	23

Tabela 6.8: Precisão, Abrangência e F-Measure

Versão	A. resolvidas	A. bem resolvidas	Precisão	Abrangência	F-Measure
v0	62	27	0,44	0,26	**0,33**
v1.0	60	32	0,53	0,31	**0,39**
v1.1	60	28	0,47	0,27	**0,34**
v1.2	60	24	0,4	0,24	**0,3**
v1.3	60	30	0,5	0,29	**0,37**

Tabela 6.9: Taxas de sucesso

Versão	Taxa sucesso$_x$	Taxa s. trivial	Taxa s. crítica
v0	**45%**	**70%**	**40%**
v1.0	**53,33%**	**53,85%**	**53,19%**
v1.1	46,67%	53,85%	44,68%
v1.2	40%	53,85%	36,17%
v1.3	50%	53,85%	48,94%

Para este corpus, verificam-se exactamente as mesmas observações do anterior. A taxa de sucesso crítica, comparativamente com o corpus jurídico é mais baixa. Contudo a melhoria, apresenta entre a versão v0 e a versão v1.0 é mais acentuada.

6.2.3 Corpus Literário

Tabela 6.10: Análise das Anáforas dos textos

Texto	Manual	*parser*	sem candidato	a Resolver
1	119	90	3	87
3	81	58	4	54
4	224	188	16	172
Totais	**424**	**336**	**23**	**313**

Tabela 6.11: Valores obtidos de acordo com número de candidatos

Tipo	v0	v1.0	v1.1	v1.2	v1.3
A. com um candidato	37	39	39	39	39
A. correctamente resolvidas, com apenas um candidato	14	14	14	14	14
A. com mais do que um candidato	276	274	274	274	274
A. correctamente resolvidas, com mais do que um candidato	92	104	85	88	96

Tabela 6.12: Precisão, Abrangência e F-Measure

Versão	A. resolvidas	A. bem resolvidas	Precisão	Abrangência	F-Measure
v0	300	106	0,35	0,25	**0,29**
v1.0	297	118	0,4	0,29	**0,34**
v1.1	297	99	0,33	0,23	**0,27**
v1.2	297	102	0,34	0,24	**0,28**
v1.3	297	110	0,37	0,26	**0,31**

Tabela 6.13: Taxas de sucesso

Versão	Taxa sucesso$_x$	Taxa s. trivial	Taxa s. crítica
v0	**33,87%**	**37,83%**	**33,33%**
v1.0	**37,70%**	**35,90%**	**37,96%**
v1.1	31,63%	35,90%	31,02%
v1.2	32,59%	35,90%	31,75%
v1.3	35,14%	35,90%	34,67%

A diferença entre número de anáforas com um antecedente e anáforas com mais do que um antecedente é muito elevada neste corpus, também é nele que ocorre o maior número de anáforas, 424, marcadas manualmente. As taxas trivial e crítica para as versões v1.1 e v1.2 estão muito próximas entre si, e abaixo dos valores obtidos na versão v1.3. Contudo nenhum deles, ultrapassa os 35,9% e os 37,98% encontrados em v1.0.

6.2.4 Corpus Infantil

Tabela 6.14: Análise das Anáforas dos textos

Texto	Manual	*parser*	sem candidato	a Resolver
1	6	6	0	6
2	20	20	8	12
3	10	10	0	10
4	14	14	3	11
5	7	27	1	6
Totais	**57**	**57**	**12**	**45**

Tabela 6.15: Valores obtidos de acordo com número de candidatos

Tipo	v0	v1.0	v1.1	v1.2	v1.3
A. com um candidato	27	21	21	21	21
A. correctamente resolvidas, com apenas um candidato	20	18	18	18	18
A. com mais do que um candidato	18	24	24	24	24
A. correctamente resolvidas, com mais do que um candidato	12	14	13	15	14

Tabela 6.16: Precisão, Abrangência e F-Measure

Versão	A. resolvidas	A. bem resolvidas	Precisão	Abrangência	F-Measure
v0	52	32	0,62	0,56	**0,59**
v1.0	49	32	0,65	0,56	**0,60**
v1.1	49	31	0,63	0,54	**0,58**
v1.2	49	33	0,67	0,58	**0,62**
v1.3	49	32	0,65	0,56	**0,60**

Tabela 6.17: Taxas de sucesso

Versão	Taxa sucesso$_x$	Taxa s. trivial	Taxa s. crítica
v0	**71,11%**	**74,07%**	**66,67%**
v1.0	**71,11%**	**85,71%**	**58,33%**
v1.1	68,89%	85,71%	54,17%
v1.2	73,33%	85,71%	62,5%
v1.3	71,11%	85,71%	58,33%

Este, é o único corpus, onde o número de anáforas com apenas um candidato, para a versão v0 é superior e nas restantes está muito próximo do número de anáforas com mais do que um candidato. É também o único que regista um decréscimo da taxa de sucesso crítica, e um aumento da taxa de sucesso trivial entre a versão v0 e a v1.0. No entanto as medias de precisão, abrangência e *f-measure* mantêm a tendência já verificada pelos restantes corpora.

6.3 Interpretação dos Resultados

Relativamente às medidas de precisão, abrangência e *f-measure* entende-se que esta última é a mais importante das três, isto porque conjuga as outras duas, ou seja é uma medida que não só considera o grau de confiança do sistema como tem em conta a abrangência do mesmo. Em todos os corpora analisados observou-se uma melhoria da versão v0 para a v1.0, quer na precisão, quer na abrangência, quer na medida *f-measure*. Observando os vários valores calculados de *f-measure*, verifica-se que os três melhores variam entre 0,39 e 0,60 e, o mais baixo, é de 0,34, dado pelo corpus literário. Fazendo a média destes quatro valores obtém-se 0,45.

Já foi referido que a medida *f-measure*, avalia o sistema no seu todo, incluindo a fase de pré-processamento. Pelos quadros que comparam as anáforas encontradas manualmente, com as encontradas pelo analisador sintáctico percebe-se a "incompatibilidade"dos valores apresentados, o analisador pode reduzir o número real, ou no extremo aumentá-lo em cerca do dobro. Considera-se por isso que as três primeira medidas são menos "precisas"que as taxas de sucesso na tarefa de avaliar o algoritmo propriamente dito,uma vez que não permitem contornar os erros de pré-processamento tão evidentes. É por isso que se dá especial importância às taxas de sucesso, que permitem distinguir a avaliação do sistema da avaliação do algoritmo. Contudo estas taxas continuam a sofrer influência negativa da fase de pré-processamento, pois erros como marcações erradas de sintagmas nominais, ou não determinísticas quanto ao género, não foram possíveis de excluir.

No corpus jurídico a taxa de sucesso$_x$ cresce 1% entre a versão v0 e a v1.0. Se, se considerar a taxa crítica, há um aumento de 4%, a taxa trivial baixa. Seja neste corpus ou em qualquer outro, a maior importância é reservada à taxa crítica por ser aquela que traduz o processo de escolha de anáforas, quando há mais do que um candidato a antecedente, é portanto a mais ilustrativa do sucesso real do algoritmo.

No corpus jornalístico, assim como no literário, verifica-se a mesma tendência entre as duas versões acima mensionadas. De notar que no segundo corpus, o crescimento da

taxa de sucesso crítica é mais acentuado, cerca de 13%.

Relativamente ao último corpus (infantil) a taxa de sucesso$_x$é idêntica em ambas as versões, no entanto a taxa trivial sobe na versão v1.0 e a crítica baixa. Observando as tabelas 6.3, 6.7, 6.11 e 6.15 percebe-se que em todos os corpus excepto no último, o número de anáforas com apenas um antecedente é muito inferior às restantes. É natural que assim seja, porque o corpus infantil é constituído por textos de estrutura muito simples e pouco elaborados, com poucas entidades diferentes. O facto da percentagem crítica baixar de v0 para v1.0 está relacionado com essa diferença, isto é o número de anáforas bem resolvidas com apenas um candidato está muito próximo do total de anáforas bem resolvidas, daí o crescimento da taxa trivial , portanto os erros ocorrem principalmente nas situações em que há mais que um candidato. Contudo neste corpus específico, dada a sua dimensão foi possível fazer-se uma análise mais exacta para se tentar compreender este resultado. Viu-se que a grande maioria das anáforas resolvidas com antecedente errado, são consequência da situação em que o sintagma nominal correcto, não estava nas três frases onde é feita a procura de antecedentes. Logo todas estas situações, independentemente do algoritmo não seriam correctamente resolvidas, no que resulta uma interferência negativa. Uma desvantagem deste corpus é o número reduzido de anáforas que torna a sua avaliação, comparativamente com as restantes, menos exacta.

Já foi referida a melhoria de resultados de versão v0 para v1.0, essa melhoria, observada em qualquer uma das medidas, é a clara justificação, para que os testes com as extensões preferência de gramatical e /ou de centro, tenham sido aplicados à versão 1.0. Os resultados mostram que ambas as preferências, não só individualmente mas também em conjunto, baixam a *f-measure* e a taxa de sucesso do algoritmo de resolução de anáforas. Entre elas, mesmo não melhorando os resultados da versão v1.0, a preferência de centro é a que obtém melhores resultados.

Das tabelas 6.3, 6.7 e 6.11 observa-se que o número de anáforas com apenas um candidato é menor para a versão v1.0 do que para a primeira. Estes valores são naturais, uma vez que a versão v1.0 onde o algoritmo está implementado por orações, há um maior

cuidado na escolha dos candidatos, de acordo com a sua profundidade na árvore sintáctica, da frase em que os sintagmas nominais ocorrem. Daqui resulta também que para a versão v0, o número de candidatos, é superior às situações que ocorrem em v1.0. Estes valores, também justificam a ligeira descida da taxa de sucesso trivial entre a versão v0 e a v1.0.

Observando as várias percentagens individualmente verifica-se, que as três melhores variam entre os 53% e os 67% e, a mais baixa, dada pelo copus literário, é de 37,96%, o género literário pode influênciar os resultados. Fazendo a média da taxa de sucesso crítica, para a melhor das cinco versões (v1.0) dos quatro corpora aqui apresentados obtem-se uma taxa de 53,93%.

Embora a *f-measure* e a taxa de sucesso crítico sejam medidas diferentes, comparando-as percebe-se que a taxa dá-nos valores mais elevados, o que era de esperar uma vez que esta tenta eliminar, o mais possível, erros induzidos ao algoritmo. De qualquer forma os resultados nos vários corpus são confirmados pelas diferentes medidas.

Comparar os resultados destas implementações, com os de implementações do *centering* para inglês, ou com outros algoritmos que se propões a resolver diferentes tipos de anáforas, é tarefa ingrata e dela não se podem tirar conclusões, isto porque não só à diferenças entre os corpora, como na língua dos mesmos. O ideal seria comparar com outras implementação para o português, o mais próximo que existe é a implementação realizada por Thiago Coelho [Coelho(2005] que desenvolveu o algoritmo *Lappin* e *Leass* para língua portuguesa. Propõe resolver não só, anáforas pronominais na terceira pessoa, como também, pronome reflexivos e recíprocos. Os resultados obtidos rondam os 43,56% um pouco menos que os 53,93% aqui apresentados. Consciente das "limitações"desta comparação referem-se aqui os 59,35% obtidos por Mitkov na sua *knowledge-poor* abordagem. E os cerca de 76.50% obtidos por Brennan, Friedman e Pollard, num ambiente ideal.

7 Conclusão e Trabalho Futuro

Sempre que ocorre uma anáfora num discurso, esta é facilmente associada ao seu antecedente pelo leitor/ouvinte. O grande objectivo deste trabalho é simular este processo, construindo um sistema de resolução de anáforas pronominais para a língua portuguesa. Este sistema é baseado na metodologia do *Centering* e foca a sua atenção num grupo restrito de pronomes pessoais, aqueles que realmente têm função anafórica. A integração de sistemas como este, em aplicações de processamento de língua natural é fundamental para o desenvolvimento das mesmas. O processo de extracção de informação de textos, por exemplo, é com certeza mais eficiente se as anáforas que aí ocorrem tiverem a entidade que referem, correctamente identificada.

Desenvolveram-se cinco versões do algoritmo. A primeira resolve as anáforas por frase, isto é, em cada frase é encontrado um *forward-looking center (Cf)*, *backward-looking center* e o elemento mais saliente do *Cf* corresponde ao *Preferred center*. O primeiro contém todas as entidades que ocorrem na frase, pronomes, ou sintagmas nominais. No fim de aplicar o algoritmo, os pronomes estão substituídos pelo antecedente escolhido entre os candidatos. A versão v1.0, por sua vez, aplica o algoritmo por oração, para que isso seja possível utiliza-se uma estrutura em árvore de cada frase que permite não só a divisão da mesma nas referidas orações, como a implementação da restrição de contra-indexação, esta restrição permite descartar antecedentes e pronomes contra-indexados, que por regra não podem co-referir. Esta é uma vantagem fundamental sobre a primeira versão, pois a primeira não contemplava a existência deste filtro, o que faz, por si só, da versão v.1.0 uma melhor abordagem ao problema de resolução de anáforas pronominais, independentemente dos resultados. As restantes versões dizem respeito à aplicação da implementação v1.0 com preferência gramatical, de centro e finalmente uma extensão que contempla ambas as preferências. O seu desenvolvimento foi motivado pela falta de critério quando ocorria um empate entre candidatos, isto é quando chegada à fase final, depois de verificar quais os candidatos com classificação preferível, eram encontrados dois ou mais, nas mesmas condições.

Dos resultados conclui-se um aumento no desempenho da versão v.0 para a versão 1.0, como seria de esperar, uma vez que a segunda resulta da análise dos resultados da primeira, e portanto não é mais que uma tentativa de melhorar o desempenho da versão inicial, esta observação é prova do sucesso dos objectivos inicialmente apresentados. Relativamente às extensões nenhum dos dois critérios, concordância de função gramatical entre o candidato e o pronome, ou preferência pelo candidato com função gramatical mais saliente, melhorou o desempenho da versão em que foram aplicadas. Das três extensões apresentadas na secção 4.4, conclui-se que em caso de empate, a preferência pelo candidato mais próximo é a que sobressai, pois estando ela incorporada na versão 1.0, e dado que os resultados das versões v1.1, v1.2 e v1.3 baixam sensivelmente o desempenho do algoritmo, nota-se que é mais vantajoso a situação em que o antecedente está mais próximo da anáfora.

Relativamente aos corpora, a taxa de sucesso crítico é mais elevada para o corpus jornalístico (66,22%) e mais baixa para o literário (37,93%), portanto dado um intervalo tão significativo percebe-se que o corpus tem influência nos resultados. O que também era de esperar, uma vez, que foram usados corpora de géneros literários que por sua vez, influênciam o tipo de anáforas que aí ocorrem. Esta conclusão é secundada pela observação da tabela 3.1, onde são apresentados os resultados obtidos pelo algoritmo "pai". Aí observa-se que, a aplicação do *centering* desenvolvida por Brennan, Friedman e Pollard, está sujeita a variações, dependendo do corpus onde é aplicada.

A comparação com outros resultados, faz parte da avaliação de um sistema de resolução de anáforas, contudo ela só pode ser feita correctamente, se o objecto em estudo, nomeadamente os corpora, aos quais se submetem as aplicações forem os mesmos, assim como as medidas de avaliação utilizadas. Essa comparação só pode ser feita, com algum cuidado, relativamente ao corpus jurídico e literário que é parcialmente comum ao trabalho realizado por [Coelho(2005], centrado na implementação do algoritmo *Lappin* e *Leass*. Para o corpus jurídico a percentagem de sucesso, obtida por este autor foi de 35,15% e no corpus literário foi de 31,32%. Embora o total de anáforas não seja exactamente o mesmo, os valores aqui obtidos para estes dois corpus, 66,22% e 37,96%,

principalmente para o jurídico, denotam uma taxa de sucesso significativamente mais elevada. A comparação com o algoritmo original, é feita com igual cuidado, pois, neste caso trata-se de avaliações para línguas distintas. Nas percentagens gerais, está-se a comprar os 53,93% aqui conseguidos com os 76,59% obtidos por [Brennan *et al.*(1987], o que faz concluir que se conseguiu um algoritmo bastante eficiente, para a língua portuguesa, comparativamente com a versão para o inglês. Uma diferença de cerca de 23% é bastante satisfatória, quando se comparam dois algoritmos sendo, o que obteve melhores resultados, testado num ambinete ideal, em contrate com os corpora sujeito aos erros de pré-processamento aqui utilizado.

O facto da taxa de sucesso ser diferente de 100% demostra que o sistema de resolução de anáforas aqui apresentado tem limitações. Algumas dessas limitiações foram encontradas e são aqui enumeradas:

- O elevado número de candidatos por pronome.

  Provoca, sempre que há mais que um pronome, numa oração ou frase (dependendo da versão), um grande número de combinações. Além disso, quantos mais candidatos para o mesmo pronome, mais provável é, que mais do que um, tenha a mesma classificação, no que resultam várias situações de empate, entre as quais geralmente se encontra a solução correcta. Para estas situações, nenhuma extensão, se mostrou extremamente útil a solucionar. Dadas as regras de escolha de candidatos ser mais restrita em v1.0, esta versão gera menos possibilidades o que é, além da regra de contra-indexação, um dos factores para justificar a melhoria de desempenho, nela verificada.

- Os erros de pré-processamento

  São vários, não só na marcação das anáforas, que se viu ser bastante insuficiente pelos quadros 6.2, 6.6, 6.10 e 6.14. Outras situações além desta se verificam:

  Marcação indefinida de género nos sintagmas nominais

  Este exemplo é retirado do texto 1 do corpus literário, e mostra como o sintagma nominal *D.Evarista* não têm género definido.

```
subj(np,h(prop('D._Evarista','M/F','S'),'D._Evarista',3388)
```

Identificação incorrecta de pronomes

O analisador sintáctico identificou a palavra *a* como um pronome pessoal, como mostra o exemplo,

```
acc(pron_pers(ela,'F','3S','ACC'),a,1580)
```

contudo esta situação não corresponde a um pronome, como se pode ver pelo excerto do texto retirado do corpus literário, onde esta palavra ocorre.

> *...Costa não se deteve um minuto, foi ao devedor e perdoou-lhe* ***a*** *dívida...*

Os erros de pré-processamento, constituem um problema, que pode baixar em muito, as taxas de sucesso, independentemente da eficiência do algoritmo em questão.

- Situações de concordância entre género e número impossíveis.
  Acontecem, por exemplo, quando o antecedente é constituído por dois sintagmas nominais que individualmente têm número singular, mas a anáfora refere-se aos dois sendo o seu número plural. Ou ainda, quando os géneros são incompatíveis, por exemplo o antecedente é feminino, e a anáfora masculino, é o mostra a situação descrita em baixo, extraída do corpus jurídico.

  > *"...Mas* ***decisiva*** *nesta matéria, - por Regulamento a dever, obviamente, respeitar- sê-lo-a, decerto a Lei, nº 12/91..."*

  O pronome *lo* nunca será resolvido com *decisiva* pois este antecedente não passaria no filtro morfossintáctico.

- Anáforas que necessitam de conhecimento do mundo real para serem correctamente resolvidas.
  Note-se o seguinte excerto,

*"...Acontece porém, em dois aspectos carecerá ainda de ser completado o citado sistema, conforme este diploma regulamentar **o** configura..."*

o pronome refere-se à exigência que o diploma impõe, contudo essa inferência é feita pela pessoa que lê o texto, a marcação automática, exigiria outro tipo de ferramentas que não as usadas neste trabalho, para encontrar o antecedente correcto, caso seja possível.

- Frequência de pronomes, cujo género é ambíguo, marcados correctamente pelo analisador sintáctico como 'M/F'
  O pronome pessoal *lhe* e a sua forma no plural não têm género específico, são marcados como o exemplo extraído do corpus literário.

  ```
  dat(pron_pers(ele,'M/F','3S','DAT'),lhe,3483)
  ```

  Neste caso, na procura de candidatos a antecedente, são aceites não só os sintagmas nominais no feminino como no masculino, aumenta o número de candidatos, aumenta a probabilidade de erro. A taxa de sucesso do corpus literário, parece estar intimamente relacionada com este facto, pois a grande parte dos pronomes aí presentes são *lhe (s)*[Coelho(2005].

- Menor número de anáforas com apenas um candidato entre a versão v0 para v.1.0
  Uma vez que o "campo"de pesquisa é o mesmo para as duas versões (três frases inclusive), o número de anáforas seria, intuitivamente o mesmo. Contudo a versão v1.0 é mais restritiva na escolha dos candidatos, sendo que esta é dependente da oração em que o possível antecedente ocorre. No entanto a taxa de sucesso trivial baixa da versão original para a segunda, o que demonstra que não só se estão a excluir os casos em que o único candidato não correspondia à solução correcta, como alguns em que ele seria realmente o antecedente em questão. Não é, uma descida extremamente acentuada, mas acontece.

Como trabalho futuro, pretende-se minimizar ao máximo as limitações aqui apresentadas. As que resultam do analisador sintáctico, que são a maioria, são difíceis de

contornar, uma vez que o algoritmo necessita de pré-processamento, a tarefa será construir um corpus, que após passar pelo analisador é verificado manualmente a fim de lhe retirar erros e ambiguidades. Poder-se-á então testar o sistema num ambiente ideal, facilitando, entre outras coisas o cálculo das medidas de avaliação.

Outra tarefa importante é o estudo em pormenor dos resultados, principalmente dos casos de anáforas incorrectamente resolvidas, há que tentar modelar regras para limitar os candidatos, como se tentou fazer da versão v0 para a v1.0, com garantia de que não se perdem as hipóteses que correspondiam ao antecedente correcto.

A cooperação mais íntima com linguistas, também permitirá, afinar detalhes específicos da língua portuguesa e incorporá-los no algoritmo, que foi inicialmente pensado para o inglês.

Pretende-se numa próxima extensão à implementação, alargar os candidatos (até agora correspondem apenas aos sintagmas nominais), a outros pronomes que concordem em género e número e não infringam a restrição de contra-indexação. A grande vantagem desta medida será, manter as cadeias de referência entre entidades, permitindo não só considerar o antecedente imediatamente anterior ao pronome, mas qualquer um que lhe seja equivalente.

A integração deste sistema, em aplicações de processamento de língua natural, está igualmente pensada. O sistema de Pergunta/Resposta para a Língua Portuguesa também desenvolvido na Universidade de Évora, corresponde a uma dessas aplicações.

O baixo número de estudos nesta área, para a língua portuguesa, é prova viva do quanto ainda há a fazer, e da importância da análise detalhada que se deve dedicar aos resultados obtidos, em cada aplicação. Muitos das limitações foram aqui identificadas, resta principalmente, dedicar-lhes mais atenção e tentar modelar estratégias que permitam um aperfeiçoamento do algoritmo. A definição de um critério, que escolha o melhor candidato em caso de empate, parece ser chave do problema. De um bom critério de preferência, resultará, com certeza o aumento da taxa de sucesso deste algoritmo de resolução de anáforas pronominais.

Referências

[A. McEnery(1997] A. McEnery, S.B., I. Tanaka (1997). Corpus annotation and reference resoluiton. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Pratical, Robust Anaphora Resolution*, 67–74, Madrid, Spain.

[Al-Kofani *et al.*(1999] Al-Kofani, K., Grom, B. & Janckson, P. (1999). Anaphora resolution in extracting of treatment history language from court opinions by partial parsing. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, 138–146, Oslo, Norway.

[Ariel(1990] Ariel, M. (1990). *Accessing noun phrase antecedents*. London: Routledge.

[Baldwin(1997] Baldwin, B. (1997). Cogniac: high precision coreference with limited knowledge and linguistics resourses. In *Proceedings of the ACL'97/EACL'97 workshop on Operational Factors in Pratical, Robust Anaphora Resolution*, 38–45, Madrid, Spain.

[Bick(2000] Bick, E. (2000). *The Parsing System "PALAVRAS"*. Ph.D. thesis, University of Arhus, DK.

[B.J. Grosz(1986] B.J. Grosz, S.W., A.K. Joshi (1986). Towards a computational theory of discourse interpretation, preliminary draft.

[Brennan *et al.*(1987] Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987). A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics*, 155–162, Association for Computational Linguistics, Morristown, NJ, USA.

[Carbonell & Brown(1988] Carbonell, J.G. & Brown, R.D. (1988). Anaphora resolution: a multi-stategy approach. In *Proceedings of the 12th International Conference on Computational Linguistics (COLLING'88)*, vol. I, 96–101, Budapest, Hungray.

[Coelho(2004] Coelho, J. (2004). Estudo da anaforicidade pronominal em textos jurídicos. In *Salão de Iniciação Científica da UFGRS-SIC'2004*, Porto Alegre, Brasil.

[Coelho(2005] Coelho, T.T. (2005). *Resolução de anáfora pronominal em português utilizando o algoritmo de Lappin e Leass*. Master's thesis, Universidade Estadual de Campinas.

[Dahl & Ball(1990] Dahl, D. & Ball, C. (1990). Reference resolution in pundit. Tech. rep., Paoli: Center for Advanced information Technology.

[de Abreu(2005] de Abreu, S.C. (2005). *Análise de Expressões Referenciais em Corpus Anotado da Língua Portuguesa*. Master's thesis, Universidade do Vale do Rio dos Sinos.

[Hobbs(1978] Hobbs, J.R. (1978). Resolving pronoun reference. Tech. rep.

[H.Wada(1990] H.Wada (1990). Discourse processing in mt:problems in pronominal translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, vol. 1, 73–75, Helsonki Finland.

[J.K. Gundel(1993] J.K. Gundel, R.Z., N. Hedberg (1993). *Language*, chap. 69, 274–307.

[Kheler(1997] Kheler, A. (1997). Probabilistic coreference in information extraction.

[Kibble(2001] Kibble, R. (2001). A reformulation of rule 2 of centering theory. *Comput. Linguist.*, **27**, 579–587.

[Lappin & Leass(1994] Lappin, S. & Leass, H. (1994). *An Algorithm for Pronominal Anaphora Resolution*, vol. 20, 535–361. Computational Linguistics.

[Lappin & McCord(1990a] Lappin, S. & McCord, M. (1990a). *An Algorithm for Pronominal Anaphora Resolution*, vol. 16, 197–212. Computational Linguistics.

[Lappin & McCord(1990b] LAPPIN, S. & MCCORD, M. (1990b). A syntactic filter on pronominal anaphora in slot grammar. In *28th Annual Meeting of the Association for Computational Linguistics*, 135–142.

[Market *et al.*(2003] MARKET, K., NISSIM, M. & MODJESKA, N.N. (2003). Using the web for anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*, Budapest, Hungary.

[Mitkov(1998] MITKOV, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of Colling-ACL'98*, 869–875.

[Mitkov(2000] MITKOV, R. (2000). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC-2000)*, 96–107.

[Mitkov(2001] MITKOV, R. (2001). Outstanding issues in anaphora resolution (invited talk). In *CICLing '01: Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, 110–125, Springer-Verlag, London, UK.

[Mitkov(2002] MITKOV, R. (2002). *Anaphora Resolution*. Pearson Education, `www.history-minds.com`.

[Paraboni & de Lima(1998] PARABONI, I. & DE LIMA, V.L.S. (1998). Possessive pronominal anaphor resolution in portuguese written texts. In *36th Annual Meeting of the Associtation for Computational Linguistics and 17th Internatinal Conference on Computational Linguistics*.

[Poesio *et al.*(2000] POESIO, M., CHENG, H., HENSCHEL, R., HITZEMAN, J., KIBBLE, R. & STEVENSON, R. (2000). Specifying the parameters of centering theory: a corpus-based evaluation using text from application-oriented domains. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 400–407, Association for Computational Linguistics, Morristown, NJ, USA.

[Reinhart(1983] REINHART, T. (1983). *Anaphora and semantic intterpretation.*

[Rich & LuperFoy(1988] RICH, E. & LUPERFOY, S. (1988). An architecture for anaphora resolution. In *Proceedings of the second conference on Applied natural language processing*, 18–24, Association for Computational Linguistics, Morristown, NJ, USA.

[Sidner(1979] SIDNER, C.L. (1979). Towards a computational theory of definite anaphora comprehension in english discourse. Tech. rep., Cambridge, MA, USA.

[Soon *et al.*(2001] SOON, W.M., NG, H.T. & LIM, D.C.Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, **27**, 521–544.

[Strube(1998] STRUBE, M. (1998). Never look back: an alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, Montreal, Canada.

[Strube & Hahn(1999] STRUBE, M. & HAHN, U. (1999). vol. 25, 309–344. Computational Linguistics.

[Stys & Zemke(1995] STYS, M. & ZEMKE, S. (1995). Incorporating discourse aspects in english-polish mt: towards robust implementation. In *Proceedings of the International Conference 'Recent Advances in Natural Language Processing' (RANLP'95)*, 95–102, Tzigov Chark, Bulgaria.

[Tetreault(1999] TETREAULT, J. (1999). Analysis of sintax-based pronoun resolution methods. In *Proceedings of the 37th annual Meeting of the Association for Computational Linguistcs (ACL'99)*, 602–606, Maryland, USA.

[Walker(1989] WALKER, M. (1989). Evaluationg discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the ACL (ACL'97)*, 251–161, Vancouver, Canada.

[Walker(1997] WALKER, M. (1997). Centering anaphora resolution, and discourse structure. Tech. rep., ATT Labs Research, 180 Park Ave. Florham Park, N.J.07932.

[Webber(1988] WEBBER, B.L. (1988). Discourse diexis reference to discourse segments. In *Proceedings of the 26th annual meeting on Association for Computer Linguistics*, 113–122.

Anexos

A Exemplo Corpus Jurídico

Não se encontram obstáculos de natureza jurídica a que Portugal assine a Convenção em análise - "Convenção sobre Conciliação e Arbitragem no âmbito da Comissão para a Segurança e Cooperação na Europa" -, redigida na Reunião de Peritos sobre Resolução Pacífica de Diferendos que teve lugar em Genebra de 12 a 23 de Outubro do corrente ano.

Senhor Secretário de Estado Adjunto do Ministro da Justiça, Excelência:

O Gabinete de Sua Excelência o Ministro dos Negócios Estrangeiros enviou ao Gabinete de Sua Excelência o Ministro da Justiça o texto da Convenção sobre Conciliação e Arbitragem cuja redacção foi concluída na Reunião de Peritos sobre Resolução Pacífica de Diferendos que teve lugar em Genebra de 12 a 23 de Outubro último.

Informando que aquela Convenção deverá ser assinada por Estados participantes da Comissão para a Segurança e Cooperação na Europa no decorrer do Conselho de Ministros que terá lugar em 14 e 15 de Dezembro, próximo futuro, em Estocolmo, e que Portugal, por ter apoiado politicamente a iniciativa e ter participado na redacção da Convenção, "deverá encontrar-se entre os primeiros signatários por ocasião da reunião de Estocolmo", aquele Gabinete solicitou informação "sobre se o texto da Convenção suscita alguma objecção face ao ordenamento jurídico interno".

Vossa Excelência determinou que fosse ouvida, com urgência, a Procuradoria-Geral da República.

Tendo sido ordenada a distribuição pelo Conselho Consultivo, cumpre prestar a Informação solicitada, no quadro do disposto na alínea a) do artigo 34° da Lei n° 47/86, de 15 de Outubro, examinando a referida Convenção no estrito plano da legalidade, especialmente a sua conformidade à Constituição da República.

Consta do preâmbulo da referida "Convenção sobre Conciliação e Arbitragem no âmbito da Comissão para a Segurança e Cooperação na Europa":
Os Estados partes na presente Convenção, sendo Estados que participaram na Conferência sobre Segurança e Cooperação na Europa;
Conscientes da sua obrigação, conforme prevista nos artigos 2°, n° 3, e 33° da Carta das Nações Unidas 2, de resolver os respectivos diferendos de forma pacífica;
Sublinhando que não pretendem, de modo algum, depreciar outras instituições ou mecanismos, incluindo o Tribunal Internacional de Justiça, o Tribunal Europeu dos Direitos Homem, o Tribunal de Justiça das Comunidades Europeias e o Tribunal Permanente de Arbitragem;
Reiterando o seu compromisso solene em resolver os seus diferendos através de meios pacíficos, bem como a sua decisão de desenvolver mecanismos para resolver disputas entre Estados participantes;
Relembrando que a implementação total de todos os princípios da Comissão para a Segurança e Cooperação na Europa e dos compromissos constitui, em si mesma, um elemento essencial na prevenção de diferendos entre os Estados que participam na Comissão para a Segurança e Cooperação na Europa;
Preocupados em desenvolver e fortalecer os compromissos referidos, nomeadamente, no Relatório do Encontro de Peritos sobre Resolução Pacífica de Diferendos adoptado em Vallette e apoiado pelo Conselho de Ministros dos Negócios Estrangeiros da Comissão para a Segurança e Cooperação na Europa, na sua reunião realizada em Berlim, a 19 e 20 de Junho de 1991, "Acordam no seguinte:
Como se escreveu no parecer n° 21/59, deste corpo consultivo, a propósito de um projecto de Convenção relativa ao processo de arbitragem internacional, elaborado pela Comissão de Direito Internacional da Organização das Nações Unidas.
Além de ser um dos meios mencionados na Carta das Nações Unidas, artigo

33°, para a resolução pacífica dos diferendos internacionais, a arbitragem corresponde há muito a um anseio das nações pacíficas e várias foram as tentativas de lhe dar organização cada vez mais geral, perfeita e eficaz. Portugal tem especial autoridade para dar o seu voto à finalidade desta nova iniciativa:

além de haver proposto e aceitado recorrer à arbitragem voluntária em numerosos casos ((x) V.F.

DE CASTRO CALDAS, "Portugal e a Arbitragem Internacional", páginas 85 e seguintes.

subscreveu com os Países Baixos, em 5 de Julho de 1894, o primeiro tratado entre países europeus em que se estabeleceu uma cláusula compromissória de carácter geral (-), isto é, abrangendo quaisquer questões que viessem a suscitar-se entre esses Estados, e proclamou na sua Constituição Política o princípio de que "Portugal preconiza a arbitragem como meio de dirimir os litígios internacionais" (artigo 4°, § único), depois de enunciar o princípio da obediência às convenções internacionais.

Somente é de deplorar que a comunidade internacional ainda hoje não esteja suficientemente evoluída, a ponto de quebrar antigas resistências à adopção geral da arbitragem obrigatória, permanente e institucional dos litígios entre os Estados, que já fizeram gorar a tentativa de estabelecer a obrigatoriedade da arbitragem na Convenção da Haia de 1907 para a solução pacífica dos conflitos internacionais, em revisão da de 1899, depois de larga discussão e propostas sobre a lista de matérias a que seria delimitada essa obrigatoriedade, a mais extensa das quais fora apresentada pela delegação portuguesa à Conferência (-).

Assim a Convenção (da Haia para a Resolução Pacífica de Conflitos Internacionais) de 1907, depois de repetir no artigo XXXVIII o princípio, já estabelecido na de 1899, de que "nas questões de ordem jurídica, e em primeiro lugar nas questões de interpretação ou de aplicação de convenções

internacionais, a arbitragem é reconhecida pelas potências contratantes como o meio mais eficaz, e ao mesmo tempo mais equitativo de determinar os litígios que não forem resolvidos pela vias diplomáticas", limitou-se a acrescentar o seguinte voto:

Por consequência, seria para desejar que nos litígios sobre as questões acima mencionadas as potências recorressem, dado o caso, à arbitragem sempre que as circunstâncias o permitissem".

É sabido que a referida Convenção da Haia, concluída em 1899 e revista em 1907, criou o Tribunal Permanente de Arbitragem, designação, aliás, enganadora, visto que, na realidade, não passa de uma lista permanente de árbitros, designados pelos Estados signatários, quatro membros por cada Estado.

O Tribunal propriamente dito tem de ser constituído em cada caso - o que ocorreu em apreciável número de casos importantes 3 -, designando cada Estado litigante dois árbitros, dos quais apenas um pode ser seu nacional.

Por outro lado, foram instituídos tribunais judiciais internacionais, particularmente o Tribunal Judicial de Justiça, o Tribunal de Justiça das Comunidades Europeias e o Tribunal Europeu dos Direitos do Homem.

No que toca à arbitragem, ao longo do século XX têm sido concluídos numerosos tratados permanentes (de arbitragem) dos quais o mais antigo é o que foi celebrado entre a Grã-Bretanha e a França, em 1903, onde se dispunha que as divergências de carácter jurídico ou relativamente à interpretação dos tratados seriam submetidas ao Tribunal Permanente de Arbitragem da Haia, desde que não afectem os "interesses vitais, a independência ou a honra dos dois Estados e não digam respeito aos interesses de terceiras potências".

Como já se disse, Portugal já subscreveu numerosos tratados de arbitragem.

E a actual Constituição da República favorece a subscrição de convenções como a ora em análise, ao dispor no seu artigo 7º, nº 1, que Portugal se rege "nas relações internacionais pelos princípios da [...] solução pacífica dos conflitos internacionais [...]".

Daí que se deva concluir, em tese geral, que nada há a opor à subscrição, por parte de Portugal de uma convenção como a ora em análise.
Ponto é que não se ofendam, salvaguardando-se outros princípios fundamentais, consignados naquele artigo 7° e no artigo 8° nomeadamente a independência nacional, a igualdade entre os Estados, a não ingerência nos assuntos internos do Estado Português, e o respeito por outros acordos internacionais que já vinculem o Estado Português.
Importa, pois, proceder à análise da Convenção em causa, citando, em especial, as disposições mais relevantes.
Dispõe-se no Capítulo I (parte geral):
Artigo 1°.
Estabelecimento do Tribunal).
Será estabelecido um Tribunal de Conciliação e Arbitragem a fim de resolver, por meio de conciliação e, se apropriado, de arbitragem, quaisquer diferendos que lhe sejam submetidas nos termos do disposto na presente Convenção".
Artigo 2°.
Comissões de Conciliação e Tribunais de Arbitragem).
Uma Comissão de Conciliação, criada especificamente para cada diferendo, ficará encarregue da conciliação.
A Comissão será constituída por conciliadores retirados de uma lista elaborada em conformidade com o disposto no artigo 3°.
Um Tribunal Arbitral, criado especificamente para cada diferendo, ficará encarregue da arbitragem.
O Tribunal será constituído por árbitros retirados de uma lista elaborada em conformidade com o disposto no artigo 4°.
Os conciliadores e os árbitros, em conjunto, constituirão o Tribunal de Conciliação e Arbitragem no âmbito da Comissão para a Segurança e Cooperação na Europa, a seguir designado por "o Tribunal".
Artigo 3°.

Designação de Conciliadores).
Cada Estado Parte na presente Convenção designará, num prazo de dois meses a contar da data de entrada em vigor da Convenção, dois conciliadores sendo um deles, pelo menos, nacional desse Estado.
O outro conciliador pode ser nacional de outro Estado que participe na Comissão para a Segurança e Cooperação na Europa".
Artigo 4°.
Designação de Árbitros).
Cada Estado Parte na presente Convenção designará, num prazo de dois meses a contar da data de entrada em vigor da Convenção, um árbitro e um suplente (substituto), os quais podem ser nacionais desse Estado ou nacionais de um outro Estado que participe na Comissão para a Segurança e Cooperação na Europa".
Artigo 5°.
Independência dos Membros do Tribunal e do Secretário).
Os conciliadores, os árbitros e o Secretário desempenharão as suas funções com total independência".
Artigo 6°.
Privilégios e Imunidades).
Os conciliadores, os árbitros, o Secretário, bem como os agentes e advogados das partes num diferendo gozarão, durante o exercício das respectivas funções no território dos Estados Partes na presente Convenção, dos privilégios e imunidades concedidos às pessoas ligadas ao Tribunal Internacional de Justiça".
O artigo 7° refere-se à eleição e constituição do Bureau do Tribunal (Presidente, Vice-Presidente e três outros membros).
O artigo 8° refere-se às decisões do Tribunal, do Bureau, das Comissões de Conciliação e dos Tribunais Arbitrais, tomadas por maioria dos membros participantes.

O artigo 9º refere-se ao Secretário e ao staff da secretaria.
O artigo 10ª reporta-se à sede do Tribunal, em local ainda não determinado.
O artigo 11º dispõe que o Tribunal adoptará o seu próprio Regulamento, que será sujeito à aprovação dos Estados Partes na Convenção.
O Artigo 12º determina que o Regulamento do Tribunal estabelecerá regras quanto ao uso de línguas.
O artigo 13º refere-se aos custos do Tribunal, a suportar pelos Estados Partes.
O artigo 14º refere-se ao relatório (periódico) de actividades.
O artigo 15º reporta-se à informação - pelo Secretário do Tribunal ao secretariado da Comissão para a Segurança e Cooperação na Europa - de todos os pedidos de conciliação ou arbitragem para imediata transmissão aos Estados participantes.
O artigo 16º dispõe que, durante os procedimentos, as Partes em disputa deverão abster-se de qualquer acção que possa agravar a situação ou prejudicar a resolução do conflito.
O artigo 17º dispõe que as Partes em disputa devem suportar os seus próprios custos.
É manifesto que as disposições citadas não ofendem qualquer norma ou princípio fundamentais, nomeadamente os atrás citados.
São disposições que, no essencial (com as adaptações adequadas a cada caso), regem as arbitragens internacionais e que, como tal, são incorporadas nos respectivos tratados ou convenções.
Vejam-se, a este propósito, os artigos seguintes da Convenção para Solução Pacífica dos Conflitos Internacionais - ratificada pelo Estado Português - que criou o já referido Tribunal Permanente de Arbitragem 4:
Artigo 42º.
O Tribunal permanente será competente para todos os casos de arbitragem, a menos que haja acordo entre as Partes para o estabelecimento de uma

jurisdição especial".

Artigo 44°.

Cada Potência contratante designará o número máximo de quatro pessoas, de competência reconhecida nas questões de direito internacional, gozando da mais alta consideração moral, e dispostas a aceitar as funções de árbitro.

As pessoas assim designadas serão inscritas com o título de Membros do Tribunal numa lista que será notificada a todas as potências contratantes, por intermédio da repartição".

Artigo 45°.

Quando as potências contratantes quiserem dirigir-se ao Tribunal permanente para a resolução de uma divergência ocorrida entre Elas, a escolha dos árbitros que devem constituir o Tribunal competente para se pronunciar sobre essa divergência, deverá ser feita de entre a lista geral dos Membros do Tribunal.

Na falta da constituição do Tribunal por acordo das Partes, proceder-se-á da maneira seguinte:

Cada uma das Partes nomeará dois árbitros, dos quais só um poderá ser seu nacional ou escolhido de entre os que foram designados por Ela como membros do Tribunal permanente.

Estes árbitros escolherão juntamente um árbitro de desempate".

Mais concretamente:

as normas do artigo 1° a 5° da Convenção ora em análise respeitam o princípio da igualdade e oferecem a garantia de uma decisão que salvaguarde os interesses - nomeadamente a independência e a honra - dos Estados partes, tratando-se, como é o caso, de conciliadores e de árbitros com elevadas qualificações em Direito Internacional - confira artigos 3°, n° 2, e 4°, n° 2 - que desempenham as suas funções com total independência - artigo 5°.

A norma do artigo 6° (Privilégios e Imunidades) é usual em convenções deste tipo, não havendo, também aqui, obstáculos de ordem jurídica.

E as demais disposições do Capítulo I não suscitam qualquer reparo.

O capítulo II refere-se à competência da Comissão de Conciliação do Tribunal Arbitral.

Nos termos do artigo 18°, qualquer Estado parte pode submeter a uma Comissão de Conciliação qualquer diferendo com outro Estado parte que não tenha sido resolvido através de negociação dentro de um prazo razoável e, de igual modo, qualquer diferendo pode ser submetido a um Tribunal Arbitral em conformidade com o estipulado no artigo 26°.

O artigo 19°, salvaguardando outros meios existentes de resolução dos conflitos, prevê que a Comissão de Conciliação ou Tribunal Arbitral constituídos para resolver um diferendo não tomem quaisquer medidas verificando-se:

a anterior sujeição do diferendo a outro órgão de arbitragem cuja jurisdição tenha de ser aceite pelas partes;

a anterior aceitação da jurisdição exclusiva de um outro órgão jurisdicional que tenha competência para decidir com efeito vinculativo;

o encaminhamento (posterior) do diferendo, por uma ou todas as partes, para um tribunal cuja jurisdição sobre o diferendo tenha de ser legalmente aceite pelas partes;

a resolução (posterior) do diferendo pelas partes.

No caso de desentendimento entre as partes no diferendo, relativamente à competência da Comissão ou do tribunal, a decisão da questão levantada recairá na Comissão ou no Tribunal.

Estas normas não suscitam dúvidas no plano em que nos movemos, observando o princípio do respeito pelos acordos anteriormente firmados pelo Estado Português.

Refira-se, ainda, que o n° 4 do artigo 19° permite aos Estados, no momento da assinatura, ratificação ou adesão à presente Convenção, fazer uma reserva de modo a garantir a compatibilidade do mecanismo de resolução de diferendos previsto pela Convenção com outros meios de resolução de diferendos decorrentes

de compromissos internacionais aplicáveis a cada Estado.
A economia do parecer dispensa-nos de averiguar a possibilidade (e necessidade) de o Estado Português fazer qualquer reserva, neste campo.
O capítulo III refere-se à "Conciliação".
O artigo 20° trata do pedido para a constituição de uma Comissão de Conciliação.
Qualquer Estado parte da presente Convenção pode solicitar a constituição de uma Comissão relativamente a um diferendo com um ou mais Estados.
Por outro lado, a constituição da Comissão pode também ser pedida por acordo entre dois ou mais Estados partes ou entre Estados partes e um ou mais Estados participantes na Comissão para a Segurança e Cooperação na Europa.
Nos termos do artigo 21° cada Estado parte deve indicar um conciliador (da lista estabelecida nos termos do artigo 3°), prevendo esta disposição como resolver (a indicação e o número de conciliadores) no caso de haver mais que dois Estados partes em litígio.
Mais se regula a constituição da Comissão, cabendo ao Bureau indicar, em regra, três conciliadores.
O artigo 22° regula os trâmites da constituição da Comissão, a partir do pedido de constituição.
O artigo 23° refere-se ao procedimento, dispondo-se que será confidencial e que todas as partes em disputa terão o direito a serem ouvidas.
A Comissão determinará o procedimento a seguir, de acordo com os artigos 10° e 11° e o Regulamento do Tribunal, e depois de consultar as partes.
O artigo 24° dispõe que a Comissão ajudará as partes a encontrarem uma resolução de acordo com o direito internacional.
O artigo 25° refere-se aos resultados da conciliação.
O n° 1 prevê a hipótese de as partes alcançarem uma resolução aceitável, com a ajuda da Comissão.
Os números seguintes regulam os trâmites a seguir, quando as partes não

atingiam essa resolução amigável:
a Comissão lavrará um relatório final contendo as propostas da Comissão para a resolução da disputa;
as partes, notificadas desse relatório, não são obrigadas a aceitar essas propostas.
Perante tais normas deve concluir-se não haver lugar a objecções de ordem jurídica, muito menos a nível constitucional, pois não se mostra atingido qualquer princípio fundamental ou de ordem pública.
Nomeadamente, o procedimento aqui regulado decorre em plena igualdade de armas, e as partes não são obrigadas a aceitar as propostas da Comissão, salvaguardando-se, assim, a independência e os interesses dos Estados partes.
Nada a objectar, pois, nesta parte.
O Capítulo IV trata da "Arbitragem".
Nos termos do artigo 26° o pedido de arbitragem pode ser feito a qualquer momento por acordo entre dois ou mais Estados partes ou entre um ou mais Estados partes e um ou mais Estados participantes na Comissão para a Segurança e Cooperação na Europa.
Por outro lado, os Estados partes podem em qualquer momento declarar que reconhecem a jurisdição de um Tribunal Arbitral, sujeita a reciprocidade.
Quando um litígio é submetido a um Tribunal Arbitral, o Tribunal pode indicar medidas que devem ser observadas para evitar o agravamento do litígio ou maiores dificuldades na procura da solução.
O artigo 27° regula os termos (conteúdo) do pedido de arbitragem.
O artigo 28° regula a constituição do Tribunal Arbitral em termos próximos dos fixados para a constituição da Comissão de Conciliação.
O artigo 29° regula o procedimento de arbitragem.
Importa destacar que todas as partes têm o direito a ser ouvidas durante o procedimento, que será conforme aos princípios de um "fair trial" (julgamento amigável).

O procedimento será constituído de uma parte escrita e de uma parte oral.
O artigo 30° refere-se à função do Tribunal Arbitral:
decidir, de acordo com o direito internacional, os litígios apresentados, sem prejuízo do poder do Tribunal de decidir o caso "ex aequo et bono", se as partes nisso convierem.
O artigo 31° dispõe sobre a "decisão" do Tribunal, que deverá indicar razões em que se baseia, e apenas terá força vinculativa entre as partes em litígio e relativamente à questão posta.
A decisão é irrecorrível, mas é possível pedir a sua revisão, baseada na descoberta de algum facto decisivo até aí desconhecido pelo Tribunal e pelas partes.
O artigo 32° dispõe que a decisão será publicada pelo secretário, bem assim notificada às partes.
Também aqui se não devem levantar dúvidas quanto à conformidade destas normas com os princípios fundamentais atrás referidos.
No tocante ao valor e efeitos da decisão, as normas indicadas (artigo 31°) são muito próximas, essencialmente idênticas, às dos artigos 81° e 83° da Convenção referida pelo n° 3.1.2 .
Daqui que se deva concluir, também nesta parte, que nada há a objectar.
O Capítulo V - "Disposições Finais" - contém normas de estilo que não levantam quaisquer dificuldades.
Lembre-se, apenas, que o artigo 34° estipula que a Convenção não pode ser objecto de qualquer reserva não expressamente autorizada, caso da prevista no artigo 19°, n° 4 .
Conclusão:
Na sequência do exposto deve concluir-se que não se encontram obstáculos de natureza jurídica a que Portugal assine a Convenção em análise - "Convenção sobre Conciliação e Arbitragem no âmbito da Comissão para a Segurança e Cooperação na Europa" -, redigida na Reunião de Peritos sobre Resolução

Pacífica de Diferendos que teve lugar em Genebra de 12 a 23 de Outubro do corrente ano.

Tradução (da versão inglesa) da nossa responsabilidade.

Carta das Nações Unidas:

Todos os membros deverão resolver as suas controvérsias internacionais" por meios pacíficos, de modo que não sejam ameaçadas a paz, a segurança e a justiça internacionais" - nº 3 do artigo 2º.

As partes num conflito, que possa vir a constituir uma ameaça à paz e à segurança internacionais, procurarão, antes de tudo, chegar a uma solução por negociação, inquérito, mediação, conciliação, arbitragem, solução judicial, recurso a entidades ou acordos regionais ou a qualquer outro meio pacífico à sua escolha.

Conselho da Segurança convidará, quando julgar necessário, as referidas partes, a resolver, por tais meios, os seus conflitos" - artigo 33º.

DE CASTRO CALDAS, "Portugal e a Arbitragem Internacional", páginas 85 e seguintes.

ANTÓNIO JOSÉ FERNANDES, "Relações Internacionais", Editorial Presença, página 388 . 4) A título meramente exemplificativo, cite-se o Tratado de Conciliação e Arbitragem entre o Estado Português o Grão-Ducado do Luxemburgo, assinado 15.8.1929 (Diário Geral, I Série, de 29/8/31), cujos artigos 1º a 4º dispõem:

Artigo 1º.

Todos os litígios cujo objecto seja um direito, de qualquer natureza, alegado por uma das Partes Contratantes e contestado pela outra, e, especialmente, as divergências mencionadas no artigo 13º do Pacto da Sociedade das Nações, que não tenham podido ser resolvidos num prazo razoável pelos processos diplomáticos ordinários, serão submetidos ao Tribunal Permanente de Justiça Internacional".

Artigo 2º.

As Partes Contratantes estabelecerão, para cada caso particular, um compromisso

especial determinando nitidamente o objecto da divergência, as competências particulares que poderiam ser devolvidas ao Tribunal Permanente de Justiça Internacional como todas as outras condições ajustadas entre si.
Artigo 3°.
Antes de qualquer processo perante o Tribunal Permanente de Justiça Internacional, a divergência será, a pedido de qualquer das Partes, submetida, para conciliação, a uma comissão internacional permanente, designada Comissão Permanente de Conciliação, constituída nos termos do presente Tratado".
Artigo 4°.
A Comissão Permanente de Conciliação será composta de cinco membros.
As Partes Contratantes nomearão cada uma um comissário da sua livre escolha e designarão, de comum acordo, as outras três e, entre estes, o presidente da comissão.
Estes três comissários não deverão ser súbditos das Partes Contratantes, nem ter domicílio no seu território nem estar ao serviço delas.
Deverão ser todos os três de nacionalidade diferente.

B Exemplo Corpus Jornalístico

O presidente nacional do PTB, Roberto Jefferson, comparece nesta terça-feira ao Congresso Nacional para falar sobre as acusações que fez sobre o esquema de mensalão ao Conselho de Ética da Câmara .
Há nove dias, Jefferson denunciou uma operação em que o tesoureiro do PT, Delúbio Soares, seria o responsável pelo pagamento de mesadas de 30000 a congressistas do PP e do PL .
Caso não conseguir provar o suposto esquema, o deputado, que já responde a um processo disciplinar, pode também ser cassado .

Os integrantes do Conselho de Ética devem pedir a Jefferson provas de que

parlamentares receberiam o mensalão .
Até o momento ele tem afirmado que não há provas .
Alguns julgam que Jefferson não seria tão ingênuo e estaria blefando para apresentar as provas apenas na hora certa .
O parlamentar, porém, é alvo de acusação em outro escândalo .
Ele será investigado sobre as denúncias de corrupção na Empresa Brasileira de Correios e Telégrafos e no Instituto Resseguros Brasil .

O governo deve aguardar o depoimento de Jefferson para tomar decisões .
Uma eventual reforma ministerial e medidas administrativas de reação à crise dependerão dos desdobramentos de seu discurso .
" Estamos todos esperando alguma prova material . Se o depoimento se resumir a um testemunho, teremos de tentar comprovar com outros depoimentos . " , disse o presidente do Conselho , deputado Ricardo Izar .

Nos últimos dias , petebistas que conversaram com Jefferson alimentaram a versão de que ele possuiria gravações comprometedoras de aliados e ministros .
À Folha, ele afirmou não ter provas .
" Ele está consciente de que vai ser cassado . E me disse que , no último discurso , antes da cassação , o Brasil vai ser outro . " , disse um dos aliados do petebista .
O depoimento deve começar às 14hs30.

C Exemplo Corpus Literário

As crônicas da vila de Itaguaí dizem que em tempos remotos vivera ali um certo médico, o Dr. Simão Bacamarte, filho da nobreza da terra e o maior dos médicos do Brasil, de Portugal e das Espanhas. Estudara em Coimbra e

Pádua. Aos trinta e quatro anos regressou ao Brasil, não podendo el-rei alcançar dele que ficasse em Coimbra, regendo a universidade, ou em Lisboa, expedindo os negócios da monarquia.

A ciência, disse ele a Sua Majestade, é o meu emprego unico; Itaguaí é o meu universo.

Dito isso, meteu-se em Itaguaí, e entregou-se de corpo e alma ao estudo da ciência, alternando as curas com as leituras, e demonstrando os teoremas com cataplasmas. Aos quarenta anos casou com D. Evarista da Costa e Mascarenhas, senhora de vinte e cinco anos, viúva de um juiz de fora, e não bonita nem simpática. Um dos tios dele, caçador de pacas perante o Eterno, e não menos franco, admirou-se de semelhante escolha e disse-lho. Simão Bacamarte explicou-lhe que D. Evarista reunia condições fisiológicas e anatômicas de primeira ordem, digeria com facilidade, dormia regularmente, tinha bom pulso, e excelente vista; estava assim apta para dar-lhe filhos robustos, sãos e inteligentes. Se além dessas prendas, únicas dignas da preocupação de um sábio, D. Evarista era mal composta de feições, longe de lastimá-lo, agradecia-o a Deus, porquanto não corria o risco de preterir os interesses da ciência na contemplação exclusiva, miúda e vulgar da consorte.

D. Evarista mentiu às esperanças do Dr. Bacamarte, não lhe deu filhos robustos nem mofinos. A índole natural da ciência é a longanimidade; o nosso médico esperou três anos, depois quatro, depois cinco. Ao cabo desse tempo fez um estudo profundo da matéria, releu todos os escritores árabes e outros, que trouxera para Itaguaí, enviou consultas às universidades italianas e alemãs, e acabou por aconselhar à mulher um regímen alimentício especial. A ilustre dama, nutrida exclusivamente com a bela carne de porco de Itaguaí, não atendeu às admoestações do esposo; e à sua resistência, explicável,

mas inqualificável, devemos a total extinção da dinastia dos Bacamartes.

Mas a ciência tem o inefável dom de curar todas as mágoas; o nosso médico mergulhou inteiramente no estudo e na prática da medicina. Foi então que um dos recantos desta lhe chamou especialmente a atenção, o recanto psíquico, o exame de patologia cerebral. Não havia na colônia, e ainda no reino, uma só autoridade em semelhante matéria, mal explorada, ou quase inexplorada. Simão Bacamarte compreendeu que a ciência lusitana, e particularmente a brasileira, podia cobrir-se de "louros imarcescíveis", expressão usada por ele mesmo, mas em um arroubo de intimidade doméstica; exteriormente era modesto, segundo convém aos sabedores.

A saúde da alma, bradou ele, é a ocupação mais digna do médico.

Do verdadeiro médico, emendou Crispim Soares, boticário da vila, e um dos seus amigos e comensais.

A vereança de Itaguaí, entre outros pecados de que é argüida pelos cronistas, tinha o de não fazer caso dos dementes. Assim é que cada louco furioso era trancado em uma alcova, na própria casa, e, não curado, mas descurado, até que a morte o vinha defraudar do benefício da vida; os mansos andavam à solta pela rua. Simão Bacamarte entendeu desde logo reformar tão ruim costume; pediu licença à Câmara para agasalhar e tratar no edifício que ia construir todos os loucos de Itaguaí, e das demais vilas e cidades, mediante um estipêndio, que a Câmara lhe daria quando a família do enfermo o não pudesse fazer. A proposta excitou a curiosidade de toda a vila, e encontrou grande resistência, tão certo é que dificilmente se desarraigam hábitos absurdos, ou ainda maus. A idéia de meter os loucos na mesma casa, vivendo em comum, pareceu em si mesma sintoma de demência e não faltou quem o insinuasse à própria mulher do médico.

Olhe, D. Evarista, disse-lhe o Padre Lopes, vigário do lugar, veja se seu marido dá um passeio ao Rio de Janeiro. Isso de estudar sempre, sempre, não é bom, vira o juízo.

D. Evarista ficou aterrada. Foi ter com o marido, disse-lhe "que estava com desejos", um principalmente, o de vir ao Rio de Janeiro e comer tudo o que a ele lhe parecesse adequado a certo fim. Mas aquele grande homem, com a rara sagacidade que o distinguia, penetrou a intenção da esposa e redargüiu-lhe sorrindo que não tivesse medo. Dali foi à Câmara, onde os vereadores debatiam a proposta, e defendeu-a com tanta eloqüência, que a maioria resolveu autorizá-lo ao que pedira, votando ao mesmo tempo um imposto destinado a subsidiar o tratamento, alojamento e mantimento dos doidos pobres. A matéria do imposto não foi fácil achá-la; tudo estava tributado em Itaguaí. Depois de longos estudos, assentou-se em permitir o uso de dois penachos nos cavalos dos enterros. Quem quisesse emplumar os cavalos de um coche mortuário pagaria dois tostões à Câmara, repetindo-se tantas vezes esta quantia quantas fossem as horas decorridas entre a do falecimento e a da última bênção na sepultura. O escrivão perdeu-se nos cálculos a itméticos do rendimento possível da nova taxa; e um dos vereadores, que não acreditava na empresa do médico, pediu que se relevasse o escrivão de um trabalho inútil.

Os cálculos não são precisos, disse ele, porque o Dr. Bacamarte não arranja nada. Quem é que viu agora meter todos os doidos dentro da mesma casa?

Enganava-se o digno magistrado; o médico arranjou tudo. Uma vez empossado da licença começou logo a construir a casa. Era na Rua Nova, a mais bela rua de Itaguaí naquele tempo; tinha cinqüenta janelas por lado, um pátio no centro, e numerosos cubículos para os hóspedes. Como fosse grande arabista,

achou no Corão que Maomé declara veneráveis os doidos, pela consideração de que Alá lhes tira o juízo para que não pequem. A idéia pareceu-lhe bonita e profunda, e ele a fez gravar no frontispício da casa; mas, como tinha medo ao vigário, e por tabela ao bispo, atribuiu o pensamento a Benedito VIII, merecendo com essa fraude aliás pia, que o Padre Lopes lhe contasse, ao almoço, a vida daquele pontífice eminente.

A Casa Verde foi o nome dado ao asilo, por alusão à cor das janelas, que pela primeira vez apareciam verdes em Itaguaí. Inaugurou-se com imensa pompa; de todas as vilas e povoações próximas, e até remotas, e da própria cidade do Rio de Janeiro, correu gente para assistir às cerimônias, que duraram sete dias. Muitos dementes já estavam recolhidos; e os parentes tiveram ocasião de ver o carinho paternal e a caridade cristã com que eles iam ser tratados. D. Evarista, contentíssima com a glória do marido, vestiu-se luxuosamente, cobriu-se de jóias, flores e sedas. Ela foi uma verdadeira rainha naqueles dias memoráveis; ninguém deixou de ir visitá-la duas e três vezes, apesar dos costumes caseiros e recatados do século, e não só a cortejavam como a louvavam; porquanto, e este fato é um documento altamente honroso para a sociedade do tempo, porquanto viam nela a feliz esposa de um alto espírito, de um varão ilustre, e, se lhe tinham inveja, era a santa e nobre inveja dos admiradores.

Ao cabo de sete dias expiraram as festas públicas; Itaguaí, tinha finalmente uma casa de orates.

Três dias depois, numa expansão íntima com o boticário Crispim Soares, desvendou o alienista o mistério do seu coração.

A caridade, Sr. Soares, entra decerto no meu procedimento, mas entra

como tempero, como o sal das coisas, que é assim que interpreto o dito de São Paulo aos Coríntios: "Se eu conhecer quanto se pode saber, e não tiver caridade, não sou nada". O principal nesta minha obra da Casa Verde é estudar profundamente a loucura, os seus diversos graus, classificar-lhe os casos, descobrir enfim a causa do fenômeno e o remédio universal. Este é o mistério do meu coração. Creio que com isto presto um bom serviço à humanidade.

Um excelente serviço, corrigiu o boticário.

Sem este asilo, continuou o alienista, pouco poderia fazer; ele dá-me, porém, muito maior campo aos meus estudos.

Muito maior, acrescentou o outro.

E tinha razão. De todas as vilas e arraiais vizinhos afluíam loucos à Casa Verde. Eram furiosos, eram mansos, eram monomaníacos, era toda a família dos deserdados do espírito. Ao cabo de quatro meses, a Casa Verde era uma povoação. Não bastaram os primeiros cubículos; mandou-se anexar uma galeria de mais trinta e sete. O Padre Lopes confessou que não imaginara a existência de tantos doidos no mundo, e menos ainda o inexplicável de alguns casos. Um, por exemplo, um rapaz bronco e vilão, que todos os dias, depois do almoço, fazia regularmente um discurso acadêmico, ornado de tropos, de antíteses, de apóstrofes, com seus recamos de grego e latim, e suas borlas de Cícero, Apuleio e Tertuliano. O vigário não queria acabar de crer. Quê! um rapaz que ele vira, três meses antes, jogando peteca na rua!

Não digo que não, respondia-lhe o alienista; mas a verdade é o que Vossa Reverendíssima está vendo. Isto é todos os dias.

Quanto a mim, tornou o vigário, só se pode explicar pela confusão das línguas na torre de Babel, segundo nos conta a Escritura; provavelmente, confundidas antigamente as línguas, é fácil trocá-las agora, desde que a razão não trabalhe...

Essa pode ser, com efeito, a explicação divina do fenômeno, concordou o alienista, depois de refletir um instante, mas não é impossível que haja também alguma razão humana, e puramente científica, e disso trato...

Vá que seja, e fico ansioso. Realmente!

Os loucos por amor eram três ou quatro, mas só dois espantavam pelo curioso do delírio. O primeiro, um Falcão, rapaz de vinte e cinco anos, supunha-se estrela-d alva, abria os braços e alargava as pernas, para dar-lhes certa feição de raios, e ficava assim horas esquecidas a perguntar se o sol já tinha saído para ele recolher-se. O outro andava sempre, sempre, sempre, à roda das salas ou do pátio, ao longo dos corredores, à procura do fim do mundo. Era um desgraçado, a quem a mulher deixou por seguir um peralvilho. Mal descobrira a fuga, armou-se de uma garrucha, e saiu-lhes no encalço; achou-os duas horas depois, ao pé de uma lagoa, matou-os a ambos com os maiores requintes de crueldade.

O ciúme satisfez-se, mas o vingado estava louco. E então começou aquela ânsia de ir ao fim do mundo à cata dos fugitivos.

A mania das grandezas tinha exemplares notáveis. O mais notável era um pobre-diabo, filho de um algibebe, que narrava às paredes (porque não olhava nunca para nenhuma pessoa) toda a sua genealogia, que era esta:

Deus engendrou um ovo, o ovo engendrou a espada, a espada engendrou Davi, Davi engendrou a púrpura, a púrpura engendrou o duque, o duque engendrou o marquês, o marquês engendrou o conde, que sou eu.

Dava uma pancada na testa, um estalo com os dedos, e repetia cinco, seis vezes seguidas:

Deus engendrou um ovo, o ovo, etc.

Outro da mesma espécie era um escrivão, que se vendia por mordomo do rei; outro era um boiadeiro de Minas, cuja mania era distribuir boiadas a toda a gente, dava trezentas cabeças a um, seiscentas a outro, mil e duzentas a outro, e não acabava mais. Não falo dos casos de monomania religiosa; apenas citarei um sujeito que, chamando-se João de Deus, dizia agora ser o deus João, e prometia o reino dos céus a quem o adorasse, e as penas do inferno aos outros; e depois desse, o licenciado Garcia, que não dizia nada, porque imaginava que no dia em que chegasse a proferir uma só palavra, todas as estrelas se despegariam do céu e abrasariam a terra; tal era o poder que recebera de Deus.

Assim o escrevia ele no papel que o alienista lhe mandava dar, menos por caridade do que por interesse científico.

Que, na verdade, a paciência do alienista era ainda mais extraordinária do que todas as manias hospedadas na Casa Verde; nada menos que assombrosa. Simão Bacamarte começou por organizar um pessoal de administração; e, aceitando essa idéia ao boticário Crispim Soares, aceitou-lhe também dois sobrinhos, a quem incumbiu da execução de um regimento que lhes deu, aprovado pela Câmara, da distribuição da comida e da roupa, e assim também da escrita, etc. Era

o melhor que podia fazer, para somente cuidar do seu ofício. A Casa Verde, disse ele ao vigário, é agora uma espécie de mundo, em que há o governo temporal e o governo espiritual. E o Padre Lopes ria deste pio trocado, e acrescentava, com o único fim de dizer também uma chalaça: Deixe estar, deixe estar, que hei de mandá-lo denunciar ao papa.

Uma vez desonerado da administração, o alienista procedeu a uma vasta classificação dos seus enfermos. Dividiu-os primeiramente em duas classes principais: os furiosos e os mansos; daí passou às subclasses, monomanias, delírios, alucinações diversas.

Isto feito, começou um estudo aturado e contínuo; analisava os hábitos de cada louco, as horas de acesso, as aversões, as simpatias, as palavras, os gestos, as tendências; inquiria da vida dos enfermos, profissão, costumes, circunstâncias da revelação mórbida, acidentes da infância e da mocidade, doenças de outra espécie, antecedentes na família, uma devassa, enfim, como a não faria o mais atilado corregedor. E cada dia notava uma observação nova, uma descoberta interessante, um fenômeno extraordinário. Ao mesmo tempo estudava o melhor regímen, as substâncias medicamentosas, os meios curativos e os meios paliativos, não só os que vinham nos seus amados árabes, como os que ele mesmo descobria, à força de sagacidade e paciência. Ora, todo esse trabalho levava-lhe o melhor e o mais do tempo. Mal dormia e mal comia; e, ainda comendo, era como se trabalhasse, porque ora interrogava um texto antigo, ora ruminava uma questão, e ia muitas vezes de um cabo a outro do jantar sem dizer uma só palavra a D. Evarista.

Ilustre dama, no fim de dois meses, achou-se a mais desgraçada das mulheres: caiu em profunda melancolia, ficou amarela, magra, comia pouco e suspirava a cada canto. Não ousava fazer-lhe nenhuma queixa ou reproche, porque respeitava

nele o seu marido e senhor, mas padecia calada, e definhava a olhos vistos. Um dia, ao jantar, como lhe perguntasse o marido o que é que tinha, respondeu tristemente que nada; depois atreveu-se um pouco, e foi ao ponto de dizer que se considerava tão viúva como dantes. E acrescentou:

Quem diria nunca que meia dúzia de lunáticos...

Não acabou a frase; ou antes, acabou-a levantando os olhos ao teto, os olhos, que eram a sua feição mais insinuante, negros, grandes, lavados de uma luz úmida, como os da aurora. Quanto ao gesto, era o mesmo que empregara no dia em que Simão Bacamarte a pediu em casamento. Não dizem as crônicas se D. Evarista brandiu aquela arma com o perverso intuito de degolar de uma vez a ciência, ou, pelo menos, decepar-lhe as mãos; mas a conjetura é verossímil. Em todo caso, o alienista não lhe atribuiu intenção. E não se irritou o grande homem, não ficou sequer consternado. O metal de seus olhos não deixou de ser o mesmo metal, duro, liso, eterno, nem a menor prega veio quebrar a superfície da fronte quieta como a água de Botafogo. Talvez um sorriso lhe descerrou os lábios, por entre os quais filtrou esta palavra macia como o óleo do Cântico:

Consinto que vás dar um passeio ao Rio de Janeiro.

D. Evarista sentiu faltar-lhe o chão debaixo dos pés. Nunca dos nuncas vira o Rio de Janeiro, que posto não fosse sequer uma pálida sombra do que hoje é, todavia era alguma coisa mais do que Itaguaí, Ver o Rio de Janeiro, para ela, equivalia ao sonho do hebreu cativo. Agora, principalmente, que o marido assentara de vez naquela povoação interior, agora é que ela perdera as últimas esperanças de respirar os ares da nossa boa cidade; e justamente agora é que ele a convidava a realizar os seus desejos de menina e moça. D. Evarista não pôde dissimular o gosto de semelhante proposta. Simão Bacamarte pagou-lhe na

mão e sorriu, um sorriso tanto ou quanto filosófico, além de conjugal, em que parecia traduzir-se este pensamento: "Não há remédio certo para as dores da alma; esta senhora definha, porque lhe parece que a não amo; dou-lhe o Rio de Janeiro, e consola-se". E porque era homem estudioso tomou nota da observação.

Mas um dardo atravessou o coração de D. Evarista. Conteve-se, entretanto; limitou-se a dizer ao marido que, se ele não ia, ela não iria também, porque não havia de meter-se sozinha pelas estradas.

Irá com sua tia, redargüiu o alienista.

Note-se que D. Evarista tinha pensado nisso mesmo; mas não quisera pedi-lo nem insinuá-lo, em primeiro lugar porque seria impor grandes despesas ao marido, em segundo lugar porque era melhor, mais metódico e racional que a proposta viesse dele.

Oh! mas o dinheiro que será preciso gastar! suspirou D. Evarista sem convicção.

Que importa? Temos ganho muito, disse o marido. Ainda ontem o escriturário prestou-me contas. Queres ver?

E levou-a aos livros. D. Evarista ficou deslumbrada. Era uma via-láctea de algarismos. E depois levou-a às arcas, onde estava o dinheiro.

Deus! eram montes de ouro, eram mil cruzados sobre mil cruzados, dobrões sobre dobrões; era a opulência.

Enquanto ela comia o ouro com os seus olhos negros, o alienista fitava-a,

e dizia-lhe ao ouvido com a mais pérfida das alusões:

Quem diria que meia dúzia de lunáticos...

D. Evarista compreendeu, sorriu e respondeu com muita resignação:

Deus sabe o que faz!

Três meses depois efetuava-se a jornada. D. Evarista, a tia, a mulher do boticário, um sobrinho deste, um padre que o alienista conhecera em Lisboa, e que de aventura achava-se em Itaguaí cinco ou seis pajens, quatro mucamas, tal foi a comitiva que a população viu dali sair em certa manhã do mês de maio. As despedidas foram tristes para todos, menos para o alienista. Conquanto as lágrimas de D. Evarista fossem abundantes e sinceras, não chegaram a abalá-lo. Homem de ciência, e só de ciência, nada o consternava fora da ciência; e se alguma coisa o preocupava naquela ocasião, se ele deixava correr pela multidão um olhar inquieto e policial, não era outra coisa mais do que a idéia de que algum demente podia achar-se ali misturado com a gente de juízo.

Adeus! soluçaram enfim as damas e o boticário.

E partiu a comitiva. Crispim Soares, ao tornar a casa, trazia os olhos entre as duas orelhas da besta ruana em que vinha montado; Simão Bacamarte alongava os seus pelo horizonte adiante, deixando ao cavalo a responsabilidade do regresso. Imagem vivaz do gênio e do vulgo! Um fita o presente, com todas as suas lágrimas e saudades, outro devassa o futuro com todas as suas auroras.

D Exemplo Corpus Infantil

Era uma vez, três irmãos que viviam no campo com sua mãe.
A mãe chamava-se Susana e era muito velha.
Os três levaram um saco de milho para o moinho para amassar e fazer papas.
Mas o vento, cada vez que iam ao moinho, levava sempre o milho.
Certo dia, os três irmãos foram à procura do vento e encontraram o vento num castelo.
Os três ralharam muito com o vento mas não tinha sido ele o culpado.
Como o vento teve pena deles deu-lhes uma toalha mágica.
Eles só precisavam de dizer: Toalha. põe a mesa e aparecia um banquente.
No caminho para casa, e depois de muito andarem, ficaram com sono e pararam numa estalagem.
O dono da estalagem perguntou-lhes o que é que eles queriam e eles disseram que queriam dormir mas que não tinha dinheiro,mas que podiam dar-lhe um banquete.
Então,colocaram a toalha na mesa e disseram as palavras mágicas.
Apareceu logo um banquete.
À noite, o estalajadeiro Diogo roubou a toalha e colocou uma igual, em vez da verdadeira.
Quando o Fernando, o Rui Pedro e o Carlos chegaram a casa, a toalha já não era mágica.
Resolverem ir procurar de novo o vento que lhes ofereceu um burro especial.
No dia seguinte, vinham a caminho de casa com um burro.
Tiveram sono e pararam na estalagem.
O estalajadeiro perguntou-lhes o que é que eles queriam.
Eles disseram que queriam dormir mas agora podiam pagar.
Disseram ao burro o seguinte:-Burro, burro cospe dinheiro.
E o burro cuspiu muito dinheiro.
À noite, o estalajadeiro Diogo roubou o burro e colocou outro igual no

mesmo sitio.

Quando chegaram a casa os três irmãos disseram ao burro para cuspir dinheiro mas o burro não cuspiu.

No dia seguinte, os três irmãos foram ao castelo do vento e o vento desta vez deu-lhes um pau mágico.

No caminho para casa os três pararam na estalagem e o estalajadeiro perguntou -lhes o que eles queriam.

Os três disseram que queriam as suas coisas e o estalajadeiro Diogo disse que não sabia das coisas deles.

Então os três irmãos disseram:- Pau, pau sai do saco e bate sem parar.

O pau saiu do saco e bateu na cabeça do Diogo.

O Diogo disse-lhes onde estava as coisas deles.

O Carlos, o Rui Pedro e o Fernando foram buscar as suas coisas,a toalha e o burro, e disseram ao pau para ele parar e o pau parou.

Chegeram a casa e os três mostraram a mãe o que eles conseguiam fazer.

A toalha deu um banquete e o burro cuspiu dinheiro.

Os quatro viveram felizes para sempre.