**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

**CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

# A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov

Amanda Rocha Chaves

**São Carlos**
**Agosto/2007**

# Universidade Federal de São Carlos

## Centro de Ciências Exatas e de Tecnologia

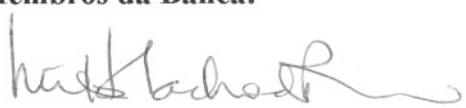
### Programa de Pós-Graduação em Ciência da Computação

# *"A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov"*

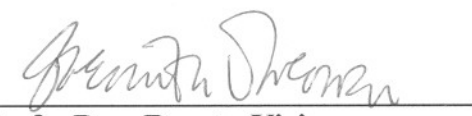
**AMANDA ROCHA CHAVES**

**Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.**

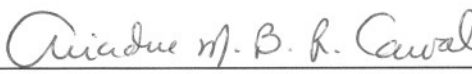
**Membros da Banca:**



**Profa. Dra. Lucia Helena Machado Rino**
**(Orientadora – DC/UFSCar)**



**Profa. Dra. Renata Vieira**
**(UNISINOS)**



**Profa. Dra. Ariadne Maria B. Rizzoni Carvalho**
**(UNICAMP)**

**São Carlos**
**Agosto/2007**

**Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar**

C512ra

Chaves, Amanda Rocha.
A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov / Amanda Rocha Chaves. -- São Carlos : UFSCar, 2007.
116 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Inteligência artificial. 2. Processamento da linguagem natural (Computação). 3. Resolução anafórica automática. I. Título.

CDD: 006.3 (20ª)

Existe em mim algo que não pode ser ferido nem enterrado,
algo que rompe rochedos...
a minha vontade.

*Nietzsche*

Nenhum computador tem consciência do que faz,
mas, na maior parte do tempo, nós também não.

*Marvin Minsky*

# Agradecimentos

À minha família, especialmente à minha mãe, pelo apoio incondicional, incentivo, carinho, dedicação, amor e por tudo que ainda tem feito para o meu crescimento, amadurecimento e aprendizado. A Arista, pelo amor, amizade, carinho e principalmente pela compreensão a mim dispensada. Pelo seu apoio, paciência e tolerância acima de tudo no primeiro ano do mestrado. À Rose, uma pessoinha especial, que segurou as pontas financeiras várias vezes durante esse período para que eu pudesse permanecer aqui trabalhando tranqüilamente.

À minha orientadora Profa. Lucia, pela confiança, pela orientação sem igual, pela dedicação e pelos valiosos direcionamentos e contribuições que tornaram possível a realização deste trabalho.

A todos os meus amigos, em especial à Val, que, como um anjo, apareceu em minha vida no primeiro dia em que cheguei nesta cidade: deu-me um lar, um ombro amigo e um emprego. Sem ela esse mestrado não teria se constituído. “Obrigada Valzinha, por tudo! Você é realmente muito especial.” Agradeço também ao meu amigão Marcos Laia, por toda a força e amizade, pelas caroninhas, pelos almoços juntos e, principalmente, pela paciência em me ouvir discursar sobre o meu trabalho.

Aos colegas do mestrado que se reuniam para estudar na primeira fase desse mestrado e aos amigos que conquistei nestes dois anos, especialmente a Lu, a minha companheirinha de casa e de estudos.

A todos que participaram da minha vida durante essa jornada, especialmente algumas funcionárias do DC que tornaram a minha estada aqui mais agradável, me fazendo dar boas risadas (Socorrinho).

# Resumo

Um dos problemas encontrados em sistemas de processamento de línguas naturais é conseguir manter a coesão referencial de um texto, propriedade que permite estabelecer as ligações entre os seus constituintes, tornando-o inteligível. Dentre os fatores de coesão referencial destacamos a anáfora, que ocorre quando duas ou mais expressões de um texto estabelecem uma relação de referência entre si, isto é, a interpretação da anáfora depende de um antecedente ao qual ela se refere no texto. Diversos algoritmos na literatura foram propostos para a resolução automática de anáforas pronominais, que consiste em: 1) identificar a anáfora, 2) determinar o conjunto de possíveis antecedentes e 3) identificar e selecionar o antecedente da anáfora. A ausência da resolução anafórica em aplicações como extração de informação, tradução automática e sumarização textual, dentre outras, pode levar à descontinuidade referencial de seus resultados, tornando-os não-coesos. Nesse contexto, apresentamos uma adaptação do algoritmo de Mitkov, originalmente aplicado no inglês, para resolver anáforas da língua portuguesa, especialmente as determinadas por pronomes pessoais de terceira pessoa cujo antecedente seja um sintagma nominal. Essa abordagem foi avaliada com base em corpora anotados com informações morfossintáticas e co-referenciais, utilizando-se como medida de avaliação de desempenho a taxa de sucesso, que determina o número de anáforas resolvidas corretamente pelo sistema automático em relação ao número de anáforas presentes no corpus avaliado. Além disso, fez-se uma comparação de desempenho entre essa abordagem e o algoritmo de Lappin & Leass adaptado para o português. Os resultados dessa avaliação são discutidos ao final do trabalho.

# Abstract

One of the problems of natural language processing systems is to assure referential cohesion in a text. This property allows connecting the text constituents and making it readable. We address the anaphoric phenomenon as one of the main factors of referential cohesion. Anaphors depict a reference relationship between two or more text components, and the interpretation of the anaphor is dependent upon the interpretation of its antecedent. This work is limited to pronominal anaphors, thus, to automatic pronoun resolution. Several algorithms have been proposed to this end. They usually involve (1) identifying the anaphoric component; (2) determining the set of its possible antecedents; and (3) identifying and selecting the most likely antecedent of the anaphor. The lack of anaphora resolution in, e.g., information extraction and automatic translation or summarization may yield non-cohesive texts. Herein we present an adaptation of the Mitkov´s algorithm for pronoun resolution. 3rd person pronouns for Brazilian Portuguese are especially addressed, whose antecedents are noun phrases. This approach has been intrinsically evaluated on annotated corpora. It has also been compared with Lappin and Leass algorithm for pronoun resolution, adapted to Portuguese. Annotations embed morphological, syntactic and co-referential information. The evaluation measure adopted was the success rate. This is defined as the ratio between the number of anaphors correctly resolved by the system and the total number of anaphors in the text. The results of both evaluations are discussed here.

# Índice de figuras

# Índice de tabelas

# Lista de siglas

API: *Application Program Interface*
CENTRO: Agente "Centro de Sentença"
DD: Descrição Definida
DR: Distância Referencial
DR_I: Distância Referencial Impeditiva
DR_P: Distância Referencial Promocional
ESG: *English Slot Grammar*
FDG: *Funcional Dependency Grammar*
IS: Instruções Seqüenciais
MARS: *Mitkov Anaphora Resolution System*
MMAX: *Multi-Modal Annotation in XML*
NP: Nome Próprio
NILC: Núcleo Interinstitucional de Lingüística Computacional
PADSUP: Agente de "Padrões de Superfície"
PC: Padrões de Colocação
PEG: *Parsing Expression Grammar*
PLN: Processamento de Línguas Naturais
POS: *Part-of-Speech*
PS: Paralelismo Sintático
PSN: Primeiro Sintagma Nominal da sentença
PSNTS: Preferência por SNs em Título de Seção
RA: Resolução Anafórica
RAP: *Resolution of Anaphora Procedure*
RAPM: Resolução Anafórica do Português baseada no algoritmo de Mitkov
RELPOS: Agente de "Relações de Posse"
RESRPP: Agente de "Resolução de RPPs"
RI: Referência Imediata
RL: Reiteração Lexical
RPP: Referência Pronominal Possessiva
SN: Sintagma Nominal
SNI: Sintagma Nominal Indefinido
SNMP: Sintagma Nominal Mais Próximo
SNP: Sintagma Nominal Preposicionado
TP: Termo Preferencial
VI: Verbos Indicativos

# Índice Geral

# Capítulo 1 - Introdução

Um dos problemas encontrados em sistemas de processamento de línguas naturais é conseguir manter a coesão referencial de um texto, propriedade esta que permite estabelecer as ligações entre seus constituintes, tornando-o inteligível. Dentre os fatores de coesão referencial destacamos a anáfora, que ocorre quando duas ou mais expressões de um texto estabelecem uma relação de referência entre si, sendo que sua interpretação depende do antecedente ao qual ela se refere no texto, conforme ilustra o exemplo seguinte.

(1.1) **O parlamentar**, porém, é alvo de acusação em outro escândalo. *Ele* será investigado sobre as denúncias de corrupção (...).

Nesse exemplo, o pronome 'Ele' é uma anáfora cujo antecedente é o termo 'O parlamentar', isto é, a anáfora só pode ser interpretada caso voltemos no texto e encontremos o termo ao qual ela se refere. Segundo Leffa:

> A anáfora pode ser descrita como um processo que acarreta uma volta no texto. O processo começa quando o anaforizante é conhecido (por exemplo, o pronome) e concluído quando o anaforizado é encontrado (a palavra a qual o pronome se refere). (Leffa, 2001:3)

O processo de determinação do antecedente correto ao qual a anáfora se refere é denominado resolução anafórica (RA). A complexidade dessa resolução reside em escolher o antecedente da anáfora quando muitos candidatos são possíveis, o que pode envolver o uso de conhecimento lingüístico diverso, englobando informações morfológicas, sintáticas, semânticas ou pragmáticas.

A resolução automática de anáforas, em geral, consiste nas seguintes etapas: 1) identificar a anáfora, 2) determinar o conjunto de possíveis antecedentes e 3) identificar e selecionar o antecedente da anáfora. O conhecimento heterogêneo necessário para a resolução vai ser utilizado em cada uma das etapas de forma diversa. Na etapa (2), por exemplo, esse conhecimento pode ser aplicado com base em heurísticas que consideram a concordância

morfológica entre anáfora e candidato, restrições sintáticas, como o paralelismo sintático, dentre outras.

Várias abordagens computacionais foram propostas para determinar o antecedente anafórico, utilizando-se estratégias e conhecimento lingüístico distintos, normalmente, aplicando-se fatores de resolução (Mitkov, 2002; 1997), que podem ser restritivos ou preferenciais. Os fatores restritivos definem propriedades que devem ser satisfeitas pelo candidato a antecedente, descartando aqueles candidatos que não os satisfazem. Os fatores preferenciais ordenam os candidatos restantes, que já passaram por filtros restritivos, determinando sua maior ou menor probabilidade de ser o antecedente da anáfora.

Este trabalho se restringe às anáforas pronominais. Há vários trabalhos relevantes para esta dissertação de mestrado que visam à resolução de tais anáforas, cujas propostas incluem a de Hobbs (1978), a de Palomar et al. (2001), a de Lappin & Leass (1994) e a de Dagan & Itai (1991). O algoritmo de Hobbs (1978) percorre a árvore sintática representativa das sentenças do texto em busca de um sintagma[1] nominal (SN) que concorde em gênero e número com o pronome anafórico. O algoritmo de Palomar et al. (2001) se baseia em restrições sintáticas e Lappin & Leass (1994) utilizam um sistema de pesos atribuídos de acordo com a estrutura sintática da sentença. Esses três algoritmos necessitam de conhecimento sintático para o seu processamento. Já o algoritmo de Dagan & Itai (1991) utiliza padrões de co-ocorrência extraídos automaticamente de um grande corpus, que servem para filtrar candidatos improváveis de serem o antecedente da anáfora. Com exceção do algoritmo de Palomar, utilizado para a resolução anafórica do espanhol, os demais algoritmos foram desenvolvidos para o processamento da língua inglesa. Para a língua portuguesa temos poucas abordagens, dentre elas, uma proposta de adaptação do algoritmo de Lappin & Leass (Coelho, 2005), uma adaptação do algoritmo de Hobbs (Santos & Carvalho, 2007) e uma implementação multi-agentes (Paraboni, 1997). Essas propostas utilizam uma abordagem baseada em conhecimento lingüístico: Coelho (2005) e Santos & Carvalho (2007), por exemplo, utilizam a árvore sintática do texto, enquanto Paraboni (1997) utiliza conhecimentos que variam do nível morfossintático ao pragmático.

[1] Um sintagma é o resultado da combinação de um determinante e de um determinado (núcleo do sintagma) numa unidade lingüística hierarquizada. O nome do sintagma depende da classe gramatical a qual pertence o seu núcleo. Temos, portanto, os seguintes sintagmas: nominal (núcleo: nome ou substantivo, ex.: o **menino**), verbal (núcleo: verbo), preposicional (núcleo: preposição), e assim por diante. (Coelho et al., 2005; Coelho, 2005; Koch, 1996)

A proposta de Mitkov (2002) para a resolução de pronomes anafóricos é mais interessante que aquelas exploradas para o português, pois evita o uso de conhecimento sintático complexo, semântico ou pragmático, baseando-se apenas na aplicação de heurísticas, denominadas 'indicadores de antecedentes', que associadas, apontam o antecedente da anáfora. Como exemplo dessas heurísticas tem-se a preferência pelo primeiro sintagma nominal da sentença, por sintagmas nominais topicalizados ou repetidos, determinados, não-preposicionados, acompanhados de determinados verbos, etc.

Este trabalho apresenta uma proposta de resolução anafórica com base no algoritmo de Mitkov, para a língua portuguesa. Esse algoritmo resolve somente pronomes de terceira pessoa cujo antecedente é um sintagma nominal. Sua escolha deve-se, primeiramente, ao fato dele ainda não ter sido implementado para o português e do seu funcionamento ser considerado simples. Além disso, o algoritmo foi implementado para outras línguas, como o árabe e o polonês, e sua avaliação apresentou bons resultados, o que demonstra sua portabilidade de língua, como será visto no Capítulo 4.

Esta dissertação está organizada em sete capítulos: no Capítulo 2 é apresentado o fenômeno da referenciação, os tipos de anáforas existentes, os problemas encontrados na sua resolução automática, as heurísticas utilizadas no processo de resolução anafórica e algumas abordagens que visam à implementação de algoritmos que identifiquem o antecedente da anáfora. O Capítulo 3 apresenta os trabalhos desenvolvidos no Brasil para a resolução anafórica. O Capítulo 4 apresenta a abordagem original do algoritmo de Mitkov, bem como um exemplo que ilustra a sua execução, a caracterização da sua natureza multilíngüe e a abordagem atual desse algoritmo, que faz parte de um sistema totalmente automático de resolução anafórica. O Capítulo 5 apresenta a proposta deste trabalho, seguida de um estudo de caso que permitiu determinar as heurísticas a serem implementadas para o português dentre as propostas por Mitkov. A avaliação dessa proposta de trabalho é mostrada no Capítulo 6, e o Capítulo 7 descreve algumas conclusões sobre este trabalho e suas perspectivas futuras.

# Capítulo 2 - Resolução Anafórica

Neste capítulo definimos o fenômeno da referenciação e sua relação direta com a coesão textual. Particularmente em foco estarão as relações anafóricas, que são o objeto de estudo desse trabalho. Descrevemos a diferença entre anáfora e co-referência, os problemas encontrados na resolução anafórica e os fatores de resolução geralmente utilizados na identificação do antecedente da anáfora. Ao final do capítulo são apresentadas algumas abordagens relacionadas ao processo de resolução anafórica automática.

## 2.1 - O fenômeno da referenciação

A coesão textual é definida por Halliday & Hasan (1976) como um conceito semântico que se refere às relações de sentido existentes no interior do texto e que o definem como 'texto'. O texto é considerado aqui como um construto da língua falada ou escrita de tamanho variado que constitui um todo unificado, composto de sentido e intenções próprias.

A coesão ocorre quando a interpretação de algum elemento *i* no discurso é dependente da interpretação de um outro elemento *j* no mesmo discurso. Diz-se que o elemento *i* pressupõe o elemento *j*, e *i* não pode ser efetivamente decodificado a não ser por recurso de *j* (Halliday & Hasan, 1976). A coesão é uma relação semântica entre um elemento do texto e algum outro elemento essencial para a sua interpretação. Ela estabelece relações de sentido que realizam a ligação entre os constituintes do texto (uma sentença se liga à anterior) através de recursos ou elos coesivos (Koch, 1994).

Halliday & Hasan (1976) distinguem cinco mecanismos tidos como principais recursos de coesão na construção de um texto: a referência, a substituição, a elipse, a conjunção e a coesão lexical. Este trabalho focaliza apenas o recurso coesivo referência.

Por referência entende-se "os elementos da língua que não podem ser interpretados semanticamente por si mesmos, mas remetem a outros itens do discurso necessários à sua interpretação" (Koch, 1994). Aos primeiros elementos dá-se o nome de

formas referenciais; aos últimos, referentes textuais. Estes podem ser representados por um sintagma, um fragmento de oração, uma oração ou todo um enunciado.

A referência pode ser i) situacional (exofórica), quando o referente é representado por algum elemento da situação comunicativa, isto é, quando o antecedente está fora do texto; ou ii) textual (endofórica), quando o referente se encontra expresso no próprio texto. Na sentença (2.1) a seguir, verifica-se uma referência exofórica representada pela palavra 'você', a qual só será interpretada corretamente caso se conheça o contexto situacional no qual essa frase ocorre.

(2.1) Por que **você** não fez a tarefa?

A referência endofórica pode estar implícita ou explícita no texto e pode ser expressa por uma catáfora ou uma anáfora. Pode, ainda, ser omitida, como no caso de elipse.

A catáfora ocorre quando o referente se encontra após o item coesivo, como na sentença (2.2), na qual o referente 'o meu marido' aparece posposto ao item coesivo representado pelo pronome 'ele'.

(2.2) *Ele*$_i$ era tão bom, **o meu marido**$_i$!

Já a anáfora ocorre quando o referente precede o item coesivo, como mostrado nas sentenças ($s_2$ e $s_4$) do texto (2.3)[2]. Neste, o referente 'Milton Nascimento' aparece antes dos itens coesivos (anáforas) 'sua', 'o artista brasileiro' e 'ele'.

(2.3) **Milton Nascimento$_i$** vive uma fase feliz$_{(s_1)}$. *Sua*$_i$ carreira ganhou, nos últimos anos, impulso internacional$_{(s_2)}$ (...). (Φ)$_i$ É amigo de Sting e das maiores feras do jazz do mundo$_{(s_3)}$. (...) dificilmente o *artista brasileiro*$_i$ mostra lá fora o que *ele*$_i$ é aqui$_{(s_4)}$.[3]

A elipse ocorre quando há omissão de algum constituinte da sentença, recuperável pelo contexto, como na sentença ($s_3$) do texto (2.3), no qual o termo elíptico está representado pelo símbolo (Φ) e se refere também ao antecedente 'Milton Nascimento'.

---

[2] Texto extraído do Correio de Domingo (Koch & Travaglia, 1996: 18), aqui transcrito somente com ênfase nas cadeias referenciais.

[3] Índices iguais indicam co-referência; $s_i$: sentença i; (Φ): representa uma referência elíptica (neste caso, o sujeito da sentença $s_3$ está omisso e, portanto, se refere ao antecedente Milton Nascimento).

A Figura 1 sintetiza a classificação das referências apresentadas.

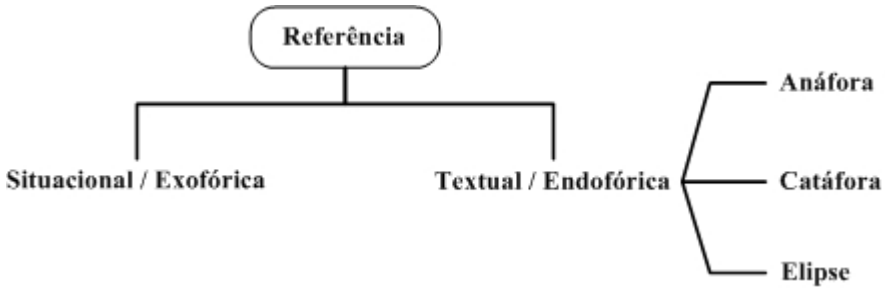



Figura 1: Classificação das referências

Dentre as referências dessa classificação, a referência endofórica, mais precisamente, as anáforas constituem o objeto de investigação dessa dissertação de mestrado e serão discutidas a seguir.

## 2.2 - Anáfora e co-referência

Em uma relação anafórica, o referente e o item coesivo são denominados, nesta ordem, antecedente e anáfora. Quando a anáfora e seu antecedente remetem a uma mesma referência no mundo real, dizemos que eles são co-referentes, e quando várias anáforas se referem a um mesmo antecedente, elas formam a chamada cadeia anafórica.

Os conceitos de co-referência e anáfora são similares, mas possuem pequenas diferenças: em uma relação anafórica, a anáfora pode fazer referência à mesma entidade introduzida por seu antecedente (neste caso, a anáfora e o antecedente são co-referentes) ou pode ativar um novo referente e estar ligada ao antecedente através de alguma relação semântico-discursiva (neste caso, eles não são co-referentes).

Exemplos[4]:

(2.4) Co-referência: Comprei três **livros**$_i$ excelentes. *Os livros*$_i$ estão lá em casa.

(2.5) Não co-referência[5]: Entrei no **restaurante**$_i$ e *o garçom*$_j$ veio me atender.

Mitkov (2002) classifica as anáforas da seguinte maneira:

a) <u>De acordo com a sua forma</u>: i) anáfora nominal – ocorre quando a expressão referencial (pronome, sintagma nominal definido[6] ou nome próprio)

[4] Extraídos de (Haag & Othero, 2003: 3).
[5] Índices diferentes indicam que os sintagmas nominais que os precedem não são co-referentes.

tem como antecedente um sintagma nominal não-pronominal; ii) anáforas verbal e adverbial e iii) anáfora elíptica.

Mitkov verificou, em suas pesquisas sobre o fenômeno de referenciação pronominal na língua inglesa, que os pronomes em primeira e segunda pessoa ocorrem, na maioria das vezes, de maneira dêitica, portanto não-anafórica. O pronome '*it*', freqüentemente, é não-anafórico ou, como denominaram Lappin & Leass (1994), pronome pleonástico. Destacamos que esse fenômeno não ocorre na língua portuguesa e, portanto, não é tratado neste trabalho.

b) De acordo com a localização da anáfora e do antecedente: i) anáfora intra-sentencial, quando ambos ocorrem na mesma sentença e ii) inter-sentencial, caso contrário. Os pronomes reflexivos e possessivos são tipicamente intra-sentenciais.

Mitkov (1998) afirma que a localização do antecedente de uma anáfora também depende do seu tipo (sintagma nominal, pronome, etc.). Ele sugere, baseando-se em estudo de corpus, que a distância entre uma anáfora pronominal e seu antecedente não deve exceder duas ou três sentenças anteriores à sentença da anáfora, enquanto descrições definidas (DD) anafóricas envolvem uma busca em até dez sentenças precedentes.

Na seção seguinte são apresentados os problemas lingüísticos e computacionais envolvidos no processo de identificação do antecedente e as etapas que descrevem esse processo, considerando a sua complexidade e os fatores de resolução geralmente utilizados.

## 2.3 - A resolução automática de anáforas

Na construção de um discurso minimamente coeso, o escritor geralmente utiliza, apropriadamente, recursos como o encadeamento referencial, visando fornecer aos leitores/ouvintes informações suficientes para que cada termo anafórico possa ser interpretado de maneira unívoca. Para um sistema automático de resolução anafórica, que basicamente busca, dentre os termos já introduzidos no texto, o antecedente correto da anáfora, a interpretação desse termo anafórico não é tão trivial quanto possa ser para um leitor, pois para isso, tal sistema deverá recorrer a tipos de conhecimento distintos (morfológico, sintático,

[6] Denominado por Russell (1905) de descrição definida, esta é composta por um núcleo (substantivo) e iniciada por um artigo definido (o, a, os, as). Muitas vezes o núcleo vem acompanhado de elementos complementares, como adjetivos, etc.. Exemplos: o presidente, o professor de inglês, a garota bonita. Vide (Vieira, 1998; Vieira & Poesio, 2000; Poesio et al., 2005) para uma pesquisa detalhada sobre processamento de descrições definidas.

semântico ou até mesmo pragmático) sobre a língua natural, que, por si só, acarretam problemas computacionais expressivos.

A sentença (2.6)[7] ilustra um exemplo em que o antecedente da anáfora pode ser determinado de forma unívoca pela dependência morfológica de gênero e número. Assim, nesse exemplo, a anáfora 'ela' tem como único antecedente o termo 'Maria'.

(2.6) **João** disse à **Maria**$_i$ que *ela*$_i$ era bonita.

Diferentemente, na sentença (2.7)[8], a concordância de gênero e número não é suficiente para identificar o antecedente correto, já que os candidatos a antecedente 'casas' e 'pessoas' possuem as mesmas características morfológicas que a anáfora. A opção aqui poderia ser a aplicação de restrições em função da estrutura sintática da sentença. Podemos argumentar que existe um paralelismo sintático entre o substantivo 'casas' e o pronome 'elas', isto é, 'casas' e 'elas' estão na posição de sujeito em suas próprias orações. Já o substantivo 'pessoas' se encontra na posição de complemento verbal, não compartilhando, com o termo 'elas', de um paralelismo. Assim, o antecedente escolhido será o substantivo 'casas'.

(2.7) **Casas**$_i$ são compradas por **pessoas** porque *elas*$_i$ oferecem conforto.

A mudança do verbo 'oferecer', presente na sentença (2.7), para o verbo 'gostar', como em (2.8), introduz um novo problema para a RA, demonstrando também que alguns verbos determinam a escolha dos antecedentes. Nessa nova sentença, o conhecimento sintático já não é suficiente para resolver a anáfora: se aplicássemos o paralelismo sintático, encontraríamos como antecedente o termo 'casas', que com certeza é uma escolha incorreta. Nesse exemplo necessitamos de conhecimento semântico para encontrar o antecedente. A anáfora e o antecedente devem compartilhar o mesmo traço semântico indicado pelo verbo 'gostar': o traço '+ animado'. Assim, o antecedente escolhido será 'pessoas'.

(2.8) **Casas** são compradas por **pessoas**$_i$ porque *elas*$_i$ gostam de conforto.

Muitas vezes as restrições morfológicas, sintáticas e semânticas ainda não são suficientes para resolver a anáfora, sendo necessário o conhecimento pragmático ou situacional, como na sentença (2.9), em que o termo 'ela' tem como antecedente o substantivo 'caixa' e não 'corda'. Esta escolha depende da compreensão da relação de causa e

[7] Sentenças 2.6 e 2.8 (Paraboni, 1997: 12-13).
[8] Sentenças 2.7 e 2.9 (Leffa, 2001: 3-4).

conseqüência que se estabelece entre ‘cortar a corda’, e ‘cair a caixa que está pendurada à corda’, respectivamente.

(2.9) João cortou a **corda** que suspendia a **caixa**$_i$ e *ela*$_i$ caiu.

Existem diversas abordagens computacionais que visam à identificação do antecedente do termo anafórico, utilizando-se estratégias de conhecimento lingüístico distintos, as quais Mitkov (2002) denominou ‘fatores de resolução anafórica’, termo que será utilizado como este propõe. Muitas dessas abordagens dão suporte a aplicações de Processamento de Língua Natural (PLN) tais como sumarização automática de textos, extração e recuperação de informação e tradução automática, com o intuito de tornar os seus resultados mais coesos, tornando-os, portanto, mais inteligíveis. Em sistemas de sumarização automática, por exemplo, a ausência de resolução anafórica pode causar a descontinuidade referencial do sumário, levando-o a transmitir uma mensagem discordante com a do seu texto-fonte, ou, no caso extremo, impedindo a compreensão do sumário (Rino & Seno, 2006).

A seguir, discorremos sobre os três passos principais da RA automática.

### *2.3.1 - A identificação da anáfora*

A identificação do termo anafórico pode depender de recursos computacionais como analisadores morfológicos, etiquetadores, *parsers*, dentre outros, na identificação do termo lingüístico. Dependendo do tipo de anáfora (pronome, sintagma, etc.) é necessária ainda uma classificação mais refinada, sendo que classificações diversas são propostas por diferentes autores (por exemplo: Vieira, 1998; Mitkov, 2002).

Muñoz (2001), por exemplo, propõe um método de classificação de descrições definidas (DD) em anafóricas e não anafóricas baseando-se na geração de uma rede semântica com o auxílio da *WordNet* (Miller & Fellbaum, 1992) para o espanhol. Esse método consiste nos passos seguintes: para cada DD encontrada no texto, uma lista de possíveis antecedentes é produzida, consistindo de todos os sintagmas nominais (SNs) que a precedem. O SN que tem seu núcleo diferente da DD e que não possui compatibilidade semântica (sinonímia, hiperonímia, hiponímia) com a mesma é considerado como não anafórico. Além disso, verifica-se se os modificadores do núcleo da DD e dos candidatos são compatíveis (ex. termos anafóricos não possuem relação de antonímia). Depois, um módulo de desambiguação do sentido da palavra é usado para obter o correto significado dos núcleos, enquanto nomes

próprios são recuperados como potencialmente anafóricos quando antecedem outros nomes próprios que 'casam' em termos de primeiro e últimos nomes.

Assim como Muñoz desenvolveu um sistema de classificação de DD, Vieira & Poesio (2000) desenvolveram um sistema de tratamento de co-referência para a língua inglesa considerando que a anáfora é uma DD e sua identificação se dá por meio de sua classificação em duas grandes classes: a) descrições definidas novas no discurso e b) descrições definidas anafóricas.

Uma DD nova no discurso é aquela que introduz um novo referente no texto não relacionado a nenhum antecedente do discurso, ou seja, não tem uma âncora em que possa se apoiar semanticamente (Vieira, 1998). Ela ocorre, geralmente, no início do texto ou com SNs seguidos de sintagmas preposicionais. O processo de classificação proposto por Vieira & Poesio será detalhado na Seção 2.4.5. A implementação de um sistema de resolução de co-referência para a língua portuguesa, fundamentada nesse processo de classificação, é apresentada em Vieira (2001), Rossi, et al. (2001), Coelho et al. (2005) e Collovini et al. (2005).

Autores como Lappin & Leass (1994) e Evans (2001), diferentemente dos autores que abordam as DDs, propõem a identificação de anáforas pronominais. Para a língua inglesa, por exemplo, todo sistema de resolução deve conter um módulo que reconheça ocorrências não-anafóricas do pronome *it*. Lappin & Leass desenvolveram um módulo de identificação desse pronome no qual ele é classificado como pleonástico caso a estrutura da sentença onde o mesmo ocorre coincida com um dos padrões de estruturas sentenciais apontados por eles. Como exemplo, o padrão estrutural '***It is thanks to*** SN ***that*** S' pode ser verificado na sentença (2.10), o que torna o pronome *it* semanticamente vazio, ou seja, um pronome sem antecedente. Esse pronome é então descartado pelo sistema.

(2.10) ***It is thanks to*** *your dedication* ***that*** *our team won the race.*

Evans (2001) também descreve uma abordagem para identificação do pronome pleonástico *it;* contudo, ele acrescenta um módulo que identifica ocorrências não-nominais do *it*, isto é, suas instâncias que indicam um antecedente constituído por orações, sentenças, etc. Assim, uma ocorrência de *it* é representada como um vetor de 35 'traços' (por exemplo, localização do pronome, características relacionadas a elementos circunvizinhos do pronome, etc.) que classifica o pronome em pleonástico, não-nominal ou sintagma nominal anafórico.

### *2.3.2 - Determinação do conjunto de candidatos a antecedentes*

Uma vez identificada a anáfora, o sistema de RA deverá agora detectar o conjunto dos possíveis candidatos a antecedente. A maioria dos sistemas propõe apenas a identificação de anáforas nominais, "pois o processamento de anáforas cujo antecedente é uma cláusula, sentença ou seqüência de sentenças é uma tarefa bem mais complicada" (Mitkov, 2002: 39). Geralmente, todos os sintagmas nominais que precedem a anáfora e que estão dentro de um 'escopo de busca' são inicialmente considerados como candidatos a antecedente. Esse escopo depende do tipo de anáfora e do modelo adotado para a RA.

Usualmente, para a RA pronominal na língua inglesa, costuma-se limitar o escopo de busca a duas ou três sentenças anteriores à sentença onde ocorre a anáfora (Mitkov, 1998; Mitkov, 2002); para as descrições definidas anafóricas o escopo é maior, cerca de dez sentenças (Kameyama, 1997).

### *2.3.3 - A identificação do antecedente*

Após encontrar a anáfora e identificar o conjunto de candidatos possíveis a antecedente, o sistema de RA deverá agora identificar o antecedente, selecionando-o a partir desse conjunto de candidatos. As regras de resolução empregadas nessa identificação baseiam-se em diferentes tipos de conhecimento. Neste trabalho, considera-se somente aqueles fatores discriminados por Mitkov (2002), que são classificados em dois tipos, nomeados por ele de 'fatores de resolução': os restritivos e os preferenciais. Os fatores restritivos visam descartar candidatos do conjunto de candidatos a antecedente, e os preferenciais somente apontam o antecedente preferencial.

Uma restrição define uma propriedade que deve ser satisfeita por algum candidato a antecedente para que o mesmo possa ser considerado uma possível solução da anáfora. Por exemplo, anáforas pronominais e antecedentes, geralmente, devem concordar em pessoa, gênero e número[9]. Caso a restrição não seja satisfeita, o candidato é desconsiderado. Restrições geralmente utilizadas na resolução de anáforas são: concordância morfológica, restrições c-comando (*constituent-command*) (Reinhart, 1983) e restrições semânticas. Estas restrições são detalhadas abaixo.

[9] Esta regra não se aplica a todos os pronomes. Por exemplo, pronomes possessivos da língua portuguesa concordam em gênero e número com a coisa possuída.

Ao contrário das restrições, que servem como filtros para eliminar os candidatos que não as satisfazem, as regras preferenciais propõem um candidato a antecedente. Uma preferência é a característica que nem sempre é atendida pelo candidato a antecedente. A aplicação de preferências, usualmente, envolve o uso de heurísticas para se obter uma lista de candidatos ordenados, favorecendo aqueles que as satisfazem. As preferências geralmente usadas são: paralelismo sintático e semântico, preferência por sujeito, saliência, proximidade, etc.

Os tópicos seguintes apresentam com detalhes os tipos de restrições ($R_i$) e preferências ($P_i$) utilizados no processo de RA e exibem alguns exemplos representativos, extraídos de Mitkov (2002) e adaptados para o português.

**Restrições:**

$R_1$ - Concordância de gênero e número: esta restrição requer que a anáfora e seu antecedente concordem em gênero e número (restrição morfológica).

$R_2$ - Restrição c-comando: essa restrição se baseia na relação estrutural c-comando entre a anáfora e o antecedente. A restrição c-comando é um tipo de relação que é estabelecida entre os nós de uma árvore sintática, e é análoga à idéia de 'irmãos e todos os seus descendentes', em uma árvore genealógica. Esta relação é definida pelo conceito de dominância, que diz que um nó A domina um nó B se e somente se o nó A se encontra em uma posição superior à posição de B na árvore (por exemplo: A poderia ser o pai do filho B) e se, a partir de A, pode-se traçar uma linha descendente até B. Essa relação está fundamentada no fato da estrutura sintática da sentença impor restrições à relação de co-referência entre a anáfora e o antecedente (Reinhart, 1983). Como exemplo de restrição c-comando temos: a co-referência é proibida caso a anáfora c-comande o antecedente. A definição formal de c-comando é apresentada da seguinte maneira:

O nó A c-comanda um nó B se e somente se: i) A não domina B, ii) B não domina A e iii) o primeiro nó com ramificação x que domina A também domina B.

Na Figura 2 (Mitkov, 2002: 59) ilustramos a relação de c-comando entre os constituintes de um texto.

a) B c-comanda C e todo nó que C domina.
b) C c-comanda B e todo nó que B domina.
c) D c-comanda E e J, mas não C, ou algum dos nós que C domina.
d) H c-comanda I e nenhum outro nó.

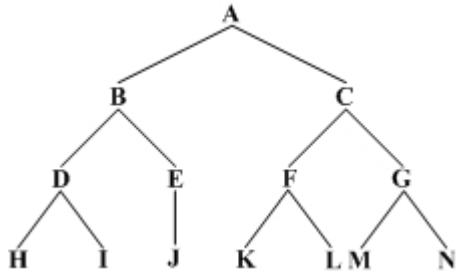



Figura 2: Árvore sintática.

Como ilustração, para a sentença (2.11) será aplicada a restrição c-comando: a co-referência é proibida caso o pronome c-comande o sintagma nominal.

(2.11) *Ela* quase procurou **Hera** para saber do assunto.

Neste exemplo, o pronome 'ela' c-comanda o sintagma nominal 'Hera' e, portanto, a co-referência entre eles é proibida.

$R_3$ - Restrição semântica: esta restrição exige que o antecedente satisfaça às mesmas restrições impostas à anáfora. Particularmente, se uma anáfora participa de uma relação sintática (funcionando como sujeito ou objeto de um verbo), então a substituição da anáfora pelo antecedente será possível desde que este satisfaça a restrição semântica estipulada pelo verbo. Na sentença (2.12), o antecedente deve ser um objeto que pode ser desconectado (o computador, mas não o disco), enquanto em (2.13), o antecedente deve ser um objeto que pode ser copiado (o disco, mas não o computador).

(2.12) George removeu **o disco** do **computador**$_i$ e em seguida desconectou-*o*$_i$.

(2.13) George removeu o **disco**$_i$ d**o computador** e em seguida copiou-*o*$_i$.

**Preferências:**

$P_1$ - Paralelismo sintático: esta preferência é dada para o SN que tem a mesma função sintática que a anáfora.

$P_2$ - Saliência: preferência pela entidade central do discurso, assim o candidato central do discurso é identificado como antecedente. Essa preferência é fundamentada na *Centering Theory* (Grosz et al., 1995), que estabelece um sistema de regras e restrições que governam as relações entre o tema do discurso e algumas escolhas lingüísticas efetuadas pelos participantes do discurso, como por exemplo, o emprego de pronomes. Essas regras

determinam que o centro da própria sentença ou centros das sentenças anteriores são candidatos altamente prováveis a termo antecedente.

$P_3$ - Preferência pelo sujeito: algumas abordagens dão preferência ao candidato que é o sujeito da sentença.

Os fatores de resolução mais comuns detalhados anteriormente não são os únicos. Outros fatores, principalmente os preferenciais, serão mencionados no momento oportuno.

As ferramentas geralmente utilizadas pela maioria dos sistemas de RA nas três etapas de RA descritas anteriormente são:

a) Analisador léxico-morfológico: decompõe a sentença em itens lexicais e realiza uma varredura, tratando item a item, decompondo-os em seus morfemas (Vieira & Lima, 2001). Além disso, contém informações úteis para análise morfológica tais como gênero, número, etc. e significado das palavras.

b) Etiquetador (POS, ou *part-of-speech*, *tagger*): é uma ferramenta baseadas em corpus que têm como finalidade identificar a categoria gramatical correta de cada palavra na sentença.

c) *Parser*: analisador sintático de sentenças. Utiliza conhecimento sobre as palavras e seus significados através de léxicos e de uma gramática (conjunto de regras de composição de sentenças).

d) Extrator de sintagmas nominais: é considerado um *parser* simples, pois não é designado para construir a árvore sintática completa de sentenças, mas apenas determinar pequenos constituintes como sintagmas nominais ou sintagmas preposicionais; é também chamado *shallow parser* (Mitkov, 2002) ou *chunker*.

Dentre essas ferramentas, o conhecimento necessário para identificação das anáforas e dos candidatos a antecedente pode ser fornecido pelos analisadores léxico-morfológicos, etiquetadores e extratores de SNs. Para a implementação dos fatores de RA, como por exemplo, para o filtro morfológico, é necessário apenas o uso do analisador léxico-morfológico; as restrições c-comando necessitam da estrutura arbórea da sentença e, portanto, requerem um *parser*; e o conhecimento semântico pode ser provido pela *WordNet* (Miller & Fellbaum, 1992).

Na próxima seção são descritas algumas das principais abordagens de resolução anafórica para a língua inglesa e espanhola, em especial, aquelas envolvidas com a resolução de anáforas pronominais.

## 2.4 - Abordagens de resolução anafórica

Esta seção apresenta algumas abordagens para resolução de anáforas pronominais, anáforas determinadas por descrições definidas e anáforas associativas, priorizando os trabalhos que resolvem anáforas pronominais da língua inglesa. Alguns modelos ricos em conhecimento lingüístico diferem das abordagens de Mitkov (2002) e da proposta apresentada na Seção 2.4.2.

### *2.4.1 - Abordagem puramente sintática*

Hobbs (1978) propõe uma abordagem para resolução de pronomes baseada na estrutura sintática correta da sentença no texto. O algoritmo de Hobbs busca na árvore sintática um SN com mesmo gênero e número que o pronome.

A busca pelo antecedente da anáfora se inicia na sentença na qual ela ocorre, isto é, no nó da árvore sintática que a representa. O algoritmo faz uma busca em largura, da esquerda para a direita, subindo em direção à raiz da árvore. Caso não seja encontrado nenhum antecedente intra-sentencial, a busca continua na árvore sintática da sentença anterior a partir da raiz, em largura, da esquerda para a direita.

Hobbs avaliou seu algoritmo sobre um conjunto de 300 pronomes obtido de três textos diferentes, sendo que de cada texto foram processados 100 pronomes: um texto literário, um texto técnico de arqueologia e outro jornalístico. Os pronomes considerados foram *he*, *she*, *it* e *they*. Ele obteve um taxa de sucesso de 88,3%, considerando que o algoritmo foi simulado manualmente utilizando uma estrutura sintática perfeita dos textos. Em um resolvedor anafórico automático, muitos erros podem ser inseridos devido, por exemplo, à incorreta análise sintática, etiquetagem e identificação de nomes próprios realizada automaticamente, degradando essa taxa de sucesso. A taxa de sucesso definida por Mitkov (2002) é uma medida que reflete o sucesso de resolução de um algoritmo frente a todas as anáforas no corpus avaliado, que foram identificadas e anotadas por um especialista humano. Ela normalmente é expressa por um valor percentual e é calculada dividindo-se o número total de anáforas resolvidas corretamente (AC) pelo número total de anáforas presentes no texto (T).

$$\text{Taxa de sucesso} = \frac{\mathbf{AC}}{\mathbf{T}}$$

Segundo Mitkov, essa medida de avaliação focaliza a performance do processo de RA e não os módulos de pré-processamento que geram os arquivos de entrada para a RA. Ele salienta que o valor exato da taxa de sucesso só pode ser obtido, caso os dados de entrada do processo de RA estejam corretos, isto é, os arquivos de entrada devem ser pós-editados por humanos ou extraídos de um corpus já etiquetado corretamente.

### *2.4.2 - Padrões de colocação*

Dagan & Itai (1991) descrevem uma abordagem para resolução de pronomes pessoais de terceira pessoa da língua inglesa baseada em padrões de agrupamento, também chamados 'padrões de co-ocorrência' ou 'colocação'. Esses padrões são coletados automaticamente de um grande corpus e são usados para filtrar diferentes candidatos a antecedentes. Após encontrar o conjunto de candidatos, inicia-se o processo de RA propriamente dito. A anáfora é substituída por cada candidato identificado e aquele que produzir o padrão de co-ocorrência mais freqüente é adotado como antecedente.

O modelo proposto compreende, assim, duas fases: 1) fase de aquisição, em que um corpus é processado para construir uma base de dados estatística e 2) fase de desambiguação, em que a base de dados é utilizada para resolver as anáforas que estejam na terceira pessoa. Na fase de aquisição, a base de dados é criada com padrões de co-ocorrência para os seguintes pares de relações sintáticas: sujeito-verbo, verbo-objeto e adjetivo-substantivo. A identificação das funções sintáticas das sentenças é feita pela gramática *PEG – Parsing Expression Grammar* (Jensen, 1986).

Como método de avaliação os autores realizaram um experimento com o corpus *Hansard*[10] para resolução do pronome anafórico *it* cujo antecedente esteja localizado na mesma sentença que a anáfora. Esse corpus é constituído de processos do parlamento canadense, totalizando 85 milhões de palavras. Os autores obtiveram para essa resolução uma taxa de sucesso de 87%. Ressalta-se que para tal experimento, uma filtragem manual foi realizada pelos autores com o intuito de descartar os pronomes pleonásticos, os pronomes cujo antecedente não era um SN e aqueles pronomes que não se encaixavam em uma das relações sintáticas citadas acima. Além disso, as anáforas com apenas um candidato a antecedente foram desconsideradas. Essa filtragem manual contribuiu consideravelmente para o alto desempenho dessa proposta.

[10] <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20>

A fim de explorar a utilidade de suas estatísticas, Dagan & Itai (1991) integraram o seu algoritmo a outros modelos que não utilizam padrões de co-ocorrência, como o de Hobbs (1978). Desta união, estatísticas foram coletadas de três corpora distintos: um corpus de artigos do jornal '*The Washington Post*' contendo cerca de 40 milhões de palavras, outro corpus composto por artigos do jornal *Associated Press* (24 milhões de palavras) e o próprio corpus *Hansard*. Foi utilizada a gramática ESG – *English Slot Grammar* (McCord, 1990) para o processamento das sentenças e depois foi aplicado o algoritmo de Hobbs. A taxa de sucesso desse experimento ficou em 64% para a implementação pura do algoritmo de Hobbs, e 86%, para a abordagem de Dagan & Itai agregada à de Hobbs.

Um problema identificado para essa abordagem é a escassez de dados necessários para a construção da base de estatísticas, o que a torna não aplicável para muitos casos de anáfora. Como alternativa a essa escassez de padrões de co-ocorrência, Mitkov (2002) propõe utilizar os padrões de colocação como um fator preferencial complementar no processo de resolução anafórica, e não como única medida de resolução.

### *2.4.3 - O algoritmo de Lappin & Leass*

O RAP (*Resolution of Anaphora Procedure*), como é conhecido o algoritmo proposto por Lappin & Leass (1994), resolve anáforas intra-sentenciais e inter-sentenciais pronominais de terceira pessoa da língua inglesa. O algoritmo é baseado em um sistema de pesos, atribuídos de acordo com a estrutura sintática da sentença gerada pela gramática ESG, também utilizada por Dagan & Itai (1991). Como descrito em Lappin & Leass (1994), o RAP contém os seguintes módulos principais:

a) Um procedimento para identificar pronomes, incluindo a identificação de pronomes semanticamente vazios (caso do *it*, no inglês).

b) Um algoritmo de ligação para identificar os possíveis antecedentes de um pronome reflexivo na mesma sentença.

c) Um filtro morfológico.

d) Um filtro intra-sentencial sintático, que descarta os SNs não co-referentes dos pronomes anafóricos, com base em informações sintáticas.

e) Uma função para atribuir os fatores de saliência adequados a cada sintagma nominal, como paralelismo sintático, preferência de sujeito, dentre outros.

f) Uma função de decisão para selecionar o antecedente.

Os componentes do RAP podem ser agrupados de acordo com as etapas de resolução anafórica descritas na Seção 2.3, desta maneira: a primeira etapa, na qual as anáforas são identificadas, pode ser representada pelo componente (a); a segunda etapa, na qual o conjunto de candidatos a antecedentes é determinado, estaria representada pelo componente (b), o algoritmo de ligação; e finalmente, a terceira etapa, onde o conjunto de possíveis antecedentes é identificado, seria representada pelos componentes (c) e (d), que servem para eliminar alguns dos possíveis candidatos identificados na etapa anterior, pelo componente (e), que promove os candidatos a antecedentes e pelo componente (f), que finalmente seleciona o antecedente da anáfora.

A avaliação do RAP se baseou em um corpus de cinco manuais técnicos de computadores, contendo um total aproximado de 80.000 palavras. Desse corpus, 560 ocorrências de pronomes de terceira pessoa (incluindo reflexivos e recíprocos) e seus antecedentes foram extraídos. Um teste aleatório foi executado com 360 pronomes escolhidos, randomicamente, do corpus acima, obtendo-se uma taxa de sucesso de 86%.

### *2.4.4 - Abordagem baseada em restrições e preferências*

Palomar et al. (2001) também apresentam um algoritmo para resolução de anáforas pronominais pessoais de terceira pessoa, incluindo, além desses, os pronomes demonstrativos e nulos, cujo antecedente seja um sintagma nominal. Diferentemente das abordagens anteriores, eles propõem um algoritmo para RA da língua espanhola. O algoritmo identifica tanto antecedentes intra-sentenciais quanto inter-sentenciais e é aplicado à estrutura sintática gerada pela gramática *Slot Unification Grammar* (Fernandes et al., 1997).

A idéia principal desse algoritmo, assim como em Hobbs (1978), Lappin & Leass (1994) e Mitkov (1998, 2002), é utilizar distintamente restrições e preferências, aplicadas nessa ordem, como estratégia de resolução. Os autores definiram uma lista de restrições e preferências para diferentes tipos de expressões pronominais considerando que as mesmas utilizam tipos de conhecimento diversos (lexical, morfológico, sintático e semântico).

O algoritmo proposto possui as seguintes funcionalidades principais:

a) Classificador pronominal: identifica qual é o tipo de pronome a ser processado.

b) Restrições:

   i) Concordância morfológica (pessoa, gênero e número).

ii) Restrições sintáticas: são baseadas nas restrições formuladas por Reinhart (1983)[11] e nas restrições de não co-referência propostas por Lappin & Leass (1994).

c) Preferências:

i) SNs não-abstratos (SNs inanimados são preteridos).

ii) SNs presentes na mesma sentença que a anáfora.

iii) SNs na mesma sentença que a anáfora e que também são soluções para outros pronomes nulos.

iv) SNs presentes na sentença anterior à da anáfora.

v) SNs não incluídos em um outro SN.

vi) SNs não preposicionados.

vii) SN que precede um verbo igual ao que acompanha a anáfora.

viii) SNs repetidos.

ix) SN que aparece mais de uma vez relacionado ao mesmo verbo da anáfora.

x) SN que aparece posposto a um verbo semelhante ao da anáfora;

Essas preferências foram agrupadas diferentemente para cada tipo de pronome a ser resolvido e aplicadas de forma ordenada. Por exemplo, para os pronomes nulos, as oito primeiras preferências são aplicadas na ordem em que foram listadas acima. Já, para a resolução de pronomes pessoais átonos, as seguintes preferências são aplicadas, nesta ordem: (i), (ii), (iv), (v), (vi) e (ix).

O processo de resolução anafórica pode ser sumarizado nas seguintes etapas:

1. Primeiramente, identifica-se o tipo de anáfora, isto é, se ela é um pronome pessoal, demonstrativo, reflexivo ou nulo.

2. Em seguida é gerada uma lista (L) dos possíveis SNs candidatos a antecedentes.[12]

3. Os candidatos da lista L são filtrados com o emprego de restrições morfológicas (concordância de gênero e número) e sintáticas. Ou seja, as restrições são aplicadas à lista L para se obter uma nova lista, L1.

[11] A utilização dessas restrições também foi feita por Coelho (2005).

[12] A lista criada depende do tipo de anáfora e do seu escopo de acessibilidade, obtido empiricamente através de um estudo aprofundado de corpus (Palomar et al., 2001). Para os pronomes reflexivos, somente são considerados os candidatos que ocorrem na mesma sentença que a anáfora. Para os demais, são considerados os SNs que aparecem em até quatro sentenças anteriores à sentença anafórica.

4. Caso reste apenas um candidato ( |L1| = 1 ), ele será escolhido como antecedente da anáfora. Caso contrário, o conjunto de preferências estabelecido é usado sobre os candidatos de L1 para se obter uma nova lista (L2) de candidatos a antecedentes.

5. Analogamente ao passo (4), se restar apenas um candidato ( |L2| = 1 ), ele será escolhido como antecedente da anáfora. Caso contrário, três regras de preferência, comuns a todos os tipos de pronomes, são aplicadas. São estas, em ordem de prioridade: SN que ocorre com maior freqüência no texto, SN que aparece mais vezes relacionado ao mesmo verbo da anáfora e o candidato mais próximo da anáfora. Finalmente, após o emprego dessas três últimas preferências, o antecedente é escolhido.

Esse algoritmo foi avaliado com manuais técnicos e textos literários retirados respectivamente da edição espanhola do corpus *Blue Book Corpus*[13] e do corpus *Lexesp*[14]. Os dois corpora totalizam 1677 pronomes dos quais 76,8% foram resolvidos corretamente pelo algoritmo de resolução proposto. Segundo os autores, essa taxa de sucesso não foi maior devido os seguintes fatores: a complexidade do corpus *Lexesp*, formado por textos narrativos, cujas sentenças possuem uma média de 16 palavras e cerca de 27 candidatos por anáfora; erros de etiquetagem (3%), erros de identificação de SNs complexos (7%), a ausência de informações semânticas (32%), erros de antecedentes divididos[15] (10%), ou seja, aqueles antecedentes formados por constituintes que não possuem nenhum tipo de ligação que permita agrupa-los em um aglomerado único, catáfora (2%), exófora (3%) e exceções na aplicação das preferências (43%).

Os autores implementaram também outras abordagens e as avaliaram com os mesmos corpora de teste: o algoritmo de Hobbs e o RAP, ambos já descritos anteriormente, com o intuito de compará-las com a sua proposta. Para essas abordagens eles obtiveram as seguintes taxas de sucesso: 62,7% e 67,4%, respectivamente, o que demonstra uma superioridade da sua abordagem.

Embora o algoritmo de Palomar et al. tenha sido construído para a língua espanhola, para este trabalho, algumas contribuições são consideradas relevantes: a aplicação de conhecimento lingüístico diverso, a distinção entre fatores restritivos e preferenciais, a

[13] CRATER (Projeto CRATER 1994-1995) *Corpus Resources and Terminology Extraction Project.* Projeto financiado pela Comunidade Européia.

[14] O corpus Lexesp pertence ao projeto de mesmo nome, construído pelo Departamento de Psicologia da Universidade de Oviedo.

[15] Uma descrição detalhada sobre antecedentes divididos (*split antecedents*) pode ser vista em Elbourne (2006) e no site < http://semanticsarchive.net/Archive/jU0ZTc2N/splitantecedents.pdf>.

ordem de aplicação desses fatores: restrições seguidas de preferências e a importância dada aos fatores restritivos no sucesso de RA, pois eles reduzem, consideravelmente, o número de possíveis candidatos a antecedentes.

### *2.4.5 - Outras abordagens de RA*

Os trabalhos apresentados nas seções anteriores concentram-se nas resoluções pronominais. Além das anáforas pronominais, trabalhos significativos para a resolução de outros tipos de anáfora envolvem o tratamento de descrições definidas (Poesio et al., 2005; Vieira & Poesio, 2000; Vieira, 1998; Munõz, 2001) e de anáforas associativas (Meyer & Dale, 2002a, 2002b; Haag & Othero, 2003).

As descrições definidas são expressões referenciais compostas por sintagmas nominais completos, isto é, artigo definido e substantivo (núcleo). Muitas vezes o núcleo vem acompanhado de elementos complementares (como adjetivos, etc.), por exemplo, a cantora, os biscoitos de chocolate. Ao serem introduzidas no texto, as DDs podem se referir a uma entidade que esteja aparecendo pela primeira vez no contexto discursivo, ou podem fazer referência a alguma entidade já mencionada no texto. Assim, elas podem ser novas no discurso ou podem apresentar uma relação de co-referência ou de referência com algum termo anterior (neste caso, descrições definidas co-referentes ou anafóricas).

Considere os exemplos seguintes (Haag & Othero, 2003: 1-2):

(2.14) Rodrigo estava exausto porque tinha acabado de voltar da escola.

(2.15) Há **um filme**$_i$ muito bom em cartaz. *O filme*$_i$ fala sobre a lenda do Rei Artur.

(2.16) Nós visitamos **um museu**$_i$ fantástico. *As esculturas*$_j$ eram belíssimas.

A descrição definida ‘a escola’, incluída no exemplo (2.14), aparece pela primeira vez no discurso, não apresentando nenhum referente ou co-referente textual. Ela pode ser classificada, portanto, como ‘nova no discurso’.

No exemplo (2.15), a relação de co-referência estabelecida entre as descrições definidas ‘o filme’ e ‘um filme’ é explicitada no texto, já que as mesmas referem-se a uma mesma entidade.

Já no exemplo (2.16), temos uma anáfora associativa. Esta ocorre quando a expressão referencial e seu antecedente não se referem à mesma entidade no discurso, mas a um mesmo contexto semântico. Nesse exemplo, a descrição definida 'as esculturas' faz referência a uma entidade diferente da DD 'um museu', mas ambas ainda mantêm uma relação entre si. A DD 'as esculturas' apresenta uma relação semântica com sua âncora textual, a expressão 'um museu', de forma que a expressão referencial definida 'as esculturas' remete ao antecedente 'um museu'. Dessa forma, as DDs não são exatamente novas no discurso, mas estão ancoradas num mesmo contexto semântico, que é dado pelo termo antecedente. Esse tipo de expressão referencial difere das anáforas pronominais porque nestas o termo anafórico referencia a mesma entidade, enquanto naquelas, a sua interpretação depende da busca por alguma relação semântico-discursiva que tome o antecedente como mera âncora textual.

O trabalho de Vieira & Poesio (2000) para a resolução de DDs, por exemplo, é baseado no processamento superficial das estruturas dos SNs que compõem as DDs. Os autores desenvolveram um sistema que não executa nenhum pré-processamento por si só, mas se beneficia de um subconjunto de dados provindos de um corpus anotado do *Penn Treebank I*[16]. Ele classifica cada DD como anáfora direta, anáfora indireta, anáfora associativa ou descrição nova no discurso.

A anáfora direta possui nome-núcleo (substantivo) igual ao do seu antecedente e refere-se à mesma entidade no discurso que ele; por exemplo: '**As listas**$_{\mathbf{i}}$ apontam quase todas as divisões e departamentos da Polícia Civil. Alguns delegados da Polícia Federal também são citados n*as listas*$_{\mathrm{i}}$'.

A anáfora indireta possui um antecedente que denota a mesma entidade do discurso que ela, porém está representada por nome-núcleo diferente e requer conhecimento extra-lingüístico para sua identificação. Assim, o núcleo da anáfora pode ser um sinônimo do antecedente, ou mesmo uma elipse; por exemplo: '**A Folha de São Paulo**$_{\mathrm{i}}$ apresentou as listas apreendidas na operação contra o crime organizado. O *jornal*$_i$ tentou ouvir o delegado encarregado'.

Já as anáforas associativas, como já visto, possuem um antecedente textual não co-referente que está relacionado a elas por uma relação semântico-discursiva.

Para avaliação desse sistema de classificação de DDs foram considerados 33 artigos do jornal *Wall Street*, escolhidos aleatoriamente, que estavam contidos no corpus *Penn*

[16] <http://www.cis.upenn.edu/~treebank/>

*Treebank I*. Essa parte do corpus contém 464 DDs e o modelo proposto conseguiu obter 62% de cobertura e 83% de precisão para resolução de anáforas diretas, enquanto a identificação de descrições novas no discurso atingiu 69% de cobertura e 72% de precisão. Nesse contexto, a cobertura representa o número de DDs classificadas corretamente frente ao total de DDs presentes no corpus, enquanto a precisão representa o número de DDs classificadas corretamente em relação ao total de DDs que se tentou classificar.

Para a resolução de anáforas associativas, por exemplo, Meyer & Dale (2002a, 2002b) utilizam a técnica de mineração de textos para realizar, de forma automática, a aquisição de relações semânticas e criar uma base de axiomas associativos que são utilizados no processo de resolução. Como recurso extra, utilizam a *WordNet* como fonte de desambiguação na construção de axiomas gerais que representem a estrutura sintática das anáforas.

Essa abordagem emprega no pré-processamento o analisador sintático *Conexor's FDG Parser* (Tapanainen & Järvinen, 1997) para identificar as construções associativas. A avaliação dos axiomas associativos utilizados nessa abordagem serviu para que os autores os considerassem um recurso a mais no processo de RA a ser usado como filtro para os possíveis candidatos a antecedentes.

### *2.4.6 - Considerações sobre as abordagens de RA*

Neste capítulo foi apresentada uma visão geral do fenômeno lingüístico de referenciação, o conceito de anáfora e suas variações, os problemas encontrados no processo de resolução anafórica, as etapas de processamento automático das anáforas, os fatores de resolução mais comumente empregados na identificação do antecedente anafórico, além de destacar-se a natureza heterogênea do conhecimento envolvido no processo de RA.

Verificou-se que um dos maiores problemas no processamento automático completo de anáforas é o pré-processamento necessário para entrada do algoritmo de resolução, que envolve diversas questões como análise morfológica, reconhecimento de nomes próprios, identificação de pronomes pleonásticos (não aplicável para o português), etc. Cada uma dessas tarefas introduz erros e pode contribuir para o insucesso do sistema de resolução como um todo.

De um modo geral, as abordagens de RA, descritas anteriormente, apresentam, com ênfase variada, as seguintes preocupações básicas: a definição do tipo de conhecimento

(fatores de resolução) a considerar (Tabela 1), a integração desses fatores (cálculo de saliência[17] dos candidatos) e uma estratégia de resolução baseada na saliência dos candidatos.

Tabela 1: Síntese dos fatores de resolução utilizados pelas abordagens de RA

| **Proposta** | **Fatores restritivos** | **Fatores preferenciais** |
|---|---|---|
| Hobbs (1978) | ✓ Concordância morfológica | – |
| Lappin & Leass (1994) | ✓ Concordância morfológica<br>✓ Restrições sintáticas | ✓ Paralelismo sintático<br>✓ Preferência de sujeito, etc. |
| Dagan & Itai (1991) | – | ✓ Padrões de co-ocorrência |
| Palomar et al. (2001) | ✓ Concordância morfológica<br>✓ Restrições sintáticas | ✓ Preferência por SNs presentes na mesma sentença que a anáfora e na sentença anterior<br>✓ Preferência por SNs repetidos<br>✓ Preferência por SNs não preposicionados, etc. |

Optou-se por apresentar uma síntese das avaliações de tais abordagens somente para ilustrar como os sistemas de RA são geralmente avaliados. Ressalta-se que todos esses trabalhos e o nosso não podem ser comparados, pois além de alguns deles serem construídos para línguas distintas, tais abordagens foram avaliadas com corpora distintos.

Os trabalhos apresentados também podem ser agrupados de acordo com o conhecimento utilizado no processo de resolução (Tabela 2): o conhecimento sintático é utilizado pelo algoritmo de Hobbs (1978), que obteve uma taxa de sucesso de 88,3%; o algoritmo de Palomar et al. (2001) se baseia em restrições sintáticas e obteve uma taxa de sucesso de 76,8% na sua avaliação; Lappin & Leass (1994) utilizaram um sistema de pesos atribuídos de acordo com a estrutura sintática da sentença e obtiveram uma taxa de sucesso de 86%. O algoritmo de Dagan & Itai (1991) difere dos anteriores por utilizar padrões de co-ocorrência extraídos automaticamente de um corpus que são usados para filtrar candidatos improváveis de serem o antecedente da anáfora, portanto, não utiliza conhecimento lingüístico complexo no seu processo de resolução. Sua taxa de sucesso na RA foi de 87%, lembrando que esse algoritmo é dependente de domínio, pois os padrões de co-ocorrência são extraídos de um domínio específico.

[17] Probabilidade de co-ocorrência.

Tabela 2: Síntese da avaliação das abordagens de RA

| Proposta | Descrição do Corpus | Tam. do Corpus (palavras) | # PRONs | Tipo de PRON[18] | Tipo de Anáfora | TS (%) |
|---|---|---|---|---|---|---|
| Puramente Sintática Hobbs (1978) | Textos de literatura, arqueologia e jornalismo | - | 300 | PP | Inter/Intra sentencial | 88,3 |
| Padrões de colocação Dagan & Itai (1991) | Pareceres do parlamento canadense | 28 milhões | 38 | P | Inter/Intra sentencial | 87 |
| Sistema de pesos de acordo com estrutura sintática Lappin & Leass (1994) | Manuais técnicos de computadores | 80 mil aprox. | 360 | PP, N, Rf e Rc | Inter/Intra sentencial | 86 |
| Baseada em restrições e preferências Palomar et al. (2001) | Manuais técnicos e textos literários | 5 milhões | 1677 | PP, N, D, Rf e Rc. | Inter/Intra sentencial | 76,8 |

Dentre essas abordagens, a proposta por Hobbs obteve o melhor desempenho, representado pela taxa de sucesso. Porém, devemos considerar que esse resultado pode ter sido influenciado pela metodologia de avaliação proposta pelo autor, que consistiu de uma simulação manual do algoritmo e utilizou uma estrutura sintática perfeita das sentenças analisadas. O processamento automático de sentenças insere erros como estrutura sintática e etiquetagem incorretas. Já a proposta com maior taxa de sucesso, utilizando um pré-processamento automático gerado por um *parser*, foi a apresentada por Dagan & Itai (1991).

A abordagem puramente sintática proposta por Hobbs continua sendo um dos trabalhos mais influentes em resolução anafórica e é freqüentemente utilizada como *benchmark* para avaliação de novas propostas. Diversos autores utilizaram-na ou implementaram-na para comparação de desempenho com suas propostas, como Lappin & Leass (1994), Dagan & Itai (1991), Palomar et al. (2001) e Mitkov (2002).

O próximo capítulo apresenta as propostas de resolução anafórica para a língua portuguesa. Algumas delas se baseiam nos trabalhos descritos neste capítulo, em especial a resolução de descrições definidas (Vieira, 1998) e de pronomes, particularmente, o algoritmo de Lappin & Leass (1994).

[18] Pessoais de 3ª Pessoa (PP), Nulos (N), Reflexivos (Rf), Recíprocos (Rc) e Demonstrativos (D).

# Capítulo 3 - A resolução anafórica na língua portuguesa

Neste capítulo são apresentados alguns trabalhos de RA desenvolvidos no Brasil. Dentre eles, detalhamos as propostas de classificação de descrições definidas[19] e enfatizamos as abordagens que lidam com o processamento de anáforas pronominais (Coelho, 2005; Coelho & Carvalho, 2005; Paraboni,1997). Coelho (2005) resolve pronomes pessoais de terceira pessoa e reflexivos/recíprocos enquanto Paraboni (1997), pronomes possessivos.

## 3.1 - Processamento de descrições definidas

Trabalhos anteriores (Vieira, 1998; Vieira & Poesio, 2000) apresentam um estudo detalhado sobre o uso de descrições definidas na língua inglesa e propõem um sistema baseado em corpus para o processamento destas expressões. Esses trabalhos têm servido de base para estudos sobre a resolução de DDs na língua portuguesa (Vieira et al., 2000; Vieira, 2001; Rossi et al., 2001; Collovini et al., 2005 e Coelho et al., 2005, 2006).

Rossi et al. (2001) desenvolveram um sistema de resolução de co-referência para o português baseado em (Vieira, 1998), com o propósito de classificar as descrições definidas como novas no discurso ou anafóricas O sistema funciona da seguinte maneira:

1. Efetua-se a leitura de um arquivo contendo uma lista de SNs[20].
2. Atribui-se um índice para cada SN extraído. Os SNs considerados antecedentes potenciais são armazenados em uma base de dados.
3. Classifica-se o SN: caso este seja uma DD, os seguintes procedimentos são executados para sua classificação: primeiramente, busca-se encontrar o núcleo desta DD, que então é comparado com os núcleos dos SNs armazenados na lista de antecedentes potenciais. Havendo um antecedente, a DD é classificada como anáfora direta. Caso contrário, investiga-se indícios (como existência de pós-modificação com preposição, núcleo formado por nome próprio, presença

[19] Desenvolvidas junto à UNISINOS-RS e coordenadas pela professora Renata Vieira.
[20] Os SNs foram extraídos de um corpus composto de 15 textos/artigos do Jornal Correio do Povo, de Porto Alegre, editado em 1999. Os textos foram processados para a extração dos sintagmas nominais e corrigidos manualmente para eliminar possíveis erros. A construção desse corpus foi realizada por Vieira et al. (2000).

de letra maiúscula e construção de aposto) de que a DD possa ser uma descrição nova no discurso; se algum desses indícios for encontrado, a DD é classificada como nova no discurso. Caso contrário, a DD é considerada como não classificada.

Para avaliar essa proposta, os autores compararam a solução gerada automaticamente pelo classificador com a classificação manual do sintagma correspondente, contabilizando assim, quantas DDs possuem a mesma classificação e quantas possuem classificação diversa. Além disso, determinaram o total de cada classe e o total das DDs classificadas, exibindo uma saída semelhante à ilustrada na Tabela 3.

Tabela 3: Resultados da classificação manual e automática de descrições definidas

| **: : : : Comparação da análise manual e automática** | | |
|---|---|---|
| **Classificação:** | **Manual** | **Automática** |
| ❖ Nº. de DDs classificadas manualmente ..... | 69 | 52 |
| ❖ Nº. de DDs classificadas como novas no discurso ..... | 33 | 31 |
| ❖ Nº. de DDs classificadas como anáforas diretas ..... | 15 | 21 |
| ❖ Nº. de DDs classificadas como anáforas indiretas ..... | 07 | 00 |
| ❖ Nº. de DDs classificadas como associativas ..... | 00 | 00 |
| ❖ Nº. de DDs não-classificadas ..... | 14 | 17 |
| Total de DDs com igual classificação ..... 30 | | |

Rossi et al. (2001) verificaram que os valores apresentados na Tabela 3 são similares aos reportados por Vieira (1998) para a língua inglesa, comprovando a portabilidade dessa metodologia para o português.

Assim como os autores acima, Collovini et al. (2005) propuseram um sistema de classificação automática de DDs baseado na classificação de Vieira (1998). A construção desse sistema fundamentou-se em um estudo de corpus sobre DDs do português desenvolvido por Coelho et al. (2005). O diferencial dessa proposta em relação à anteriormente descrita é que nessa cria-se uma base de dados para a classificação automática das descrições definidas com árvores de decisão: para a classificação das DDs, os autores adotam uma metodologia similar às etapas tradicionais de sistemas de categorização de textos, que pode ser resumida a seguir.

A etapa de coleta da base de dados consistiu na obtenção dos exemplos a serem utilizados para o treinamento do classificador. A base de dados foi constituída de um extrato

do corpus NILC[21], formado por 24 textos jornalísticos da Folha de São Paulo. Os exemplos usados pelo classificador foram as DDs presentes nesses textos, que já tinham sido anotadas manualmente com informações de co-referência. Após a obtenção desses exemplos, foi criada uma representação conceitual dessa base de dados e, posteriormente, realizada a classificação das DDs.

Na classificação das descrições definidas como novas no discurso, foi obtida uma taxa de sucesso de 70,4%, e para a classe não co-referente, a taxa de sucesso foi de 77,6%.

Os classificadores de DDs apresentados podem ser muito úteis na primeira etapa de resolução anafórica que consiste na identificação das anáforas, isto é, na classificação de um termo em anafórico ou não. Contudo, esses classificadores não são úteis para esta proposta de trabalho já que a mesma se resume à resolução de pronomes. A próxima seção apresenta algumas abordagens importantes desenvolvidas no Brasil para esse tipo específico de RA.

## 3.2 - Processamento de anáforas pronominais

Apresentamos a seguir duas abordagens para resolução de anáforas pronominais, a primeira proposta por Coelho (2005) para resolução de pronomes pessoais de terceira pessoa, reflexivos e recíprocos, e a segunda desenvolvida por Paraboni (1997) para a resolução de pronomes possessivos.

Lembramos que, na língua portuguesa, os pronomes substituem (caso dos pronomes pessoais) ou acompanham um substantivo (caso dos pronomes possessivos), indicando as pessoas do discurso: a pessoa que fala (1ª pessoa), a pessoa com quem se fala (2ª pessoa) e a pessoa de quem se fala (3ª pessoa).

Os pronomes pessoais, foco desse trabalho, subdividem-se em dois casos: retos e oblíquos. Os pronomes pessoais do caso reto são os que desempenham a função sintática de sujeito da oração. São estes: eu, tu ele/ela, nós, vós, eles/elas. Já os oblíquos desempenham a função sintática de complemento verbal (objeto direto ou indireto), complemento nominal, agente da passiva, adjunto adverbial, adjunto adnominal ou sujeito acusativo (sujeito de oração reduzida). São estes: me, mim, comigo, te, ti, contigo, o, a, lhe, se, si, consigo, nos, conosco, vos, convosco, os, as, lhes, se, si, consigo.

[21] <http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm>.

Os pronomes reflexivos são pronomes pessoais oblíquos que, embora funcionem como objeto direto ou indireto, referem-se ao sujeito da oração, por exemplo: '**Ana**$_{(sujeito)}$ desceu a escada e *se*$_{(objeto\ PRON\_Refl.)}$ machucou'. Quanto aos pronomes recíprocos, todos eles são reflexivos. Eles indicam a reciprocidade (troca de ação) entre sujeito e objeto, por exemplo: '**Os namorados**$_{(sujeito)}$ *se*$_{(objeto\ PRON\_Rec.)}$ beijaram'.

Como visto na Seção 2.2, os pronomes em primeira e segunda pessoa ocorrem, na maioria das vezes, de maneira dêitica, o que os torna não-anafóricos. Por isso, propomos neste trabalho a resolução apenas de pronomes pessoais de terceira pessoa, o que necessariamente inclui os reflexivos e recíprocos. Como veremos a seguir, Coelho (2005) distingue, no seu processo de RA, os pronomes pessoais em: pronomes de terceira pessoa, reflexivos e recíprocos a fim de resolver alguns pronomes de primeira pessoa, como é o caso do pronome reflexivo/recíproco 'me'.

### *3.2.1 - O uso do algoritmo RAP para a RA da língua portuguesa*

O algoritmo desenvolvido por Coelho (2005) em sua dissertação de mestrado é uma adaptação do algoritmo RAP (Lappin & Leass, 1994) visto na Seção 2.4.3, para resolver anáforas pronominais inter e intra-sentenciais, com foco nos pronomes reflexivos/recíprocos, utilizando uma janela[22] de quatro sentenças para a procura do antecedente no texto. Coelho implementou os principais módulos do algoritmo original, com as seguintes diferenças, algumas específicas para o processamento do português:

- O filtro sintático e o algoritmo de ligação foram substituídos pelas restrições de co-referência propostas por Reinhart (1983). O autor justifica essa substituição baseando-se na análise de exemplos encontrados em Lappin & Leass (1994) e Lappin & McCord (1999a, 1999b). Ele verifica que essas restrições são suficientes para resolver os casos anafóricos apresentados pelos autores do RAP.
- O analisador sintático utilizado foi o PALAVRAS (Bick, 2000)[23]. Ele apóia-se num léxico de 50.000 palavras e milhares de regras gramaticais para fornecer uma análise completa, tanto morfológica como sintática, de qualquer texto. Utilizando um conjunto de etiquetas gramaticais bastante diversificado, o

[22] As quatro sentenças consideradas compreendem a sentença em que ocorre a anáfora e três sentenças precedentes.
[23] O PALAVRAS pode ser utilizado via web através do site < http://visl.sdu.dk/visl/pt/ >

*parser* alcança um nível de precisão de 99% em termos de morfologia (classe de palavras e flexão), e 97-98% em termos de sintaxe.

- A ferramenta Xtractor (Gasperin et al., 2003) foi empregada para converter a saída do *parser* PALAVRAS em XML[24] (***eXtensible Markup Language***). Essa ferramenta foi utilizada para facilitar a extração da informação disponibilizada pelo PALAVRAS. Ela converte a saída do PALAVRAS em três arquivos XML: o primeiro arquivo possui extensão '*.words*'. Ele contém uma lista das palavras do texto e seus respectivos identificadores (que são *tokens* de representação interna ao sistema dos componentes textuais). O segundo arquivo, cuja extensão é '*.pos*'[25], contém informações morfológicas (por exemplo, gênero e número) sobre as palavras do texto. O terceiro arquivo, cuja extensão é '*.chunks*', contém a estrutura sintática das sentenças. Um arquivo '*chunk*' pode possuir sub-elementos '*chunks*' com informações das sub-estruturas da sentença. Para exemplificar, vejamos a descrição definida 'O presidente nacional' e seus respectivos arquivos *words* (Figura 3), *pos* (Figura 4) e *chunks* (Figura 5).

```
...
  <word id="word_1">O</word>
  <word id="word_2">presidente</word>
  <word id="word_3">nacional</word>
...
```

Figura 3: Arquivo *words*

```
...
<words>
- <word id="word_1">
- <art canon="o" gender="M" number="S">
  <secondary_art tag="artd" />
  </art>
  </word>
- <word id="word_2">
- <n canon="presidente" gender="M" number="S">
  <secondary_n tag="Hprof" />
  </n>
  </word>
- <word id="word_3">
  <adj canon="nacional" gender="M" number="S" />
  </word>
...
```

Figura 4: Arquivo *pos*

```
...
- <sentence id="sentence_1" span="word_1..word_35">
- <chunk ext="sta" form="fcl" id="chunk_1" span="word_1..word_34">
- <chunk ext="subj" form="np" id="chunk_2" span="word_1..word_8">
  <chunk ext="n" form="art" id="chunk_3" span="word_1" />
...
```

Figura 5: Arquivo *chunks*

[24] Um descrição detalhada da linguagem de marcação XML pode ser vista em <http://www.w3.org/xml>
[25] POS = *part-of-speech.*

- O módulo de identificação do uso pleonástico do pronome *it* não foi implementado, já que esse fenômeno não ocorre no português.
- Um módulo de tratamento de catáforas também não foi implementado, pois foge ao escopo proposto por Coelho (2005), que é resolver anáforas.

O sistema desenvolvido baseado nesse algoritmo foi implementado em java, sendo esta linguagem escolhida devido à sua API (*Application Program Interface*) e a seu suporte ao processamento e manipulação de documentos XML. A Figura 6 ilustra a arquitetura desse sistema.

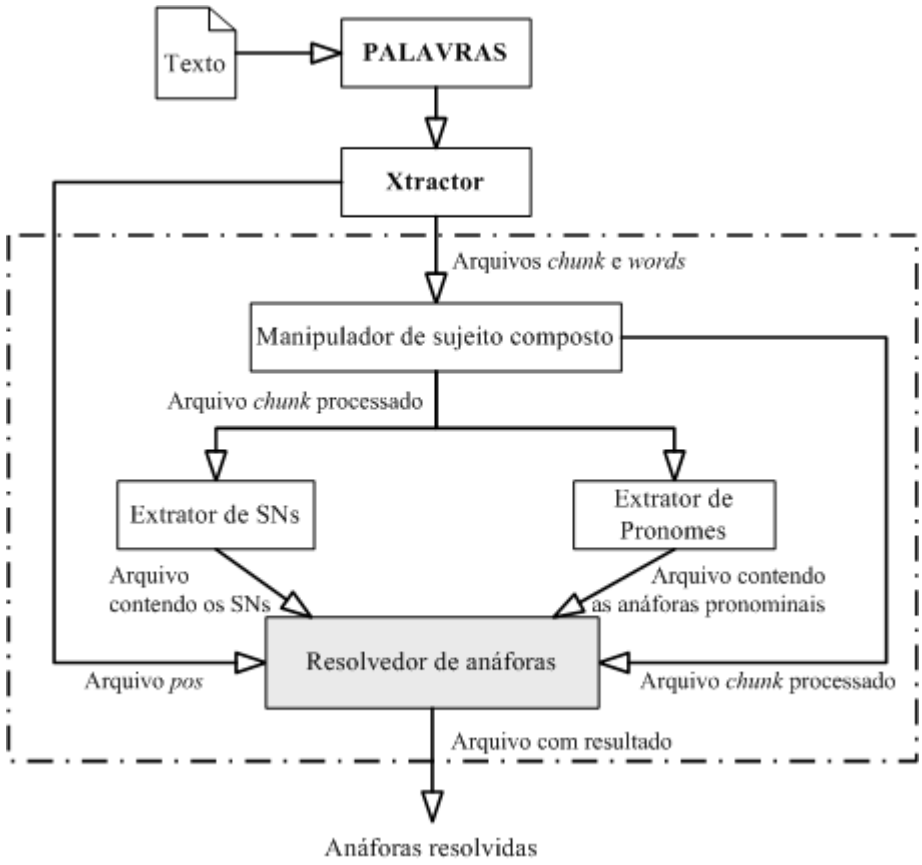



Figura 6: Arquitetura do sistema de Coelho

Nessa arquitetura, todos os arquivos do Xtractor (extensões *.words*, *.pos* e *.chunks*) são utilizados como entrada do sistema. A partir deles, ocorre a identificação e agrupamento de sujeitos compostos pelo manipulador de sujeitos compostos. Em seguida, são extraídos os possíveis candidatos à co-referência, pelo 'Extrator de sintagmas nominais'. Esse módulo gera um arquivo XML, conforme ilustrado na Figura 7, que contém todos os SNs extraídos do textos. As anáforas pronominais são extraídas pelo 'Extrator de pronomes', que gera como resultado um arquivo XML, como o da Figura 8, contendo todos os pronomes que o algoritmo tentará resolver. Para finalizar, o módulo 'Resolvedor de anáforas' resolve as

anáforas identificadas baseando-se no algoritmo de Lappin & Leass adaptado para o português, e gera um arquivo XML com as anáforas e seus respectivos antecedentes.

```
...
<np-set>
- <sentence id="sentence_1">
  <np chunk_id="chunk_2" gender="M" head_span="word_2" number="S" span="word_1..word_8" />
  <np chunk_id="chunk_8" gender="M" head_span="word_6" number="S" span="word_5..word_8" />
  <np chunk_id="chunk_15" gender="F" head_span="word_13" number="S" span="word_12..word_13" />
  <np chunk_id="chunk_20" gender="M" head_span="word_16" number="S" span="word_15..word_16" />
  <np chunk_id="chunk_29" gender="F" head_span="word_21" number="P" span="word_20..word_21" />
  <np chunk_id="chunk_37" gender="M" head_span="word_26" number="S" span="word_25..word_28" />
  <np chunk_id="chunk_42" gender="M" head_span="word_28" number="S" span="word_28" />
  <np chunk_id="chunk_45" gender="M" head_span="word_31" number="S" span="word_30..word_34" />
  <np chunk_id="chunk_50" gender="F" head_span="word_34" number="S" span="word_33..word_34" />
  </sentence>
- <sentence id="sentence_2">
...
```

Figura 7: Arquivo de sintagmas

```
...
- <pronoun-set>
  <sentence id="sentence_1" />
  ...
- <sentence id="sentence_5">
  <pron chunk_id="chunk_171" gender="M" number="3S" reci="no" refl="no" span="word_118" />
  </sentence>
  ...
- <sentence id="sentence_8">
  <pron chunk_id="chunk_229" gender="M" number="3S" reci="no" refl="no" span="word_160" />
  </sentence>
...
```

Figura 8: Arquivo de pronomes

O sistema proposto foi avaliado utilizando-se três corpora de gêneros distintos: jurídico, literário e jornalístico. Estes corpora foram anotados automaticamente pelo PALAVRAS com informações morfossintáticas, e manualmente, com o auxílio da ferramenta MMAX – *Multi-Modal Annotation in XML* (Müller & Strube, 2001) com informações de co-referência anafórica.

O corpus jurídico é composto por pareceres da Procuradoria Geral da República de Portugal[26], constituído de sentenças longas e complexas. O corpus literário, também de natureza complexa, consiste do livro 'O alienista' de Machado de Assis. Já o corpus jornalístico é constituído de 14 textos da revista Veja, cujas sentenças são mais simples que as dos dois corpora anteriores.

A anotação do corpus jurídico não englobava todos os pronomes de terceira pessoa e não incluía os pronomes reflexivos e recíprocos. Já os demais corpora foram anotados por Coelho, abrangendo todas as anáforas tratadas pelo sistema e reconhecidas pelo

[26] Esse corpus foi fornecido pela professora Renata Vieira da UNISINOS-RS e já continha anotações manuais sobre as anáforas pronominais.

PALAVRAS, além de classificar todas as expressões referenciais encontradas como inter-sentenciais, intra-sentenciais ou não-anafóricas.

A Tabela 4 exibe a distribuição das anáforas nos corpora jurídico, literário e jornalístico de acordo com o tipo de pronome e tipo de expressão anafórica (apenas para os dois últimos corpora).

Tabela 4: Distribuição e classificação das anáforas nos corpora

| | **Corpus jurídico** | **Corpus literário** | **Corpus jornalístico** |
|---|---|---|---|
| **Tipo de Pronome** | | | |
| Pronomes de terceira pessoa | 297 (100%) | 595 (85,49%) | 162 (72%) |
| Pronomes reflexivos/recíprocos | Não anotados | 101 (14,51%) | 63 (28%) |
| **Total de Pronomes anotados** | **297** | **696** | **225** |
| **Tipo de anáfora** | | | |
| Inter-sentencial | - | 219 (31,46%) | 113 (50,22%) |
| Intra-sentencial | - | 372 (53,45%) | 70 (31,11%) |
| Pronomes não-anafóricos | - | 105 (15,09%) | 42 (18,67%) |
| **Total de anáforas anotadas** | **297** | **696** | **225** |

A avaliação do sistema de Coelho consistiu de uma comparação automática entre as soluções geradas automaticamente e as soluções manuais. Considerou-se que o resultado gerado pelo sistema estaria correto caso fosse idêntico ao anotado manualmente ou se o mesmo fosse um sintagma nominal contido no SN dado pelos anotadores. Para tal avaliação foram obtidas como resultado global para cada corpus, as seguintes taxas de sucesso: jurídico (35%), literário (32,61%) e jornalístico (43,56%). Essa avaliação também foi feita para cada tipo de pronome individualmente, constatando-se uma melhor identificação dos antecedentes de anáforas pronominais reflexivas/recíprocas, isto é, anáforas determinadas por pronomes reflexivos ou recíprocos. Os autores justificam esse melhor desempenho baseando-se no fato de que o processo de resolução proposto coleta apenas candidatos intra-sentenciais, o que reduz consideravelmente o número de candidatos a serem analisados.

Constatou-se que o desempenho obtido nesses experimentos foi inferior à proposta original para o inglês, que obteve uma taxa de sucesso de 86% (Tabela 2). Essa diferença pode ser justificada pelo fato da abordagem original, que resolve anáforas da língua inglesa, ter utilizado para avaliação, um corpus mais simples (manuais de computadores). Além disso, ao ser adaptada para o português, alguns erros de pré-processamento foram

inseridos pelas ferramentas PALAVRAS e Xtractor (informações morfossintáticas incorretas, identificação incorreta de pronomes, dentre outros) e tais erros não foram contabilizados.

### *3.2.2 - A resolução de pronomes possessivos*

Outro importante trabalho em resolução pronominal foi desenvolvido por Paraboni (1997). Ele propôs uma arquitetura para resolução de pronomes possessivos em textos escritos em língua portuguesa considerando um corpus no domínio da legislação ambiental.

Em sua análise de corpus, o autor constatou que a referência pronominal possessiva apresenta algumas dificuldades de interpretação não presentes em outros tipos de anáfora da língua portuguesa, ou em equivalentes da língua inglesa. Como exemplos dessas dificuldades temos: a ausência de concordância de gênero e número entre a anáfora e o antecedente[27], a variedade de funções sintáticas exercidas, a natureza ambígua de alguns pronomes de terceira pessoa como o pronome 'sua' na sentença (3.1) (a casa pertence a quem? Maria? Ou ao pai de Maria?), e muitas vezes, o caráter abstrato da relação anafórica estabelecida.

(3.1) Vi **Maria** com **seu pai**, à porta de *sua casa*.

Devido à natureza complexa desse tipo de anáfora, Paraboni propôs uma arquitetura multi-agentes que também faz uso de fatores de resolução restritivos e preferenciais, os quais utilizam conhecimento lingüístico heterogêneo: sintático (padrões de superfície), semântico (relações de posse) e pragmático (centro da sentença)[28]. Ele também considerou a ordem de aplicação desses fatores e o peso relativo de cada um deles na determinação da solução global. A Tabela 5 apresenta os fatores de resolução utilizados, diferenciando a sua natureza (restrição ou preferência) e o tipo de conhecimento utilizado.

[27] Os pronomes possessivos são palavras que fazem referência às pessoas do discurso, apresentando-as como possuidoras de alguma coisa (Rocha Lima, 1978). Eles fazem parte do sintagma nominal da coisa possuída. Portanto, pronomes possessivos da língua portuguesa concordam em gênero e número com a coisa possuída e não com o termo a que se referem.

[28] A noção pragmática de centro da sentença é abordada por Brennan et al (1987), Brennan (1995), Sidner (1983) e Allen (1995).

Tabela 5: Fatores considerados na resolução de referências pronominais possessivas

| Fator | Natureza | Conhecimento | Enunciado |
|---|---|---|---|
| F1 | restritiva | Sintático | Um termo candidato ligado à RPP (referência pronominal possessiva) por meio de conjunção constitui o próprio termo antecedente da RPP. |
| F2 | restritiva | Sintático | Nas RPPs regidas por preposição, o termo candidato também regido por preposição é o termo antecedente da RPP. |
| F3 | restritiva | Sintático | Termos candidatos diretamente ligados à RPP por meio de preposição não são válidos para co-referência. |
| F4 | restritiva | Sintático | Somente as extremidades de cadeias de SNs ligados por preposição constituem candidatos válidos a termo antecedente. |
| F5 | restritiva | Semântico | Antecedente e RPP devem estabelecer uma relação de posse semanticamente aceitável. |
| F6 | preferencial | Pragmático | O centro da sentença é o candidato preferencial à co-referência. |

A arquitetura proposta por Paraboni (Figura 9) promove a distribuição do conhecimento e dos fatores de resolução em entidades autônomas (agentes reativos) especializadas em aspectos distintos do problema de RA.

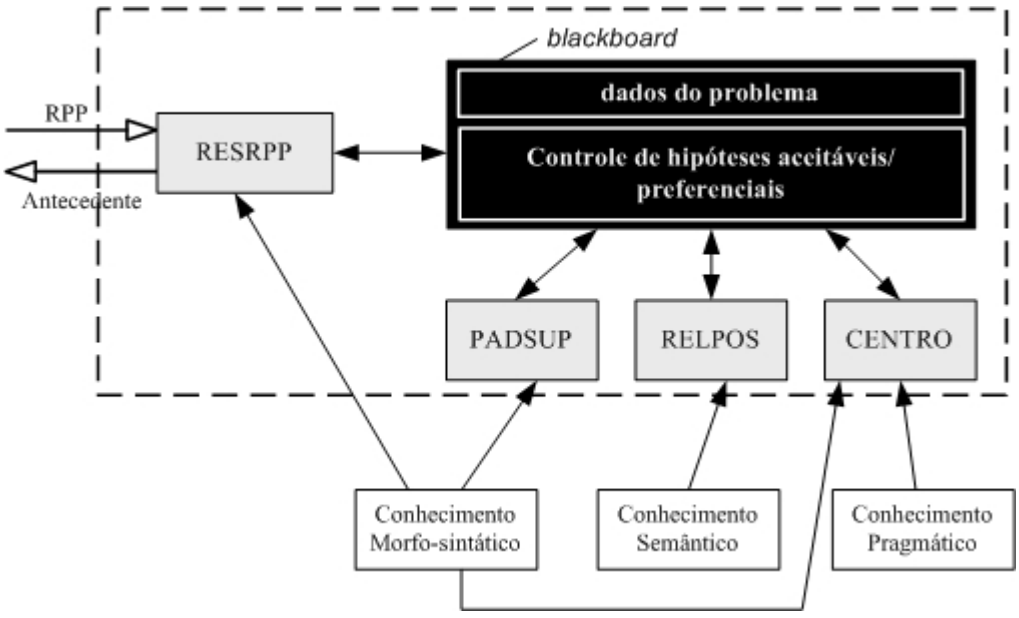



Figura 9: Arquitetura multi-agentes para resolução de RPP

Nessa arquitetura, os agentes reativos são agregados a uma estrutura global em arquitetura *blackboard* para a resolução das referências pronominais possessivas (RPPs). Essa arquitetura "é indicada para sistemas em que coexistem diversas bases de conhecimento independentes, com necessidades de integração de dados heterogêneos" (Paraboni, 1997: 57).

O processo de resolução é iniciado quando um coordenador – o agente RESRPP (Agente de '**Res**olução de **R**eferências **P**ronominais **P**ossessivas'), responsável por responder a uma requisição de resolução de RPP informando o candidato mais provável –

disponibiliza no *blackboard* os dados do problema, que incluem a própria estrutura simplificada da sentença (são identificadas as classes gramaticais das palavras e certas relações de dependências básicas entre os SNs) e os pronomes anafóricos; e as hipóteses de solução, isto é, os candidatos a antecedentes. Todos os substantivos que antecedem a anáfora e que estejam presentes na mesma sentença desta são considerados candidatos a antecedente.

Cada agente examina o *blackboard*, procurando uma oportunidade de aplicar seus conhecimentos aos dados fornecidos. Conforme mostra a Figura 9, são quatro os agentes especialistas: o agente de 'Padrões de superfície' (PADSUP), responsável por aplicar as restrições sintáticas (fatores F1 a F4), o agente de 'Relações de Posse' (RELPOS), responsável por aplicar as restrições semânticas (fator F5) e o agente 'Centro de Sentença' (CENTRO), que aplica a preferência pragmática (fator F6). Cada agente adiciona suas contribuições até que o agente coordenador (RESRPP) constate que o problema foi resolvido e termine a execução do sistema. A determinação do antecedente pode ocorrer mesmo antes de se esgotarem as possibilidades de contribuição de todos os especialistas.

Para avaliação da abordagem proposta, Paraboni utilizou como entrada para o sistema informações sobre a superfície da sentença (como classificação das palavras e algumas dependências básicas entre os SNs), obtidas diretamente do corpus e anotadas manualmente. Esse procedimento evitou a necessidade de criação de ferramentas de pré-processamento para extração de conhecimento léxico-morfológico e sintático gerados por analisadores sintáticos automáticos, porém demandou grande esforço humano para anotação. A inserção de conhecimento semântico se resumiu à inclusão de um dicionário de sinônimos como componente do agente RELPOS.

Como resultado inicial da implementação dessa proposta, obteve-se um índice de acerto de 86,87% na resolução de RPPs do corpus. Este índice é compatível com a resolução de outros tipos de anáforas, como já descrito nas seções anteriores, desconsiderando aqui, o esforço humano empregado no processo de anotação manual do corpus. Paraboni ainda realizou outro teste para determinar o grau de ambigüidade de certas RPPs, de modo a eliminar os casos de ambigüidade excessiva e melhorar os resultados obtidos no primeiro experimento. Para isso solicitou que juízes humanos marcassem as sentenças ambíguas no corpus. Como resultado desse teste, obteve um índice ainda melhor para o seu sistema de resolução, 92,97% de RPPs resolvidas, agora desconsiderando as sentenças ambíguas, que sequer são interpretadas por humanos.

## 3.3 - Considerações sobre as abordagens de RA para o português

Como mostra este capítulo, a maioria dos trabalhos sobre RA do português é de autoria do grupo de pesquisa da Prof.ª Renata Vieira (UNISINOS-RS), que desenvolveu tanto estudos de corpus quanto sistemas para classificação das descrições definidas baseando-se em estudos sobre o processamento das DDs na língua inglesa (Vieira, 1998). Somente Coelho (2005) e Paraboni (1997) propuseram sistemas para a resolução de pronomes. Coelho adaptou o RAP (Lappin & Leass, 1994) para a resolução de pronomes pessoais de terceira pessoa, enquanto Paraboni desenvolveu uma arquitetura multi-agentes para resolução de anáforas pronominais possessivas, baseada em conhecimento heterogêneo embutido em fatores de resolução restritivos e preferenciais. A maioria dessas propostas, excetuando-se a de Paraboni, de alguma maneira utilizou as mesmas ferramentas de pré-processamento (*parser* PALAVRAS, Xtractor e MMAX), seja para criação de um corpus de teste, anotado com informações morfossintáticas e referenciais, seja para servir de entrada aos sistemas de resolução, como fez Coelho (2005).

Os níveis de conhecimento empregados em cada abordagem foram: informações morfológicas e sintáticas (utilizadas por todos os trabalhos apresentados), anotações de co-referência foram úteis para as abordagens que processaram DDs e para o trabalho de Coelho, enquanto conhecimentos semântico e pragmático foram utilizados por Paraboni (1997).

Quanto aos resultados obtidos para a RA pronominal do português, sintetizados na Tabela 6, o trabalho de Paraboni (1997) mostrou uma alta taxa de sucesso, 86,87%, porém, é preciso ressaltar que a entrada do seu sistema consistiu de textos extraídos diretamente de corpus, anotados manualmente por especialista com informações sobre a superfície da sentença. Este índice é comparável aos índices obtidos por pesquisas relacionadas ao tratamento de outros tipos de anáforas, como os trabalhos apresentados para a língua inglesa (Seção 2.4). Entretanto, o trabalho de Coelho (2005) apresentou um baixo desempenho, que, já justificado, foi devido, em parte, aos erros inseridos pelas ferramentas de pré-processamento e pela natureza do corpus utilizado.

Tabela 6: Avaliação das abordagens de RA pronominal do português

| Proposta | Tipo de Corpus | Tipo de PRON | # PRONs | TS (%) |
|---|---|---|---|---|
| Coelho (2205) Adaptação do algoritmo de Lappin & Leass | Jurídico | Pessoais de terceira pessoa, reflexivos e recíprocos | 297 | 35,15 |
| | Literário | | 696 | 32,61 |
| | Jornalístico | | 225 | 43,56 |
| Paraboni (1997) Abordagem multi-agentes | Legislação ambiental | Pronomes possessivos | | 86,87 |

Esses trabalhos contribuíram para o avanço da RA na língua portuguesa, pois os mesmos cobrem uma gama considerável do fenômeno de referenciação pronominal. Apesar dos resultados insatisfatórios obtidos, por exemplo, por Coelho (2005), alguns corpora foram anotados, especificamente, para tratar os pronomes pessoais anafóricos. Além disso, ele implementou alguns componentes úteis, a saber: um extrator de sintagmas nominais, um extrator de pronomes e um manipulador de sujeito composto, permitindo assim, a utilização futura dos mesmos por novos pesquisadores, que é o que este trabalho faz, além de usar os três corpora anotados por Coelho: o jornalístico é usado como fundamento para um estudo de caso e os outros dois corpora para avaliar o sistema de RA desenvolvido. Além disso, também empregamos os arquivos gerados pelos módulos desenvolvidos por ele como entrada para o sistema, como será apresentado no Capítulo 5.

Frente às abordagens de RA pronominais apresentadas, o algoritmo de Mitkov (2002), que será apresentado no próximo capítulo, foi escolhido como proposta deste mestrado para a resolução de pronomes no português devido aos seguintes fatores: 1) esse algoritmo foi implementado para diversas línguas (inglês, polonês e árabe) e obteve um bom desempenho, demonstrando ser supostamente independente de língua e portável. Além disso, ele ainda não foi implementado para o português; 2) o algoritmo não é baseado em um modelo discursivo do fenômeno anafórico ou dependente de recursos semânticos ou pragmáticos, consistindo simplesmente de um conjunto de heurísticas que são aplicadas a um conjunto de SNs candidatos a antecedentes, o que simplifica a sua implementação.

No próximo capítulo, a abordagem original desse algoritmo é detalhada e uma versão completamente automática do mesmo é apresentada. Essa nova versão diferencia-se da maioria das abordagens que dependem de algum tipo de pré-edição da entrada do algoritmo de RA, ou daquelas abordagens que apenas foram simuladas manualmente. Como exemplos temos: a abordagem de Hobbs, que não foi implementada na sua versão original; em Dagan &

Itai (1991) os pronomes pleonásticos foram removidos manualmente; Lappin & Leass (1994) corrigiram a saída do analisador sintático, utilizado por eles, manualmente; e Paraboni (1997) utilizou como entrada para o seu sistema de RA informações sobre a superfície da sentença obtidas diretamente do corpus e anotadas manualmente.

O desenvolvimento do algoritmo original de Mitkov, segundo ele, decorreu da crescente necessidade de resolvedores anafóricos robustos e de baixo custo que operassem em ambientes de PLN automático de abordagem superficial. Ao contrário das abordagens que exigem a construção de grandes bases de conhecimento e o uso de inúmeros recursos, que poderiam tornar o processo de resolução demasiado trabalhoso e dispendioso, Mitkov propõe um algoritmo que independe da construção de uma base de conhecimento e de *parsing* e emprega um conjunto de fatores que, segundo ele, pode ser aplicado a outras línguas – ele demonstra isso fazendo uma adaptação do seu algoritmo para as línguas polonesa e árabe, como será apresentado no próximo capítulo.

# Capítulo 4 - O algoritmo de Mitkov

O algoritmo de Mitkov reproduz uma abordagem superficial do conhecimento lingüístico que tem como objetivo resolver anáforas pronominais cujos antecedentes são sintagmas nominais. Essa abordagem é superficial, pois evita análises semânticas e sintáticas complexas e utiliza como método fundamental de resolução uma lista de heurísticas denominadas 'indicadores de antecedentes', os já citados fatores de resolução.

Esse algoritmo, na sua abordagem original, é apresentado na próxima seção. As seções seguintes englobam respectivamente: os indicadores de antecedentes que constituem a base da estratégia de resolução desse algoritmo, uma ilustração da execução do mesmo com um exemplo para a língua portuguesa, a avaliação do algoritmo, o caráter multilíngüe dessa abordagem, que nos motiva a implementá-lo para o português e na última seção é mostrada uma reimplementação do algoritmo original – o MARS – *Mitkov's Anaphora Resolution System*, um sistema totalmente automático de RA cujo módulo principal é o algoritmo original de Mitkov com algumas modificações. Esse sistema resolve apenas pronomes pessoais de terceira pessoa e possessivos.

## 4.1 - A abordagem original

Sobre um texto pré-processado por um *parser* e por um extrator de SNs, a abordagem original do algoritmo de RA proposto por Mitkov (2002) realiza os seguintes passos: 1) examina a sentença corrente e as duas sentenças precedentes (se existirem) à anáfora em busca de SNs. 2) Dentre os SNs encontrados, seleciona somente aqueles que concordam em gênero e número com a anáfora e os agrupa em um conjunto de candidatos a antecedentes potenciais. 3) Os SNs desse conjunto de candidatos são pontuados pelos indicadores de antecedente e posteriormente é realizada a soma desses pontos. Essa soma é determinada pela fórmula

$$S = \sum_{i=1}^{n} I_i$$

em que I representa a pontuação atribuída por cada indicador considerado.

Por fim, o SN escolhido como antecedente da anáfora será aquele com a maior soma resultante das pontuações desses indicadores. Dessa forma a anáfora é resolvida. Em casos de candidatos com a mesma soma resultante, escolhe-se como antecedente o candidato que estiver mais próximo da anáfora. A Figura 10 ilustra esse processo de RA.

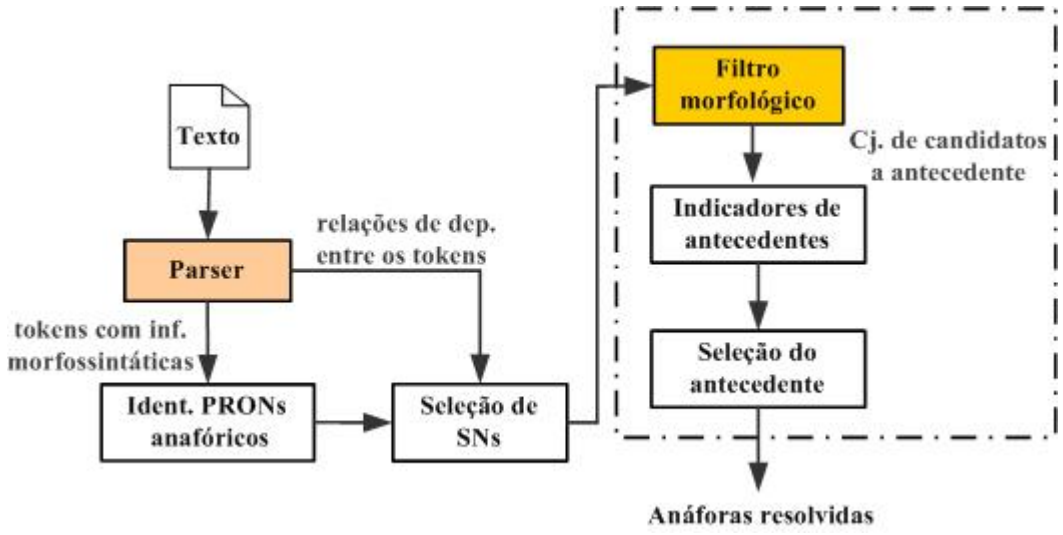



Figura 10: Arquitetura de RA com base no algoritmo de Mitkov

### *4.1.1 - Os indicadores de antecedentes*

Os indicadores de antecedentes utilizados nessa abordagem podem ser: a) promocionais, isto é, que estabelecem *scores* positivos ao candidato a antecedente, ou b) impeditivos, que estabelecem *scores* negativos. Os *scores* positivos refletem a maior probabilidade de um SN ser o antecedente do pronome anafórico e os negativos o contrário.

Os *scores* atribuídos pelos indicadores de antecedentes variam de -1 a +2, sendo que valores maiores que zero promovem o candidato e os valores menores que zero o punem, na soma total dos pesos de cada indicador. Os indicadores são os seguintes:

Primeiro sintagma nominal (PSN): um *score* positivo '+1' é atribuído ao primeiro SN de cada sentença. O uso dessa heurística pode ser justificado com base em estudos que relatam que os seres humanos expressam significados através de níveis de linguagem distintos, dentre eles, o nível denominado Metafunção textual (Ventura & Lima-Lopes, 2002) dá à sentença seu status de mensagem. De acordo com essa definição, um texto coerente deve conter uma estrutura de informação e uma organização temática que permitam que o mesmo possa transmitir alguma mensagem; além disso, essa estrutura permite determinar como a informação flui dentro do texto. A organização temática é realizada

principalmente através da escolha que se faz do elemento que ocupa a posição inicial de cada oração que é enunciada. Assim, cada oração divide-se em duas partes: a primeira, que corresponde ao início da oração, é o tema, e o restante é o rema. O tema estabelece um contexto para a compreensão do que vem a seguir no texto, o rema. E no rema são desenvolvidas as idéias que estão sendo vinculadas pelo tema. O tema representa, portanto a informação previamente dada, a qual é conhecida pelo leitor ou que é recuperável pelo contexto, e o rema constitui a parte que corresponde à sua informação nova. A relação co-referencial pode dar-se entre a informação temática e a informação remática. Uma vez que o tema representa a primeira informação dada, acredita-se que o antecedente da anáfora esteja presente no mesmo.

Verbos indicativos (VI): um *score* '+1' é atribuído àqueles SNs imediatamente seguidos de um verbo membro de um conjunto pré-definido (verbos como: analisar, acessar, apresentar, checar, considerar, cobrir, definir, descrever, desenvolver, discutir, examinar, exibir, explorar, identificar, ilustrar, investigar, revisar, sintetizar, sumarizar, etc.). Mitkov afirma que "evidências empíricas sugerem que sintagmas nominais seguidos dos verbos acima geralmente carregam mais saliência" (Mitkov, 2002: 146)[29].

Reiteração lexical (RL): um *score* '+2' é atribuído aos SNs repetidos duas ou mais vezes no parágrafo no qual o pronome ocorre e um *score* '+1' é atribuído aos SNs repetidos uma única vez nesse mesmo parágrafo. Os itens reiterados lexicalmente são identificados com base em simples semelhança de palavras (*string matching*), mas essa abordagem aceita reiterações lexicais de SNs com o mesmo nome núcleo (e.g. *a bottle, the bottle* ou *toner bottle, bottle of toner, the bottle*). Além disso, não são consideradas reiterações lexicais os SNs que possuem mesmo núcleo e que, no entanto, não são co-referentes (e.g. *the first channel and the second channel*). Por não utilizar nenhuma ontologia, tal como a *WordNet*, sinônimos, hiperônimos ou hipônimos não podem ser recuperados para a indicação de reiterações lexicais.

Este indicador pressupõe que o SN que ocorre duas ou mais vezes dentro do escopo de busca em que aparece o pronome é mais saliente, portanto, mais provável de ser o antecedente da anáfora.

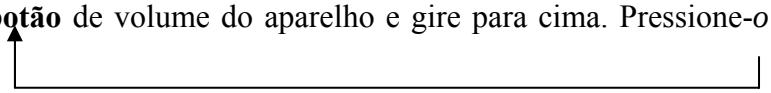
Preferência por SNs em título de seção (PSNTS): um *score* '+1' é atribuído aos SNs que ocorrem no título da seção na qual o pronome anafórico aparece. Esse *score* serve

[29] Nossa tradução.

como complemento do *score* '+1' atribuído pelo indicador reiteração lexical, pois SNs em título de seção não são considerados na delimitação do escopo de busca de tal indicador.

Padrões de colocação (PC): SNs que apresentam o mesmo padrão de ocorrência que o pronome anafórico podem ser o antecedente da anáfora. Um *score* '+2' é atribuído a estes SNs. Os padrões de colocação utilizados limitam-se aos seguintes: <SN/pronome, verbo>, <verbo, SN/pronome>; se o verbo for 'ser/estar', o seguinte padrão também é aceito: <SN/pronome, verbo, adjetivo/particípio>. Vejamos um exemplo:

(4.1) Pressione **o botão** de volume do aparelho e gire para cima. Pressione-*o* novamente.

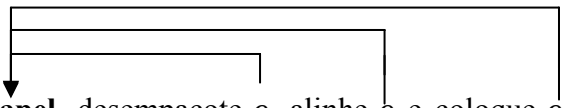


No exemplo (4.1), o padrão de ocorrência dos termos 'o botão' e 'o' é igual nas duas sentenças: <verbo, SN >, <verbo, PRON>, o que permite fazermos uma redução para o padrão <verbo, SN/PRON >. Para este caso, podemos considerar a premissa de que, se um SN possui o mesmo padrão de ocorrência do pronome, este tem o SN como termo antecedente.

Referência imediata: um *score* '+2' é atribuído aos SNs que aparecem em construções do tipo <...$V_1$ SN ...*conjunção* $V_2$ pronome (*conjunção* $V_3$ pronome)>, em que os símbolos < e > delimitam um trecho do texto constituído de orações ligadas por conjunções pertencentes ao conjunto {e, ou, antes, depois, até, ...}, cujos núcleos são os verbos $V_1$, $V_2$ e $V_3$. A primeira oração contém o SN que é o antecedente dos pronomes anafóricos presentes nas orações seguintes.

Esse indicador pode ser visto como uma especificação do anterior, contudo, ele é altamente específico de gênero e ocorre freqüentemente em construções imperativas, bastante comuns em textos de manuais técnicos. O exemplo (4.2) ilustra esse caso:

(4.2) Para imprimir **o papel**, desempacote-*o*, alinhe-*o* e coloque-*o* dentro da gaveta da impressora.



Instruções seqüenciais: um *score* '+2' é aplicado ao SN cuja posição é $NP_1$ na seguinte construção:

<'Para' $V_1$ $SN_1$, $V_2$ $SN_2$. (sentença). 'Para' $V_3$ pronome, $V_4$ $SN_4$>, sendo $SN_1$ o antecedente provável do pronome ($SN_1$ recebe *score* '+2') e a sentença entre parênteses pode ou não estar presente. Vejamos um exemplo:

(4.3) Para ligar **o aparelho de DVD**, pressione o botão *Power*. Para programá-*lo*, pressione o botão '*Programme'*.

Termo preferencial (TP): um *score* '+1' é aplicado aos SNs indicados como termos representativos do gênero textual. Esse indicador é altamente dependente do gênero textual e foi proposto para ser aplicado a textos de manuais técnicos.

Sintagma Nominal Indefinido (SNI): os SNs indefinidos recebem *score* '-1'. Segundo Mitkov, SNs indefinidos, na língua inglesa, que estejam em posição de antecedentes anafóricos são bem menos freqüentes que os SNs definidos, por isso o algoritmo pune candidatos indefinidos. Na implementação desse indicador, Mitkov considera um SN como definido se seu substantivo núcleo é modificado por um artigo definido, ou por pronomes demonstrativos ou possessivos, como mostra o exemplo (4.4):

(4.4) **O parlamentar**, porém, é alvo de acusação em **outro escândalo**. *Ele* será investigado sobre as denúncias de corrupção (...).

Nesse exemplo vemos que o SN 'outro escândalo' será punido por esse indicador, enquanto o SN 'O parlamentar' não, permitindo assim que este possa ser priorizado em relação ao outro como candidato a antecedente do pronome.

Sintagmas nominais preposicionados (SNP): um *score* '-1' é atribuído aos candidatos inseridos em um sintagma preposicional (SP), como ilustra o exemplo 4.5:

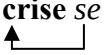
(4.5) (...) Jefferson denunciou uma operação em que o tesoureiro do PT, **Delúbio Soares**, seria o responsável pelo pagamento de mesadas de 30000 a congressistas do PP e do PL. Até o momento *ele* tem afirmado que não há provas.

Nesse trecho de texto, o pronome anafórico 'ele' tem como único antecedente o termo em negrito 'Delúbio Soares'. Entretanto outros SNs também são selecionados como candidatos a antecende ao passarem pelo filtro morfológico. Esses SNs estão representados pelos termos sublinhados.Os mesmos são punidos pelo indicador SNP pois fazem parte de sintagmas preposicionais. Como exemplo, o SN 'o PT' está incluído no sintagma

preposicional ‘do PT’. O SN ‘o PT’ de fato não é o antecedente do pronome ‘ele’, portanto é preterido da lista de candidatos ao se aplicar o indicador de antecedente SNP.

A pontuação negativa atribuída pelo indicador SNP pode ser explicada em termos de saliência, com base na *Centering Theory* (Grosz et. al, 1995). Esta estabelece um sistema de regras e restrições que governam as relações entre o tema do discurso e algumas escolhas lingüísticas efetuadas pelos participantes do discurso, como por exemplo, o emprego de pronomes. Essas regras determinam que o centro da própria sentença ou centros das sentenças anteriores são candidatos altamente prováveis a termo antecedente. Nesta teoria os constituintes da sentença: sujeito, objeto direto e objeto indireto são classificados, nessa ordem, decrescentemente por sua saliência. Esse modelo, então, considera que, se um SN está inserido em um SP, ele provavelmente será o objeto indireto da sentença, portanto é o termo menos saliente da mesma, conforme ilustra o exemplo (4.6).

(4.6) A companhia (...) precisa urgentemente de **uma injeção de capital**. **A crise** *se* arrasta desde os anos 90 (...).



Nesse exemplo vemos que o SN ‘A crise’ é priorizado em relação ao SN ‘uma injeção de capital’ para ser o antecedente da anáfora ‘se’, pois o SN ‘uma injeção de capital’ é pontuado negativamente pelo indicador SNP por fazer parte de um sintagma preposicional, e neste caso, faz parte de um objeto indireto.

Distância referencial: esse indicador pode punir ou promover um candidato a antecedente de acordo com a distância entre ele e a anáfora:

- SNs presentes na cláusula anterior à da anáfora, mas na mesma sentença, recebem *score* ‘+2’.
- SNs presentes na sentença anterior à da anáfora recebem *score* ‘+1’.
- SNs presentes a duas sentenças precedentes à da anáfora recebem *score* ‘0’.
- SNs mais distantes, presentes a mais de duas sentenças anteriores à da anáfora, são assinalados com um *score* ‘-1’. Esse *score* é atribuído somente em versões desse algoritmo que utilizam um escopo de busca de três ou mais sentenças. Portanto, na abordagem original, esse *score* não é atribuído. Contudo, no MARS e neste trabalho ele é utilizado.

Esses são todos os indicadores propostos por Mitkov para processar textos em inglês, totalizando 11 indicadores. Seu uso é ilustrado simulando-se o processo de resolução indicado na Figura 10, para o segmento de texto jornalístico (4.7)[30].

> (4.7) O flúor fortifica o esmalte, uma espécie de capa protetora dos dentes. Com a difusão de seu uso, outro problema surgiu: a fluorose, o excesso de flúor no organismo. Afinal, **a substância** não se encontra apenas na água e cremes dentais: *ela* também está presente em diversos alimentos, (...).

Para encontrar o antecedente do pronome 'ela', o sistema recebe como entrada o texto (4.7) já etiquetado com informações morfológicas e sintáticas, além de receber um arquivo contendo todos os seus sintagmas, conforme ilustra a Figura 10. Todos os SNs presentes no texto são extraídos na ordem em que os mesmos aparecem. Essa extração pode levar à determinação de SNs repetidos, como é o caso dos SNs 10 e 11, abaixo relacionados na Figura 11. É importante que SNs iguais, mas em posições distintas no texto, sejam identificados de forma distinta, pois a sua localização é importante no processo de RA, como visto na Seção 2.3.2. O conjunto de SNs extraídos para o texto em análise é:

$SN_1$ → [O flúor] {masculino, singular}
$SN_2$ → [o esmalte, uma espécie de capa protetora dos dentes] {masculino, singular}
$SN_3$ → [uma espécie de capa protetora dos dentes] {feminino, singular}
$SN_4$ → [capa protetora dos dentes] {feminino, singular}
$SN_5$ → [os dentes] {masculino, plural}
$SN_6$ → [a difusão de seu uso] {feminino, singular}
$SN_7$ → [seu uso] {masculino, singular}
$SN_8$ → [outro problema] {masculino, singular}
$SN_9$ → [a fluorose] {feminino, singular}
$SN_{10}$ → [o excesso de flúor no organismo] {masculino, singular}
$SN_{11}$ → [flúor no organismo] {masculino, singular}
$SN_{12}$ → [o organismo] {masculino, singular}
$SN_{13}$ → [a substância] {feminino, singular}
$SN_{14}$ → [a água] {feminino, singular}

Figura 11: SNs do texto 4.7

[30] Extraído do corpus jornalístico utilizado por Coelho (2005).

Após identificar o pronome ‘ela’ como anafórico, o sistema selecionará como candidatos a antecedentes, dentre os 14 SNs identificados, somente aqueles que passarem pelo filtro morfológico, isto é, os SNs cuja categoria seja feminino, singular, e que estejam presentes em até duas sentenças precedentes à da anáfora. O filtro só selecionará, assim, os SNs com os mesmos traços morfológicos do pronome. São estes os candidatos selecionados:

**Anáfora → [ela]**

$SN_3$ → [uma espécie de capa protetora dos dentes] {feminino, singular}
$SN_4$ → [capa protetora dos dentes] {feminino, singular}
$SN_6$ → [a difusão de seu uso] {feminino, singular}
$SN_9$ → [a fluorose] {feminino, singular}
$SN_{13}$ → [a substância] {feminino, singular}
$SN_{14}$ → [a água] {feminino, singular}

A última etapa de RA, representada na Figura 10, consiste na aplicação dos indicadores de antecedentes ao conjunto de candidatos que passaram pelo filtro morfológico, atribuindo-lhes uma pontuação positiva ou negativa. Posteriormente o somatório das pontuações é calculado e o candidato que está associado com o maior valor é escolhido como antecedente. Na Tabela 7 são apresentados os pesos associados aos 6 SNs anteriores, organizados de forma descendente por seus pesos.

Tabela 7: Indicadores de antecedentes aplicados no processo de RA

| | **Indicadores de antecedentes** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SN candidato** | **PSN** | **VI** | **RL** | **PSTS** | **PC** | **RI** | **IS** | **TP** | **SNI** | **SNP** | **DR** | **∑** |
| *a substância* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | **1** |
| *a água* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | **0** |
| *a fluorose* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| *a difusão de seu uso* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | **-1** |
| *uma espécie de capa protetora dos dentes* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | -1 | **-2** |
| *capa protetora dos dentes* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | **-3** |

Como pode ser visto nessa tabela, o SN ‘a substância’ é selecionado como antecedente da anáfora ‘ela’, devido à sua maior pontuação total, indicada na coluna ‘∑’. Esse resultado demonstra o sucesso do processo de RA descrito para tal texto: o antecedente do

pronome ‘ela’ é de fato ‘a substância’, ambos ocorrendo na mesma sentença no texto. Contudo, podemos perceber que muitos indicadores não contribuíram para esse sucesso, como é o caso do PSN, VI, RL, PSTS, PC, RI, IS e TP, o que poderia sugerir o descarte de tais indicadores para a resolução anafórica desse texto.

Essa abordagem é considerada probabilística já que prediz alguns comportamentos prováveis da língua. Por isso, os indicadores são denominados por Mitkov (2002) de fatores preferenciais, isto é, não são absolutos, mas sim desejáveis. Temos vários exemplos da língua portuguesa em que os mesmos pontuam incorretamente os antecedentes, entretanto, de um modo geral, quando aplicados conjuntamente, os mesmos demonstram eficiência ao apontar o antecedente anafórico.

As próximas seções descrevem a avaliação dessa abordagem, sua adaptação para outras línguas e uma implementação modificada e totalmente automática para o inglês.

### *4.1.2 - Avaliação do algoritmo de Mitkov*

A abordagem apresentada por Mitkov não incorpora conhecimento sintático ou semântico, o que poderia sugerir que os resultados da RA não alcançassem o sucesso das abordagens mais ricas em conhecimento lingüístico. A ausência de informação sintática, por exemplo, não permite a utilização de restrições como c-comando e paralelismo sintático, muito útil na seleção do antecedente, já que para a aplicação desses fatores é necessário conhecer a estrutura sintática do texto.

A abordagem original foi avaliada com vista à obtenção do valor da taxa de sucesso. Esta foi obtida com base em corpora constituídos de textos pré-processados através de etiquetagem e identificação de SNs automáticas e pós-edição manual, para garantir que a entrada do algoritmo fosse correta. Os corpora de teste incluíam diferentes manuais técnicos (*Minolta Photocopier*, *Portable StyleWriter* (PSW), *Alba Twin Speed Video Recorder*, *Seagate Medalist Hard Drive*, *Haynes Car Manual* e *Sony Video Recorderr*), os quais continham um total de 223 pronomes anafóricos. Dentre esses, o processamento do sistema resultou na resolução correta de 200 pronomes, obtendo uma taxa de sucesso de 89,7%.

O resultado dessa avaliação mostrou que, mesmo com suas limitações de conhecimento (sintático ou semântico) no processo de resolução, a abordagem proposta por Mitkov é comparável às abordagens que utilizam conhecimento sintático como a de Lappin & Leass (1994), que obtiveram 86% de acerto sobre os pronomes resolvidos (vide Tabela 2).

Contudo, ressalta-se que a simplicidade do corpus de teste e a pós-edição manual dos arquivos de entrada do sistema contribuíram para tal sucesso na RA.

Outra medida utilizada para avaliação foi a 'taxa de sucesso crítica', definida como a razão entre o número de anáforas resolvidas corretamente (AC) e o número de anáforas com mais de um candidato a antecedente (T+).

$$\text{Taxa de sucesso crítica} = \frac{AC}{T+}$$

Essa medida avalia a eficiência dos indicadores de antecedentes em apontar o antecedente correto da anáfora. Por isso, a taxa de sucesso crítica cobre somente as anáforas que, após terem seus candidatos a antecedentes apontados pelo filtro morfológico, possuem mais de um candidato a antecedente[31]. Ela foi utilizada na avaliação dos textos do corpus da PSW, obtendo-se um resultado de 82%.

Com o intuito de verificar a eficiência de sua proposta e demonstrar que a mesma é superior a modelos *baseline* de RA, Mitkov a comparou com duas abordagens *baseline*, a saber: a) *Baseline* SN: escolhe como antecedente da anáfora o SN que estiver mais próximo da mesma e b) *Baseline* Sujeito: o SN que será selecionado como antecedente se posiciona na função do sujeito em sua sentença e está mais próximo da anáfora. Essa avaliação resultou em uma taxa de sucesso de 65,9% para o primeiro modelo e 48,6% para o segundo, demonstrando assim, a superioridade da abordagem de Mitkov (89,7%).

Além dessas avaliações, Mitkov realizou uma comparação do seu algoritmo com o proposto por Hobbs (1978), com base em parte do corpus de manuais técnicos da PSW. Alcançou-se uma taxa de sucesso de 71% para a abordagem de Hobbs e 83,8 %, para a abordagem de Mitkov. Estes resultados mostram que a abordagem pobre em conhecimento lingüístico proposta por Mitkov teve um melhor desempenho que a abordagem de Hobbs, que utiliza conhecimento sintático.

## 4.2 - A natureza multilíngüe da abordagem de Mitkov

A abordagem de Mitkov, inicialmente desenvolvida para a língua inglesa, foi adaptada e testada para o polonês e árabe. Essa adaptação necessitou de algumas alterações na abordagem original, dentre elas destacamos: a construção de um filtro morfológico para o

[31] Essa medida de avaliação não foi utilizada neste trabalho porque, para os corpora avaliados, todas as anáforas possuíam mais de um candidato a antecedente selecionado pelo filtro morfológico aplicado. Portanto, a taxa de sucesso para esses casos é igual à taxa de sucesso crítica, o que permite desconsiderar o uso dessa medida.

polonês e a inclusão de um indicador de antecedente a mais para o árabe. Os indicadores utilizados na abordagem original foram todos utilizados para ambas as linguagens, modificando-se apenas alguns *scores*, específicos para cada língua.

Mitkov (2002) verificou que as regras de concordância morfológica desempenham um papel proeminente em ambas as línguas, polonês e árabe, filtrando eficientemente muitos candidatos a antecedentes, o que resulta em poucos candidatos para aplicação dos indicadores de antecedente. Talvez esse número pequeno de candidatos justifique a alta taxa de sucesso para essas línguas. A Tabela 8 resume os resultados da avaliação da abordagem de Mitkov aplicada ao polonês e árabe na sua versão original e modificada, adaptada para cada língua, demonstrando a natureza multilíngüe da mesma.

Tabela 8: Avaliação multilíngüe da abordagem de Mitkov

| **Abordagem** | **Taxa de sucesso (%)** | **Taxa de sucesso crítica (%)** |
|---|---|---|
| Inglês | 89,7 | 82 |
| Polonês direto | 90 | - |
| Polonês modificado | 93,3 | 86,2 |
| Árabe direto | 77,9 | 70,4 |
| Árabe modificado | 95,8 | 94,4 |

A avaliação da versão do algoritmo para o polonês foi baseada em manuais técnicos disponíveis na internet, contendo 180 pronomes dos quais 120 eram instâncias de referências exofóricas. A avaliação da versão polonesa original (Polonês direto), que utiliza o algoritmo original sem modificações, resultou em uma taxa de sucesso de 90% e a sua versão modificada (Polonês modificado) obteve um resultado ainda melhor 93,3%, demonstrando o seu sucesso sobre a versão original (inglês), 89,7% .

A abordagem de Mitkov para o árabe também foi avaliada operando de dois modos: o primeiro modo (Árabe direto) consistiu do uso do algoritmo original diretamente, sem nenhuma adaptação ou modificação para o árabe, enquanto o segundo modo incluiu um novo indicador de antecedente (o indicador de pronome relativo). A avaliação desses dois modos se baseou em 190 anáforas contidas nos manuais técnicos da Sony do ano de 1992. A taxa de sucesso do primeiro modo foi de 77,9% (148 de 190 anáforas foram corretamente resolvidas), um resultado bem inferior aos exibidos para o inglês e polonês, como pode ser verificado na Tabela 8. Porém, a taxa de sucesso para o segundo modo de avaliação (Árabe modificado) foi extraordinariamente maior, 95,8%, o que demonstra a sua superioridade frente às duas outras propostas.

Esses resultados e as poucas alterações realizadas demonstram que o algoritmo de Mitkov pode ser implementado para outras línguas, podendo inclusive ter uma taxa de sucesso ainda melhor, como foi comprovado para a sua versão árabe. Ressaltamos que para a avaliação das propostas apresentadas para o polonês e árabe foram utilizados corpora simples (manuais técnicos) anotados manualmente com informações morfossintáticas, portanto não foram introduzidos erros, geralmente, gerados por ferramentas de pré-processamento como etiquetadores e *parsers*.

## 4.3 - MARS: uma reimplementação do algoritmo original de Mitkov

O MARS é uma nova implementação do algoritmo de Mitkov, específico para a língua inglesa. Ele utiliza como principal ferramenta de pré-processamento o *parser Conexor's FDG Parser* (Tapanainen & Järvinen, 1997) e resolve as anáforas determinadas por pronomes pessoais de terceira pessoa e possessivos. O MARS opera de modo completamente automático, o que o torna diferente das abordagens que dependem de algum tipo de pré-edição da entrada do algoritmo de RA, ou daquelas abordagens que apenas foram simuladas manualmente. Vejamos: a abordagem de Hobbs (1978) não foi implementada na sua versão original, em Dagan & Itai (1991) os pronomes pleonásticos foram removidos manualmente e em Lappin & Leass (1994) a saída do analisador sintático utilizado foi corrigida manualmente.

O *parser* utilizado pelo MARS pretende contornar problemas inseridos por ferramentas de pré-processamento automático, como análise morfológica incorreta, o não reconhecimento de nomes próprios, etc. Esse *parser* provê informações sobre relações de dependência entre as palavras, permitindo a extração de sintagmas nominais complexos. Além disso, fornece os lemas e funções sintáticas das palavras. Também foram incluídas outras alterações no MARS: a adição de três novos indicadores, além dos 11 propostos na abordagem original, e a mudança na implementação ou cômputo de cinco desses onze indicadores: reiteração lexical, padrões de colocação, primeiro sintagma nominal, distância referencial e termo preferencial. Essas modificações foram necessárias em função das ferramentas de pré-processamento utilizadas. Os três novos indicadores incluídos no MARS são:

Pronomes: assim como os SNs, os pronomes entram na lista de candidatos de outros pronomes. A motivação para se considerar candidatos pronominais é justificada por

duas características: a) as entidades pronominalizadas tendem a ser salientes, isto é, podem ser o antecedente da anáfora; e b) o SN que corresponde ao antecedente da anáfora pode estar tão distante, que escapa ao escopo de acesso do algoritmo de resolução. Assim, se utilizarmos pronomes como candidatos a antecedentes, eles podem servir como um passo (salto) entre o pronome anafórico e seu antecedente nominal mais distante.

Um *score* '+1' é atribuído aos pronomes antecedentes salientes.

Paralelismo sintático: um *score* '+1' é atribuído ao sintagma nominal que tem a mesma função sintática que a anáfora.

Candidatos freqüentes: Este indicador premia com um *score* positivo '+1' os três SNs que ocorrem com maior freqüência como candidatos de todos os pronomes no texto.

Como o pronome *it*, no inglês, merece tratamento especial, o MARS inclui um módulo que classifica automaticamente cada ocorrência desse pronome em três tipos: anáfora nominal, anáfora não nominal e pronome pleonástico. Além desse módulo, alguns filtros sintáticos foram implementados e são utilizados antes de se aplicar os indicadores de antecedentes e depois da aplicação do filtro morfológico.

O algoritmo implementado pelo MARS segue cuidadosamente a abordagem original (Mitkov, 1998), com as modificações incorporadas nas fases 2 e 3, conforme segue:

**Fase 1:** o texto a ser processado é analisado sintaticamente usando o *parser*, que retorna a categoria gramatical, lema morfológico, função sintática, número gramatical e relações de dependência entre os *tokens* no texto. Essas relações permitem ao sistema a extração de sintagmas nominais do texto.

**Fase 2**: os pronomes anafóricos são identificados e as instâncias não anafóricas e não nominais do pronome '*it*' são filtradas através de um método de aprendizado de máquina descrito por Evans (2001).

**Fase 3**: para cada pronome identificado como anafórico, os candidatos são extraídos dos SNs encontrados nos títulos das seções nas quais os pronomes anafóricos aparecem e dos SNs presentes na sentença corrente e em até duas sentenças anteriores à sentença da anáfora dentro do parágrafo em que a mesma ocorre. Uma vez identificados os candidatos, estes são submetidos a testes morfológicos e sintáticos. Primeiramente são submetidos a um filtro morfológico; depois, os candidatos passam pelos filtros sintáticos 'i' e 'ii', também restritivos (i: um pronome não pode referenciar com um co-argumento; ii: um pronome não pode co-referenciar com um constituinte não-pronominal ao qual ele tanto comanda como precede). Esses filtros sintáticos foram adotados de Kennedy e Boguraev

(1996). Finalmente, os candidatos que passarem pelos filtros acima são agrupados no conjunto de candidatos a antecedentes.

**Fase 4**: nesta fase são aplicados os indicadores de antecedentes, isto é, os fatores promocionais e impeditivos (totalizando 14) pontuam o conjunto de candidatos. Cada fator atribui um *score* numérico a cada candidato, refletindo o quão confiável o candidato é para o sistema, para que possa ser escolhido como antecedente da anáfora.

**Fase 5**: o candidato com a maior composição de *score* é selecionado como antecedente. Caso ocorra empate entre os candidatos, o candidato que estiver mais próximo da anáfora é escolhido como antecedente.

Os *scores* propostos por Mitkov para os indicadores de antecedentes baseiam-se em observações empíricas e não devem ser considerados definitivos ou ótimos. Esses indicadores foram otimizados no MARS utilizando-se um algoritmo genético, que visa encontrar o conjunto de *scores* para os indicadores através do qual a taxa de sucesso de resolução anafórica do algoritmo seja máxima (Orasan & Evans, 2000).

O MARS foi avaliado para diferentes textos de computação: manuais técnicos de hardware e software, contendo um total de 247.401 palavras e 2.263 anáforas pronominais. Dentre estas, 1.709 são intra-sentenciais, enquanto 554, inter-sentenciais. A taxa global de sucesso obtida na resolução pelo MARS foi de 59,35 %. Após utilizar o algoritmo genético de otimização dos *scores* para os indicadores de antecedentes, essa taxa subiu para 61,55%.

Pela avaliação acima, percebe-se que o desempenho do MARS foi bem inferior se comparado ao sucesso da abordagem original. Porém, Mitkov verificou que dos antecedentes existentes nos textos, 238 não foram incluídos na lista de candidatos devido a erros de pré-processamento. Além disso, os dados de entrada para a abordagem original foram corrigidos manualmente, isto é, não houve erros de pré-processamento que contribuíssem para a redução do seu sucesso.

## 4.4 - Considerações sobre a abordagem de Mitkov

Esse capítulo apresentou duas abordagens de resolução anafórica propostas por Mitkov (2002): a abordagem original e a totalmente automática. A primeira independe de *parser*, o que a torna supostamente independente de língua e resolve anáforas pronominais em geral. A segunda é dependente de língua, pois tem um *parser* acoplado como módulo do sistema, resolve apenas pronomes pessoais de terceira pessoa e possessivos, contém um módulo para identificar quando o pronome *it* é não-anafórico, contém filtros sintáticos que

reforçam o filtro morfológico eliminando candidatos indesejáveis, e possui três indicadores a mais que a abordagem original. A avaliação dessas abordagens, conforme ilustra a Tabela 9, mostra que a versão totalmente automática teve um desempenho bem inferior em relação à original, que pode ser justificado devido a erros inseridos pelas ferramentas de pré-processamento, enquanto a abordagem original recebeu uma entrada pós-editada manualmente para correção dos erros gerados pelo *POS-tagger* e pelo extrator de sintagmas nominais utilizados. Além disso, essa abordagem quando empregada para o polonês e árabe também teve uma entrada perfeita, já que consistiu de corpora anotados manualmente por especialistas.

Tabela 9: Síntese da avaliação da abordagem de Mitkov

| **Proposta** | **Descrição do Corpus** | **# PRONs** | **TS (%)** | **TSC (%)** |
|---|---|---|---|---|
| Abordagem original (AO) para o inglês | Manuais técnicos de tecnologia | 223 | 89,7 | 82 |
| AO para o polonês (Polonês Direto) | Manuais técnicos disponíveis na Internet | 60 | 90 | - |
| Adaptação da AO para o polonês (Polonês Modificado) | | | 93,3 | 86,2 |
| AO para o árabe (Árabe Direto) | Manuais técnicos da Sony | 190 | 77,9 | 70,4 |
| Adaptação da AO para o árabe (Árabe Modificado) | | | 95,8 | 94,4 |
| MARS | Manuais técnicos de hardware e software | 2343 | 59,35 | - |
| MARS com algoritmo genético | | | 61,55 | - |

Resumidamente, as abordagens propostas por Mitkov (2002) consistem na aplicação de fatores de resolução (restritivos e preferenciais) para resolver anáforas pronominais em geral (abordagem original) e específicas (anáforas pronominais de terceira pessoa e possessivos, no MARS). Verificou-se que esses fatores de resolução podem ser aplicados a outras línguas, que não o inglês, como pôde ser visto na Seção 4.2, às vezes com desempenho ainda melhor que a proposta original (Tabela 9).

O algoritmo de Mitkov foi escolhido, como já vimos, para ser implementado e adaptado para a língua portuguesa, como esta proposta de trabalho. A escolha desse algoritmo se justifica pelos seguintes motivos: 1) os modelos de RA já explorados para o português não alcançaram índices razoáveis de sucesso, a não ser a proposta multi-agentes de Paraboni (1997), que resolve pronomes possessivos; contudo pretendemos explorar a resolução de pronomes pessoais. 2) O algoritmo de Mitkov foi implementado para diversas línguas (inglês, polonês e árabe) e obteve um bom desempenho, demonstrando ser supostamente independente de língua e portável. Além disso, ele ainda não foi implementado para o

português. 3) O algoritmo não é baseado em um modelo discursivo do fenômeno anafórico ou dependente de recursos semânticos ou pragmáticos, consistindo simplesmente de um conjunto de heurísticas que são aplicadas a um conjunto de SNs candidatos a antecedentes, o que simplifica a sua implementação.

O próximo capítulo apresenta a proposta de resolução anafórica desenvolvida neste mestrado, a qual se resume à adaptação da abordagem original de Mitkov para a língua portuguesa, seguida por um estudo de caso que descreve quais indicadores de antecedentes foram escolhidos neste trabalho considerando a língua em foco e o corpus utilizado no processo de RA.

# Capítulo 5 - A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov

Este capítulo apresenta uma proposta de resolução anafórica para pronomes pessoais de terceira pessoa da língua portuguesa, baseada no algoritmo original de Mitkov (2002). Essa solução não é apoiada em um modelo discursivo do fenômeno anafórico ou dependente de recursos semânticos ou pragmáticos. Ela consiste de uma coleção de heurísticas que são aplicadas a um conjunto de SNs candidatos a antecedentes.

A metodologia utilizada no desenvolvimento desta proposta será descrita com um estudo de caso nas próximas seções e consiste em: 1) análise de um corpus jornalístico e dos indicadores de antecedentes de Mitkov com vistas à escolha dos indicadores aplicáveis a esse corpus; 2) implementação dos indicadores definidos em 1) objetivando calcular o índice de seus acertos e erros ao pontuarem os SNs candidatos, promovendo os candidatos que são os antecedentes e punindo aqueles que não o são; e 3) verificar se tais indicadores, quando aplicados individual ou conjuntamente como estratégia de resolução anafórica, conseguem apontar o antecedente correto da anáfora.

O algoritmo implementado nesse trabalho se diferencia da abordagem original desenvolvida para língua inglesa nos seguintes pontos:

- É específico para a língua portuguesa.
- Utiliza como entrada arquivos já processados com anotações morfossintáticas e anotações sobre as anáforas e SNs[32]. A abordagem original de Mitkov utiliza como estratégia de pré-processamento um segmentador sentencial, um etiquetador e um extrator de SNs. Além disso, ele realiza a correção manual dessas entradas ante os resultados dessas ferramentas.
- O filtro morfológico utilizado, ao encontrar um candidato a antecedente que seja nome próprio, consulta um arquivo XML (se disponível) semelhante ao apresentado na Figura 12, que é gerado por um dicionário onomástico e contém a informação correta sobre o seu gênero e número.

[32] O formato desses arquivos já foi apresentado na Seção 3.2.1.

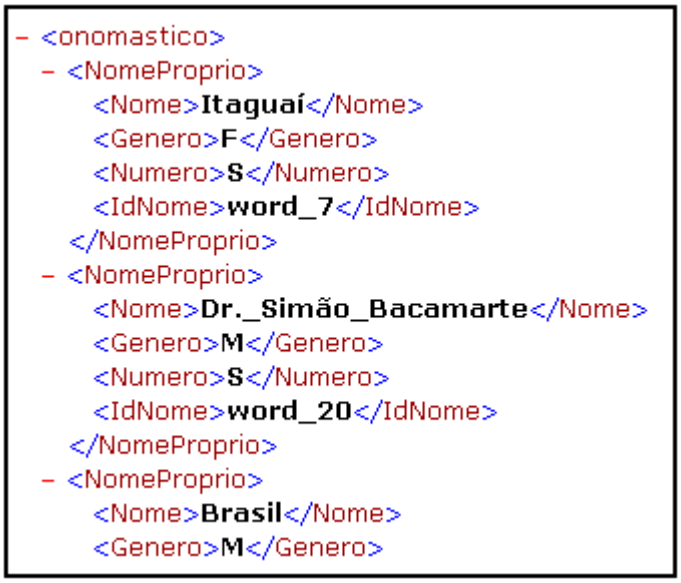



Figura 12: Arquivo gerado por dicionário onomástico

- O escopo de busca de antecedentes é de três sentenças precedentes à da anáfora, isto é, uma janela de quatro sentenças, o que inclui a sentença em que ocorre a anáfora. Já Mitkov utiliza uma janela de três sentenças. Adotou-se esse escopo, pois as propostas de RA de pronomes, para o português, geralmente utilizam esse mesmo escopo, como é o caso do trabalho de Coelho (2005).
- Dos onze indicadores utilizados na abordagem original de Mitkov, apenas cinco são considerados aqui: Primeiro Sintagma Nominal (PSN), Reiteração Lexical (RL), Sintagma Nominal Indefinido (SNI), Sintagma Nominal Preposicionado (SNP) e Distância Referencial (DR). Além desses, mais três indicadores foram incluídos: Nome Próprio (NP), SN mais Próximo (SNMP) e Paralelismo Sintático (PS). Essas escolhas são justificadas na Seção 5.2.

Esse algoritmo foi implementado como um módulo interno a um ambiente de RA, a ser detalhado no Capítulo 6, por nós denominado RAPM (**R**esolução **A**nafórica do **P**ortuguês baseada no algoritmo de **M**itkov) e tem uma arquitetura muito similar à de Coelho (2005). Adotamos tal arquitetura a fim de reutilizarmos os corpora anotados e os módulos implementados por ele, dispensando-nos do trabalho de pré-processamento necessário à entrada do módulo de RA. Essa facilidade possibilitou uma concentração de esforços na mudança do método de RA somente, ou seja, na avaliação dos fatores de resolução de Mitkov ao serem aplicados para o português. Além disso, permitiu-nos realizar uma comparação com

os resultados gerados por Coelho, podendo assim, julgar se a mudança de um único módulo (o principal) faz a diferença para a RA do português.

Como mostra a Figura 13, os arquivos de entrada do RAPM são os mesmos utilizados por Coelho e possuem, portanto, os mesmos formatos (vide Seção 3.2.1).

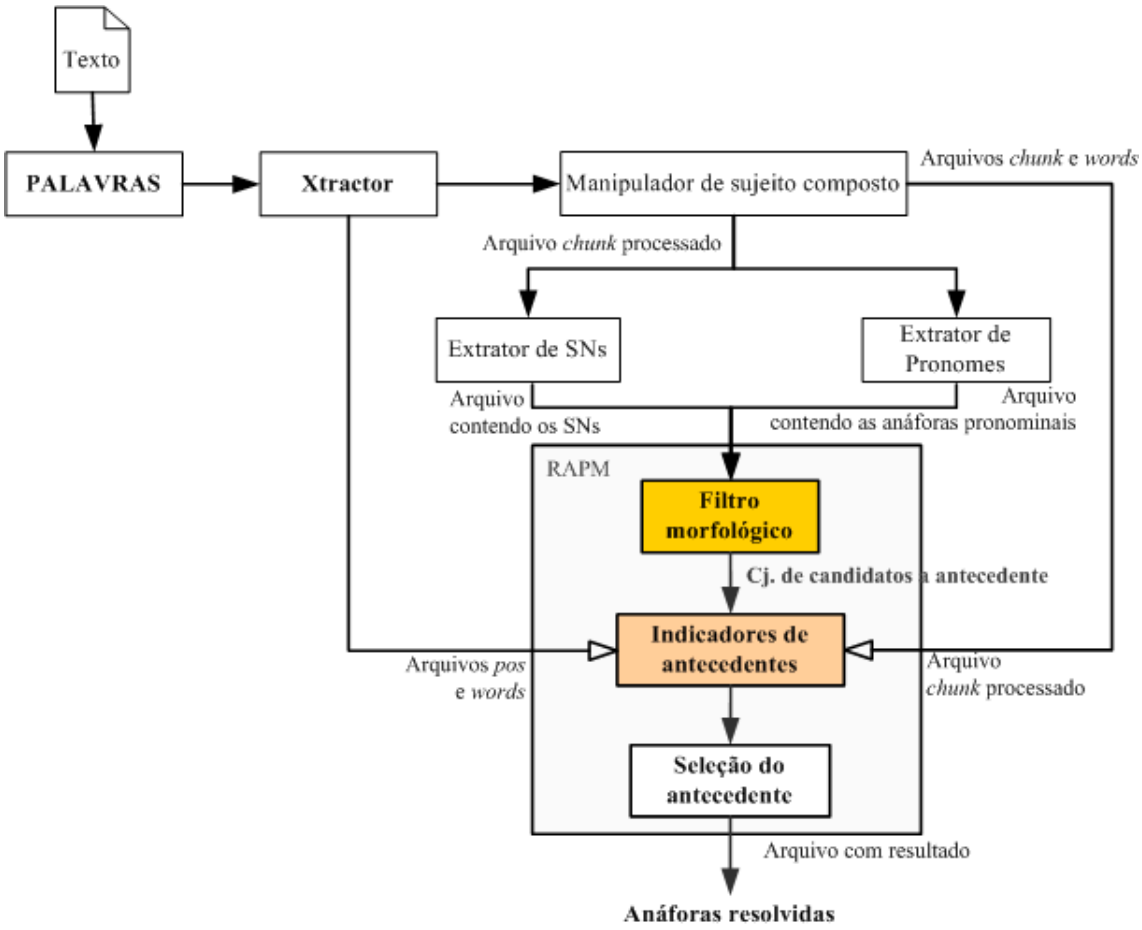



Figura 13: Arquitetura do sistema

Nessa arquitetura, os arquivos de entrada do módulo de RA passam por ferramentas de pré-processamento como o PALAVRAS e o Xtractor; a primeira ferramenta disponibiliza informações morfológicas e sintáticas do texto a ser processado; a segunda, o Xtractor, a partir da saída do PALAVRAS, gera os arquivos contendo as palavras do texto (*word*), as categorias morfológicas das palavras no texto (*pos*) e as estruturas sintáticas das sentenças (*chunks*), os quais são utilizados pelos módulos do sistema de RA. A identificação e agrupamento de sujeitos compostos é realizada pelo manipulador de sujeitos compostos, que gera o arquivo *chunk* processado. Em seguida, são extraídos todos os SNs do texto pelo 'Extrator de sintagmas nominais' e, por fim, o 'Extrator de pronomes', baseado nas etiquetas gramaticais existentes no arquivo de extensão *.pos*, identifica as anáforas pronominais.

Utilizando como entrada os arquivos gerados pelos módulos e ferramentas anteriores, o RAPM resolve as anáforas identificadas aplicando aos SNs encontrados dentro do escopo de busca de cada anáfora, o filtro morfológico, o qual seleciona apenas os SNs que concordam em gênero e número com a anáfora. Aos SNs selecionados são aplicados os indicadores de antecedentes, que apontam o antecedente da anáfora. Este é selecionado e um arquivo de saída contendo as anáforas e seus respectivos antecedentes é gerado.

A próxima seção versa sobre um estudo de caso realizado com o intuito de verificar se as heurísticas utilizadas por Mitkov poderiam ser aplicadas para a língua portuguesa da mesma maneira que foram utilizadas para a língua inglesa, se deveriam ser modificadas ou, até mesmo, se deveriam ser criadas novas heurísticas. Este estudo implicou na delimitação dos indicadores de antecedentes utilizados no RAPM.

## 5.1 - Um estudo de caso sobre indicadores de antecedentes de termos anafóricos para o português

O objetivo geral do estudo de caso foi avaliar a viabilidade da aplicação dos indicadores de antecedentes anafóricos propostos por Mitkov (2002) para a língua inglesa, na resolução de anáforas pronominais da língua portuguesa, com foco nos pronomes pessoais de terceira pessoa.

Para tal estudo as ferramentas Unitex (Paumier, 2006) e Microsoft Visual Studio[33] foram utilizadas. Outras foram desenvolvidas, especialmente, o ambiente que inclui nosso sistema de RA, o RAPM. Ele incorpora um conjunto de módulos: para análise de corpus, aplicação do filtro morfológico, implementação dos indicadores de antecedentes escolhidos, para a própria resolução anafórica e avaliação automática da RA.

Esse estudo utilizou como proposta metodológica a análise de corpus e da representatividade dos indicadores quanto à sua independência de gênero textual e de língua, que levou à escolha de cinco indicadores a serem aplicados no processo de RA para o português. Além disso, foram realizados três experimentos, também descritos nesta seção.

Nas próximas seções serão detalhados o corpus utilizado e sua análise, bem como os três experimentos, seus resultados e as contribuições de cada experimento para a resolução de anáforas pronominais do português.

[33] < http://www.msdnbrasil.com.br/visualStudio/>

### *5.1.1 - Metodologia baseada em corpus*

O corpus adotado é um corpus jornalístico composto por 14 textos contendo uma média de 961 palavras por texto, um total de 13.450 palavras, 2.710 pronomes, dos quais 222 são pronomes de terceira pessoa. Este corpus constitui-se de um conjunto de arquivos utilizados por Coelho (2005) para avaliação da sua proposta de resolução anafórica – a resolução pronominal de anáforas do português baseada no algoritmo de Lappin & Leass (Coelho, 2005; Coelho & Carvalho, 2005).

São dois os pacotes derivados desse corpus: o primeiro, aqui denominado PACOTE-1 é composto por arquivos em formato texto (.txt), texto puro e arquivos anotados automaticamente com informações morfossintáticas pelo *parser* PALAVRAS (Bick, 2000) e informações co-referenciais marcadas manualmente com o auxílio da ferramenta de anotação de discurso MMAX (Müller & Strube, 2001). Além disso, contém arquivos gerados pela ferramenta Xtractor (Gasperin et al., 2003).

O segundo pacote, por nós denominado PACOTE-2, é composto por três tipos de arquivos em formato XML, que estão relacionados com cada um dos arquivos texto (.txt) do PACOTE-1 e foram gerados, respectivamente, conforme ilustra a Figura 13, pelo Manipulador de Sujeito Composto, Extrator de Pronomes e Extrator de SNs desenvolvidos por Coelho (2005), totalizando 42 arquivos XML. O primeiro arquivo contém a estrutura sintática do texto considerado os sujeitos compostos identificados, o segundo contém os pronomes anafóricos identificados e o terceiro, os sintagmas nominais. A Tabela 10 apresenta todos os arquivos contidos em ambos os pacotes com suas respectivas extensões e conteúdos.

Tabela 10: Organização das informações do corpus

| **Arquivo** | **Extensão do arquivo** | **Conteúdo do arquivo** |
|---|---|---|
| **PACOTE-1** | | |
| Texto | .txt | Texto não processado, isto é, texto bruto. |
| Gerados pelo *parser* PALAVRAS | .visl | Texto com etiquetas morfossintáticas e estrutura sintática. |
| Gerados pela ferramenta Xtractor | .words | Palavras do texto identificadas de forma unívoca. |
| | .pos | Informações morfossintáticas das palavras do texto. |
| | .chunk | Estrutura sintática das sentenças e do texto. |
| Gerados pela ferramenta MMax | .markables | Anotações manuais de co-referência. |
| **PACOTE -2** | | |
| Gerados pelo Manipulador de Sujeito Composto | .xml | Estrutura Sintática das sentenças e do texto contendo informação sobre os sujeitos compostos do texto. |
| Gerados pelo Extrator de PRONs | .pron | Pronomes anafóricos do texto. |
| Gerados pelo Extrator de SNs | .np | Sintagmas Nominais do texto. |

Na Figura 14 é mostrado um exemplo do conteúdo de um desses arquivos listados na Tabela 10, o arquivo '.pos'.

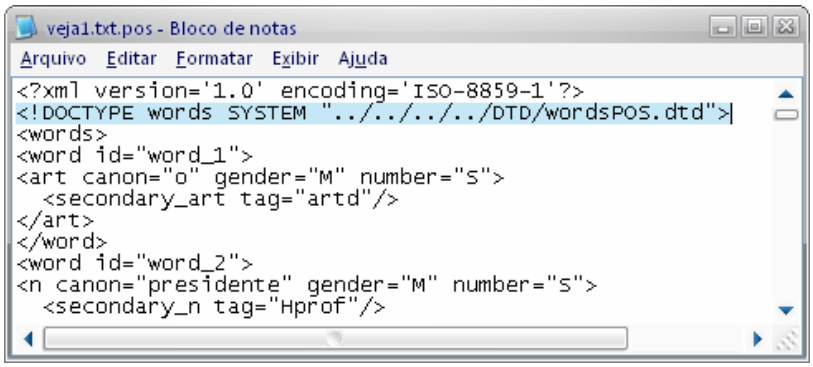



Figura 14: Exemplo de um arquivo .pos

Para que os arquivos do corpus com marcação XML pudessem ser utilizados corretamente pela ferramenta de desenvolvimento Microsoft Visual Studio, um pré-processamento manual foi necessário. Este consistiu em: ajustar-lhes o nome para conter a extensão '*.xml*' (p.ex.: o arquivo veja1.words foi modificado para veja1.words.xml). Apenas os arquivos gerados pelo extrator de sujeito composto não precisaram ser renomeados, pois já continham essa extensão.

Além disso, como é exibido na Figura 14, após o cabeçalho indicador por '<?xml ... ?>', esses arquivos contêm uma linha de código representada pelo texto <!DOCTYPE ... >. A presença desse trecho de código impede que o *Visual Studio* reconheça o arquivo como sendo um XML válido. Por isso é necessário removê-la e deixar o arquivo como mostra a Figura 15, sem essa linha de código. Ademais, esses arquivos devem ser mantidos dentro de um mesmo diretório de trabalho.

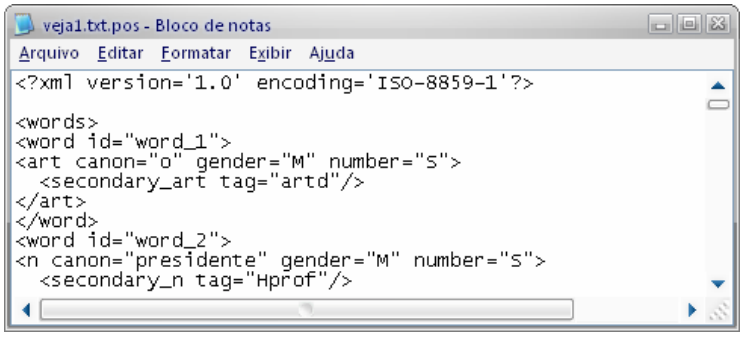



Figura 15: Arquivo .pos modificado

### *5.1.2 - Análise de corpus*

A análise do corpus jornalístico consistiu em averiguar se os indicadores de antecedentes de Mitkov (vide Seção 4.1.1, Capítulo 4) se aplicavam aos textos em português. Como resultado, foram descartados seis indicadores e selecionados cinco (Tabela 11). A escolha desses indicadores se deu pela possível independência de gênero textual e pertinência dos mesmos para a RA de textos em português, como será visto nas próximas seções.

Tabela 11: Indicadores de antecedentes aplicados no processo de RA do português

| **Indicadores escolhidos (5)** | **Indicadores descartados (6)** |
|---|---|
| Primeiro Sintagma Nominal da sentença (PSN) | Verbos Indicativos |
| Reiteração Lexical (RL) | Preferência por SN em Título de Seção |
| SN Indefinidos (SNI) | Padrões de colocação |
| SN Preposicionados (SNP) | Referência Imediata |
| Distância Referencial (DR) | Instruções Seqüenciais |
| | Termo Preferencial |

Ainda foi realizada uma modificação no indicador reiteração lexical: o escopo de busca considerado para analisar a reiteração dos candidatos a antecedentes é diferente do proposto originalmente (parágrafo em que se encontra a anáfora). A estratégia de adaptação utilizada para implementá-lo foi considerar um escopo de busca por reiteração abrangendo até 3 sentenças anteriores à que ocorre a anáfora. Esta limitação do escopo se baseia no fato de que a maioria dos sistemas de resolução anafórica pronominal para a língua inglesa (Hobbs, 1978; Mitkov, 1998; Mitkov, 2002) costuma limitar seu escopo de busca à sentença onde ocorre a anáfora e a duas ou três sentenças anteriores à da anáfora. O mesmo ocorre para a resolução pronominal do português (Coelho, 2005). Além disso, observados os arquivos do corpus que representam as informações sintáticas dos textos-fonte, percebemos que a segmentação textual não considerava a divisão do texto em parágrafos, mas tratava todo ele como um único parágrafo dividindo-o apenas em sentenças, ou seja, a segmentação é apenas sentencial.

O descarte dos demais indicadores é justificado a seguir, discriminando-se cada um deles:

**Verbos indicativos:**

Esse indicador demonstra ser dependente do gênero textual. Os textos utilizados por Mitkov com tal indicador foram manuais técnicos de computação, enquanto o

corpus desse experimento é constituído de textos jornalísticos. Pela análise dos 14 textos do corpus não foi possível localizar nenhum verbo do conjunto especificado por Mitkov e nem mesmo outros verbos que poderiam indicar o gênero jornalístico, pois este tipo textual, geralmente, aborda assuntos diversos, o que torna seu vocabulário bastante abrangente. Essa abrangência não possibilita a identificação de verbos que possam ser agrupados em um conjunto que indique o gênero textual, como ocorre com textos técnicos. Essa constatação possibilitou o descarte desse indicador para a resolução anafórica dos textos do corpus em análise.

**Preferência por título de seção:**

Esse indicador não se aplica a nosso corpus, pois seus textos não contêm títulos de seções, o que impossibilita a sua aplicação.

**Padrão de colocação:**

O conhecimento necessário para a execução do indicador 'padrão de colocação' deve ser adquirido com base em análise de corpus. A partir do corpus são colhidos padrões de ocorrência de SNs e verbos, o que leva esse indicador a uma dependência de gênero textual. Apesar de não desejarmos utilizar um indicador que seja dependente de gênero, ele foi implementado com o intuito de se verificar a ocorrência desses padrões no corpus jornalístico em análise; e pôde ser constatada uma freqüência quase nula (cerca de duas ocorrências no corpus inteiro) de tais padrões.

**Referência imediata e Instruções seqüenciais:**

Ambos os indicadores não se aplicam ao corpus porque os tipos de construções que assinalam não ocorrem em textos jornalísticos e são bem característicos de manuais técnicos.

**Termo preferencial:**

Da mesma forma que ocorreu com o indicador 'verbos indicativos', os textos jornalísticos não possuem uma lista de termos lingüísticos padrões que se repete de um texto a outro e que sirva como indicativo do gênero textual, portanto, esse indicador não pôde ser aplicado para o corpus.

A próxima seção relata o primeiro experimento executado com o intuito de avaliar os indicadores de antecedentes escolhidos nesta seção.

### *5.1.3 - Experimento E1: índices de acerto e erro dos cinco indicadores de antecedentes escolhidos*

O experimento E1 tem por objetivo verificar os índices de acerto e erro de aplicação de cada indicador de antecedente no processo geral de resolução anafórica. Para isso, tais indicadores foram incluídos individualmente no RAPM, que foi executado para cada um dos quatorze textos analisados a fim de verificarmos a pontuação atribuída por eles a todos os candidatos a antecedente que passaram pelo filtro morfológico com o intuito de mensurar os índices de acerto e erro de cada indicador.

O significado expresso pelo acerto e pelo erro varia de acordo com o tipo de indicador de antecedente utilizado, promocional ou impeditivo (restritivo). Os indicadores promocionais são PSN e RL, os impeditivos são SNI e SNP. O indicador DR pode ser promocional ou restritivo, pois as pontuações atribuídas por ele podem variar de -1 a +2. Por isso, nesse experimento, foi feita uma separação do mesmo em dois tipos: DR promocional (DR_P), cuja pontuação varia de 0 a +2, e DR impeditiva (DR_I), cuja pontuação pode ser 0 ou -1.

Para os indicadores promocionais, um acerto (A) representa um fator positivo (P) e denota que o indicador de antecedente promove corretamente o candidato que deveria promover, isto é, atribui um *score* positivo ao candidato a antecedente que também tenha sido anotado manualmente como antecedente da anáfora. Já o erro representa um fator falso positivo (FP) e estabelece que o indicador de antecedente promove candidatos que não deveria promover, isto é, atribui um *score* positivo a candidatos que não foram anotados como antecedentes da anáfora pela anotação manual de co-referência. O acerto está relacionado diretamente com o número de antecedentes válidos de cada texto, enquanto o erro se relaciona com o número total de candidatos a antecedentes que passaram pelo filtro morfológico.

Na Tabela 12 são exibidos os índices de acerto dos indicadores promocionais para cada texto do corpus. Nessa tabela verifica-se que o número total de antecedentes válidos (terceira coluna) é menor que o total de anáforas anotadas (segunda coluna). O acerto é medido somente em função dos antecedentes considerados válidos. Um antecedente é válido caso a sua anotação manual de co-referência não seja ‘nula’ (isto é, uma anáfora sem antecedente) e caso ele tenha sido incluído na lista de candidatos da anáfora. Nessa tabela, para cada indicador, exibimos o número de acertos (A) e a porcentagem (%) desse acerto frente ao número de antecedentes válidos.

Tabela 12: Índice de acerto dos indicadores promocionais

| Texto | # anáforas | # antecedentes válidos | PSN | | RL | | DR_P | |
|---|---|---|---|---|---|---|---|---|
| | | | A | % | A | % | A | % |
| veja1 | 6 | 6 | 2 | 33,33 | 2 | 33,33 | 5 | 83,33 |
| veja2 | 23 | 17 | 9 | 52,94 | 3 | 17,65 | 16 | 94,12 |
| veja3 | 26 | 23 | 6 | 26,09 | 8 | 34,78 | 16 | 69,57 |
| veja4 | 14 | 10 | 3 | 30,00 | 2 | 20,00 | 10 | 100 |
| veja5 | 12 | 4 | 3 | 75,00 | 0 | 0,00 | 3 | 75,00 |
| veja6 | 7 | 5 | 0 | 0,00 | 2 | 40,00 | 5 | 100 |
| veja7 | 24 | 15 | 4 | 26,67 | 1 | 6,67 | 12 | 80,00 |
| veja8 | 8 | 6 | 3 | 50,00 | 2 | 33,33 | 5 | 83,33 |
| veja9 | 9 | 9 | 3 | 33,33 | 2 | 22,22 | 9 | 100 |
| veja10 | 19 | 12 | 7 | 58,33 | 0 | 0,00 | 10 | 83,33 |
| veja11 | 24 | 21 | 14 | 66,67 | 6 | 28,57 | 20 | 95,24 |
| veja12 | 12 | 8 | 4 | 50,00 | 1 | 12,50 | 7 | 87,50 |
| veja13 | 6 | 3 | 2 | 66,67 | 1 | 33,33 | 3 | 100 |
| veja14 | 32 | 17 | 11 | 64,71 | 2 | 11,76 | 17 | 100 |
| **Totais** | **222** | **156** | **71** | **-** | **32** | **-** | **138** | **-** |
| **Médias** | **-** | **-** | **-** | **45,27** | **-** | **21,01** | **-** | **89,39** |

As médias de acertos, exibidas na última linha dessa tabela, demonstram que o indicador DR_P teve o melhor desempenho (89,39%) dentre os três indicadores promocionais avaliados, seguido de longe pelo indicador PSN (45,27%). Esse resultado sugere que o indicador DR_P seja, provavelmente, aquele que melhor aponta o antecedente da anáfora.

O gráfico da Figura 16 ilustra os índices de acerto dos indicadores de antecedentes promocionais para cada texto do corpus. Através dele, nota-se que o indicador DR_P é o fator que mais contribui para o sucesso da RA, pois somente ele, representado pela linha (amarela) do gráfico, ultrapassa a marca de 69 % de acerto. Já o indicador RL quase não apontou os antecedentes, inclusive, para os textos veja5 e veja10, seu índice de acerto foi nulo. Por outro lado, o indicador PSN acerta mais que o RL, mas seu desempenho ainda é considerado baixo frente ao indicador DR_P.

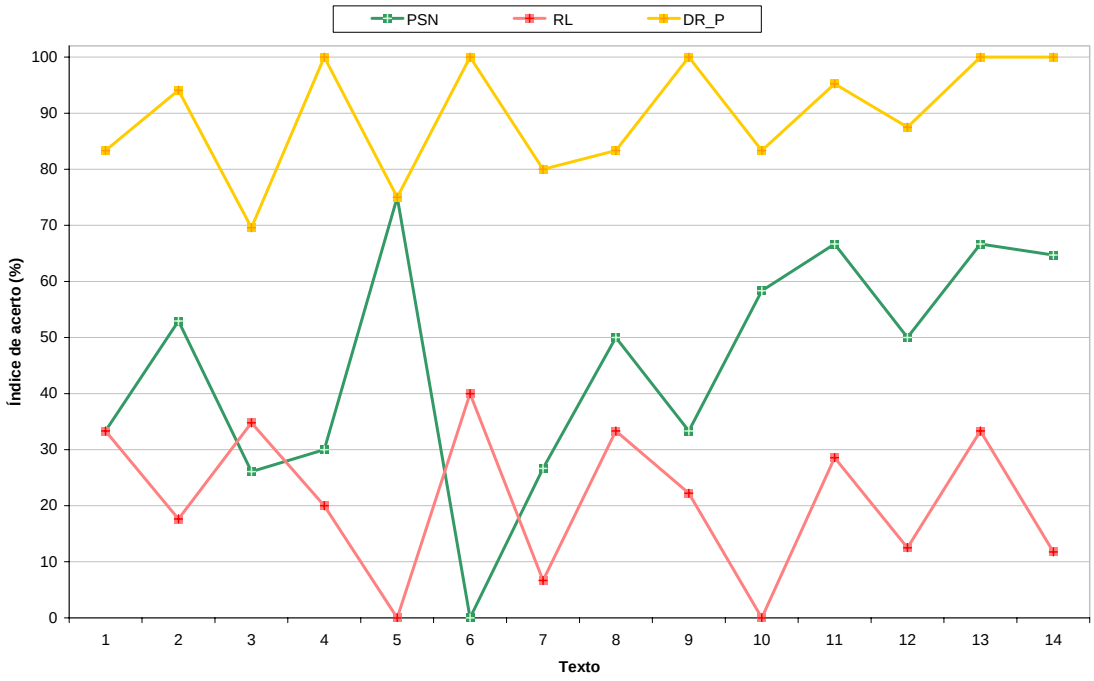



Figura 16: Índice de acerto dos indicadores promocionais

Pode-se observar também que, para o texto veja6, considerado o de melhor desempenho para os indicadores RL e DR_P, o indicador PSN obteve uma taxa de acerto nula. Esse insucesso decorreu da posição dos antecedentes no texto. Todos eles são SNs em posição de objetos em suas sentenças ou sujeitos de orações subordinadas, por isso não se posicionam como primeiro SN da sentença e logo não são promovidos por tal indicador.

Os erros gerados pela aplicação dos indicadores promocionais aos 14 textos do corpus podem ser vistos na Tabela 13. Um erro é medido em função do número de candidatos a antecedente que foi gerado para cada anáfora pelo filtro morfológico, ou seja, o indicador de antecedente erra quando pontua positivamente um candidato que não é o antecedente da anáfora.

Tabela 13: Índice de erro dos indicadores promocionais

| Texto | # candidatos a antecedente | PSN | | RL | | DR_P | |
|---|---|---|---|---|---|---|---|
| | | E | % | E | % | E | % |
| veja1 | 48 | 11 | 22,92 | 0 | 0,00 | 9 | 18,75 |
| veja2 | 157 | 15 | 9,55 | 14 | 8,92 | 68 | 43,31 |
| veja3 | 239 | 38 | 15,90 | 22 | 9,21 | 86 | 35,98 |
| veja4 | 99 | 13 | 13,13 | 4 | 4,04 | 51 | 51,52 |
| veja5 | 34 | 5 | 14,71 | 0 | 0,00 | 8 | 23,53 |
| veja6 | 30 | 2 | 6,67 | 0 | 0,00 | 13 | 43,33 |
| veja7 | 138 | 15 | 10,87 | 17 | 12,32 | 59 | 42,75 |
| veja8 | 41 | 9 | 21,95 | 2 | 4,88 | 12 | 29,27 |
| veja9 | 66 | 10 | 15,15 | 1 | 1,52 | 17 | 25,76 |
| veja10 | 84 | 10 | 11,90 | 0 | 0,00 | 27 | 32,14 |
| veja11 | 187 | 28 | 14,97 | 16 | 8,56 | 64 | 34,22 |
| veja12 | 80 | 16 | 20,00 | 4 | 5,00 | 33 | 41,25 |
| veja13 | 20 | 3 | 15,00 | 1 | 5,00 | 5 | 25,00 |
| veja14 | 230 | 36 | 15,65 | 21 | 9,13 | 93 | 40,43 |
| **Totais** | **1453** | **211** | **-** | **102** | **-** | **545** | **-** |
| **Médias** | **-** | **-** | **14,88** | | **4,90** | | **34,80** |

Pela análise dessa tabela, verifica-se que o número total de candidatos a antecedentes (segunda coluna), 1453, é bem maior que o total de antecedentes válidos (terceira coluna da Tabela 12), 156, o que equivale a uma média de 9,3 candidatos a antecedente por anáfora.

As médias de erros dos indicadores PSN, RL e DR_P são, respectivamente, cerca de 15%, 5% e 35%. Observa-se que o indicador RL, da mesma maneira que acerta pouco ao apontar o antecedente da anáfora, também erra pouco, isto é, pontua poucos candidatos que não deveria pontuar. Essa sua baixa expressividade tanto no acerto (Tabela 12) quanto no erro (Tabela 13) indica que, de fato, ele pouco contribui para o processo de identificação do antecedente. Já o indicador DR_P, apesar de apresentar uma taxa de erro significativa, possui um índice de acerto consideravelmente superior, o que nos leva a concluir que, mesmo pontuando outros candidatos que não são os antecedentes de fato, ele contribui significativamente para a indicação do antecedente correto. O gráfico da Figura17 sintetiza bem a relação entre o número de candidatos pontuados incorretamente pelos indicadores de antecedentes promocionais.

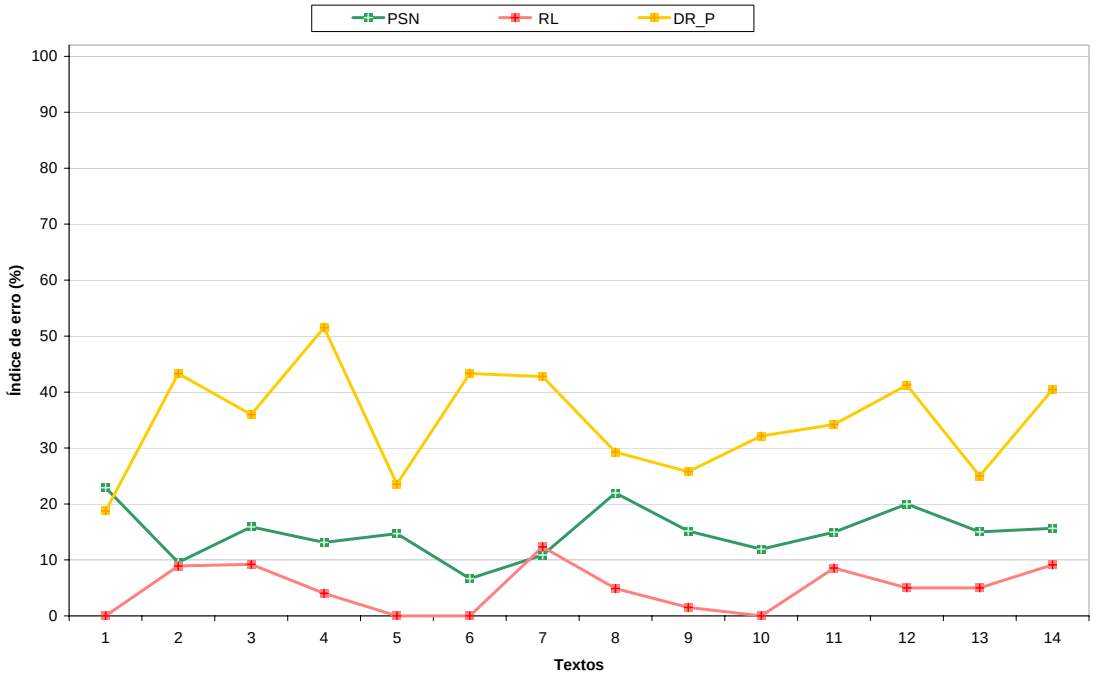



Figura 17: Índice de erro dos indicadores promocionais

Os índices de acerto e erro para os indicadores impeditivos são distintos dos já descritos para os indicadores promocionais. Na Tabela 14, um acerto A é computado por um *score* nulo, isto é, o indicador não impede o candidato que ele não deve impedir. Já o erro pode ser de dois tipos: erro E, exibido na Tabela 15, que é computado com um *score* negativo '-1'. Esse erro ocorre quando o indicador de antecedente impede o candidato que não deveria impedir. Ele representa o inverso do acerto A. Já o segundo erro é computado por uma pontuação igual à do acerto, nula. Ele determina que o indicador de antecedente não impediu o candidato que deveria impedir. Esse erro representa um fator falso negativo e está representado por (FN) na Tabela 16.

Acertos (Figura 18) e erros E (Figura 19) estão diretamente relacionados com o número de antecedentes válidos do texto, pois são calculados em função da pontuação nula ou negativa atribuída ao antecedente, que tenha sido incluído como candidato, pelos indicadores impeditivos. Já o erro FN, ilustrado na Figura 20, está relacionado com o número dos candidatos que passaram pelo filtro morfológico, pois ele é computado em função da pontuação nula atribuída pelos indicadores impeditivos a todos os candidatos que não são os antecedentes das anáforas.

Tabela 14: Índice de acerto dos indicadores impeditivos

| Texto | # antecedentes válidos | SNI | | SNP | | DR_I | |
|---|---|---|---|---|---|---|---|
| | | A | % | A | % | A | % |
| veja1 | 6 | 6 | 100,00 | 3 | 50,00 | 6 | 100,00 |
| veja2 | 17 | 14 | 82,35 | 12 | 70,59 | 17 | 100,00 |
| veja3 | 23 | 19 | 82,61 | 21 | 91,30 | 20 | 86,96 |
| veja4 | 10 | 10 | 100,00 | 5 | 50,00 | 10 | 100,00 |
| veja5 | 4 | 3 | 75,00 | 4 | 100,00 | 4 | 100,00 |
| veja6 | 5 | 4 | 80,00 | 4 | 80,00 | 5 | 100,00 |
| veja7 | 15 | 9 | 60,00 | 7 | 46,67 | 14 | 93,33 |
| veja8 | 6 | 5 | 83,33 | 5 | 83,33 | 6 | 100,00 |
| veja9 | 9 | 9 | 100,00 | 5 | 55,56 | 9 | 100,00 |
| veja10 | 12 | 8 | 66,67 | 7 | 58,33 | 10 | 83,33 |
| Veja11 | 21 | 17 | 80,95 | 19 | 90,48 | 21 | 100,00 |
| Veja12 | 8 | 8 | 100,00 | 7 | 87,50 | 8 | 100,00 |
| Veja13 | 3 | 3 | 100,00 | 2 | 66,67 | 3 | 100,00 |
| Veja14 | 17 | 15 | 88,24 | 15 | 88,24 | 17 | 100,00 |
| **Totais** | **156** | **130** | **-** | **116** | **-** | **150** | **-** |
| **Médias** | **-** | **-** | **85,65** | | **72,76** | | **97,40** |

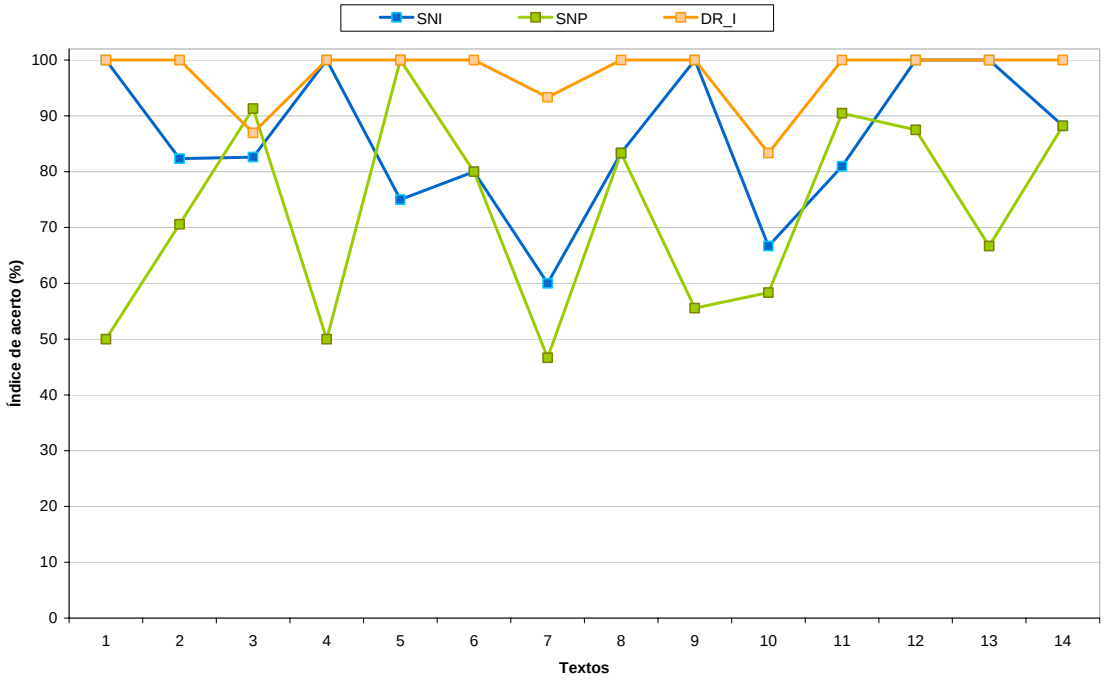



Figura 18: Índice de acerto dos indicadores impeditivos

Tabela 15: Índice de erro E dos indicadores impeditivos

| Texto | # antecedentes válidos | SNI | | SNP | | DR_I | |
|---|---|---|---|---|---|---|---|
| | | E | % | E | % | E | % |
| veja1 | 6 | 0 | 0,00 | 3 | 50,00 | 0 | 0,00 |
| veja2 | 17 | 3 | 17,65 | 5 | 29,41 | 0 | 0,00 |
| veja3 | 23 | 4 | 17,39 | 2 | 8,70 | 3 | 13,04 |
| veja4 | 10 | 0 | 0,00 | 5 | 50,00 | 0 | 0,00 |
| veja5 | 4 | 1 | 25,00 | 0 | 0,00 | 0 | 0,00 |
| veja6 | 5 | 1 | 20,00 | 1 | 20,00 | 0 | 0,00 |
| veja7 | 15 | 6 | 40,00 | 8 | 53,33 | 1 | 6,67 |
| veja8 | 6 | 1 | 16,67 | 1 | 16,67 | 0 | 0,00 |
| veja9 | 9 | 0 | 0,00 | 4 | 44,44 | 0 | 0,00 |
| veja10 | 12 | 4 | 33,33 | 5 | 41,67 | 2 | 16,67 |
| veja11 | 21 | 4 | 19,05 | 2 | 9,52 | 0 | 0,00 |
| veja12 | 8 | 0 | 0,00 | 1 | 12,50 | 0 | 0,00 |
| veja13 | 3 | 0 | 0,00 | 1 | 33,33 | 0 | 0,00 |
| veja14 | 17 | 2 | 11,76 | 2 | 11,76 | 0 | 0,00 |
| **Totais** | **156** | **26** | **-** | **40** | **-** | **6** | **-** |
| **Médias** | **-** | **-** | **14,35** | | **27,24** | | **2,60** |

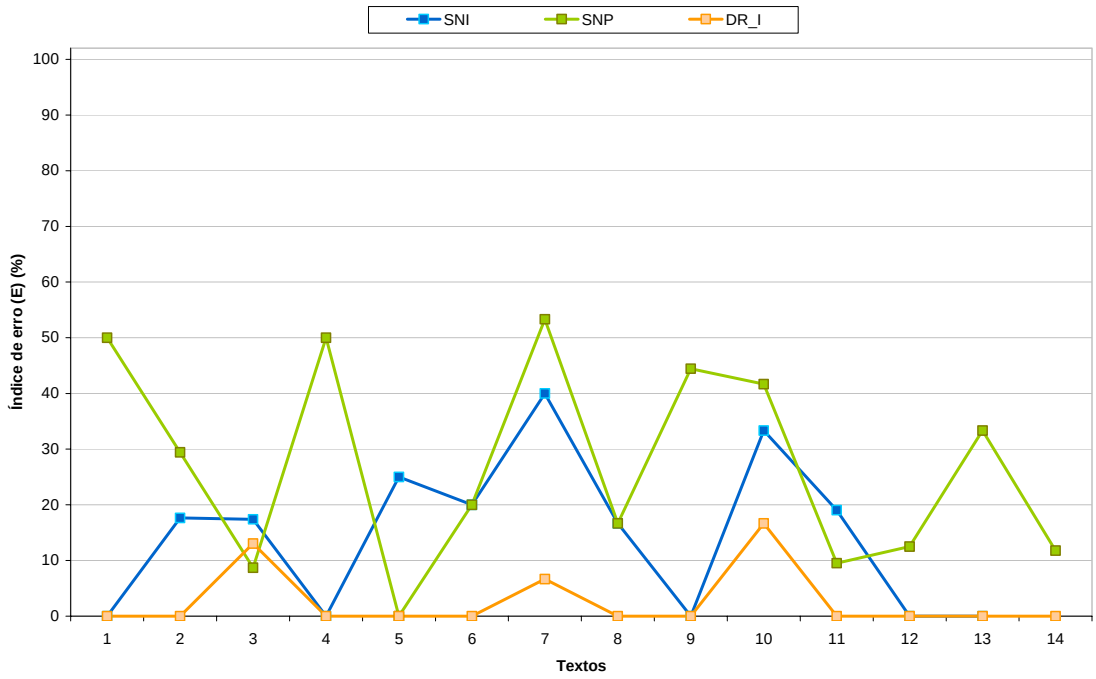



Figura 19: Índice de erro E dos indicadores impeditivos

Tabela 16: Índice de erro FN dos indicadores impeditivos

| Texto | # candidatos a antecedente | SNI | | SNP | | DR_I | |
|---|---|---|---|---|---|---|---|
| | | FN | % | FN | % | FN | % |
| veja1 | 48 | 25 | 52,08 | 21 | 43,75 | 25 | 52,08 |
| veja2 | 157 | 106 | 67,52 | 57 | 36,31 | 109 | 69,43 |
| veja3 | 239 | 140 | 58,58 | 110 | 46,03 | 156 | 65,27 |
| veja4 | 99 | 63 | 63,64 | 46 | 46,46 | 77 | 77,78 |
| veja5 | 34 | 16 | 47,06 | 14 | 41,18 | 22 | 64,71 |
| veja6 | 30 | 13 | 43,33 | 9 | 30,00 | 15 | 50,00 |
| veja7 | 138 | 62 | 44,93 | 60 | 43,48 | 86 | 62,32 |
| veja8 | 41 | 20 | 48,78 | 12 | 29,27 | 21 | 51,22 |
| veja9 | 66 | 32 | 48,48 | 24 | 36,36 | 41 | 62,12 |
| veja10 | 84 | 29 | 34,52 | 32 | 38,10 | 49 | 58,33 |
| veja11 | 187 | 101 | 54,01 | 84 | 44,92 | 129 | 68,98 |
| veja12 | 80 | 45 | 56,25 | 29 | 36,25 | 45 | 56,25 |
| veja13 | 20 | 13 | 65,00 | 7 | 35,00 | 10 | 50,00 |
| veja14 | 230 | 140 | 60,87 | 89 | 38,70 | 164 | 71,30 |
| **Totais** | **1453** | **805** | **-** | **594** | **-** | **949** | **-** |
| **Médias** | **-** | **-** | **53,22** | **-** | **38,99** | **-** | **61,41** |

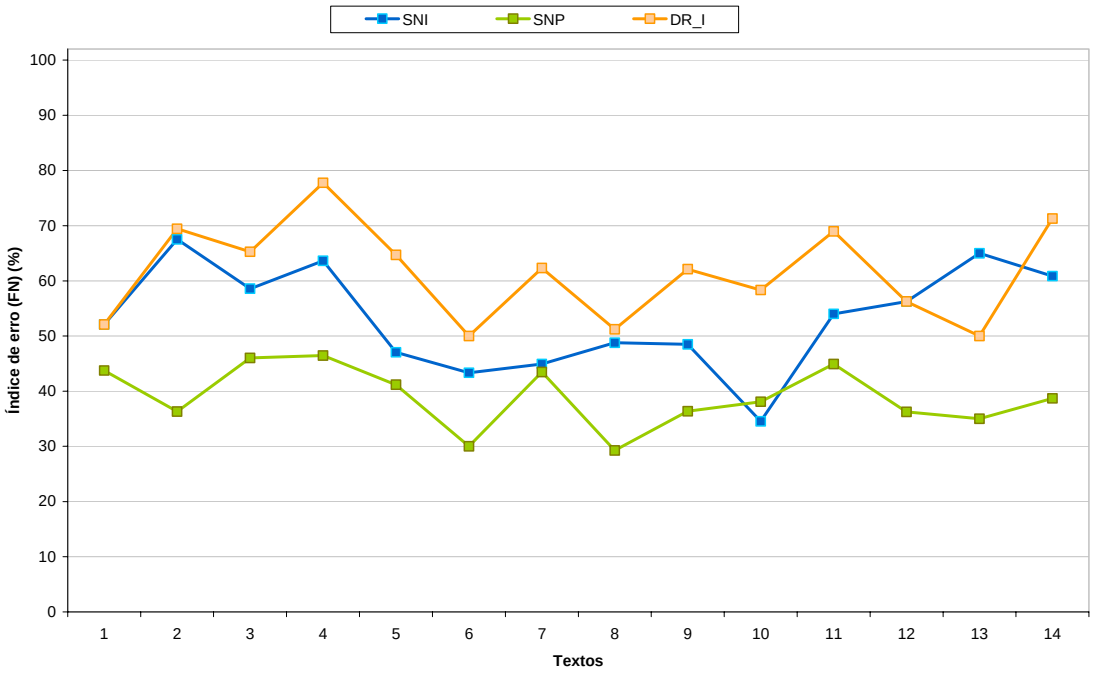



Figura 20: Índice de erro FN dos indicadores impeditivos

A média de acertos, exibida na última linha da Tabela 14 e ilustrada na Figura 18, demonstra que o indicador DR_I teve também o maior índice de acertos; contudo, teve o maior índice de erros FN, isto é, foi o indicador que menos puniu os candidatos que deveria punir, o que permite que candidatos que não sejam os antecedentes da anáfora não sejam

impedidos de concorrer a antecedentes. Esse contraste leva a crer que tal indicador pouco contribui para o sucesso de RA.

O indicador de antecedente SNI se comporta como o indicador DR_I em relação aos índices de acerto e erro; porém, seus índices são proporcionalmente inferiores aos computados pelo DR_I. Já o indicador SNP é aquele que comete mais erros do tipo E, ou seja, esse indicador pune os candidatos a antecedentes que são de fato os antecedentes anafóricos. Porém, esse indicador também gerou o menor índice de erros do tipo FN, ou seja, ele geralmente pune os candidatos que deveria punir. Além disso, vale ressaltar que o erro E, o qual representa uma punição do antecedente, é mínimo (Figura 19) na aplicação desses três indicadores de antecedentes. Por isso, acredita-se que, aplicados conjuntamente, os indicadores impeditivos possam resolver satisfatoriamente as anáforas do texto.

Em vista dessa análise, concluímos que valeria a pena utilizar os indicadores individual e coletivamente como estratégias distintas para a RA. Particularmente, são explorados os seguintes indicadores: PSN, RL, SNI, SNP e DR. A avaliação dessas estratégias é mostrada nos experimentos E2 e E3 apresentados nas seções seguintes.

Salientamos que o desempenho de RA, considerado nos experimentos E2 e E3, consiste em verificar se a solução gerada pelo sistema de RA coincide com a solução anotada manualmente ou com algum outro SN que tenha uma relação de co-referência com esta. Assim, as soluções geradas pelo sistema foram comparadas automaticamente por um módulo da ferramenta de RA desenvolvida, para os casos em que a solução seja exatamente igual à anotação manual, ou caso o núcleo do SN escolhido como antecedente seja o núcleo da solução manual ou pertença a ele. Para os casos de co-referenciação, não considerados na avaliação automática, fez-se a comparação manual dos resultados.

### *5.1.4 - Experimento E2: o uso dos indicadores de forma individual como estratégia de resolução anafórica*

O experimento E2 consistiu em avaliar o sucesso da estratégia de RA quando esta se resume ao uso do algoritmo RAPM restrito a apenas um indicador de antecedente por vez, isto é, após aplicarmos o filtro morfológico aos SNs presentes no escopo de busca da anáfora e gerar um conjunto de candidatos a antecedentes, aplicamos a esse conjunto um dos indicadores de antecedente, atribuindo aos SNs o *score* correspondente desse indicador. O candidato que tiver o maior *score* é escolhido como antecedente da anáfora. Caso haja mais

de um candidato com mesmo *score*, aquele que estiver mais próximo da anáfora é escolhido como antecedente.

A avaliação desse experimento consistiu em comparar a solução gerada automaticamente com a anotação manual de co-referência com o intuito de medir a taxa de sucesso de RA de cada indicador de antecedente, ou seja, a quantidade de anáforas resolvidas corretamente frente ao número de anáforas válidas do texto[34].

Na Tabela 17 é exposta uma síntese dos resultados desse experimento. Para cada texto do corpus são exibidos: o total de anáforas válidas encontradas no texto, o número de anáforas resolvidas corretamente (AR) para cada indicador de antecedente e seu percentual (%) frente ao número total de anáforas válidas.

Tabela 17: Taxa de sucesso de RA dos indicadores de antecedentes

| **Textos** | **Anáforas válidas** | **PSN** | | **RL** | | **SNI** | | **SNP** | | **DR** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **AR** | **TS %** | **AR** | **TS %** | **AR** | **TS %** | **AR** | **TS %** | **AR** | **TS %** |
| veja1 | 6 | 2 | 33,33 | 5 | 83,33 | 5 | 83,33 | 1 | 16,67 | 4 | 66,67 |
| veja2 | 18 | 11 | 61,11 | 7 | 38,89 | 9 | 50,00 | 9 | 50,00 | 8 | 44,44 |
| veja3 | 25 | 5 | 20,00 | 6 | 24,00 | 12 | 48,00 | 12 | 48,00 | 11 | 44,00 |
| veja4 | 12 | 4 | 33,33 | 9 | 75,00 | 11 | 91,67 | 8 | 66,67 | 11 | 91,67 |
| veja5 | 10 | 3 | 30,00 | 0 | 0,00 | 2 | 20,00 | 1 | 10,00 | 0 | 0,00 |
| veja6 | 5 | 3 | 60,00 | 4 | 80,00 | 5 | 100,00 | 5 | 100,00 | 5 | 100,00 |
| veja7 | 21 | 5 | 23,81 | 3 | 14,29 | 7 | 33,33 | 9 | 42,86 | 9 | 42,86 |
| veja8 | 6 | 3 | 50,00 | 4 | 66,67 | 3 | 50,00 | 3 | 50,00 | 4 | 66,67 |
| veja9 | 9 | 5 | 55,56 | 5 | 55,56 | 7 | 77,78 | 7 | 77,78 | 6 | 66,67 |
| veja10 | 16 | 9 | 56,25 | 5 | 31,25 | 4 | 25,00 | 5 | 31,25 | 5 | 31,25 |
| veja11 | 22 | 11 | 50,00 | 11 | 50,00 | 13 | 59,09 | 15 | 68,18 | 10 | 45,45 |
| veja12 | 8 | 4 | 50,00 | 1 | 12,50 | 5 | 62,50 | 5 | 62,50 | 3 | 37,50 |
| veja13 | 3 | 3 | 100,00 | 2 | 66,67 | 3 | 100,00 | 2 | 66,67 | 3 | 100,00 |
| veja14 | 21 | 12 | 57,14 | 6 | 28,57 | 10 | 47,62 | 12 | 57,14 | 10 | 47,62 |
| **Médias** | | | **48,61** | | **44,77** | | **60,59** | | **53,41** | | **56,06** |

Os valores exibidos na última linha dessa tabela representam a taxa de sucesso média de resolução anafórica para cada indicador de antecedente, quando este é utilizado, unicamente, como estratégia de RA. Através do gráfico da Figura 21 visualiza-se claramente essa medida para todos os indicadores avaliados para cada texto individualmente.

[34] Foi considerada uma anáfora válida aquela marcada pela anotação manual de co-referência como uma anáfora com antecedente nominal. Dos 222 pronomes de terceira pessoa anotados, apenas 182 foram considerados neste trabalho.

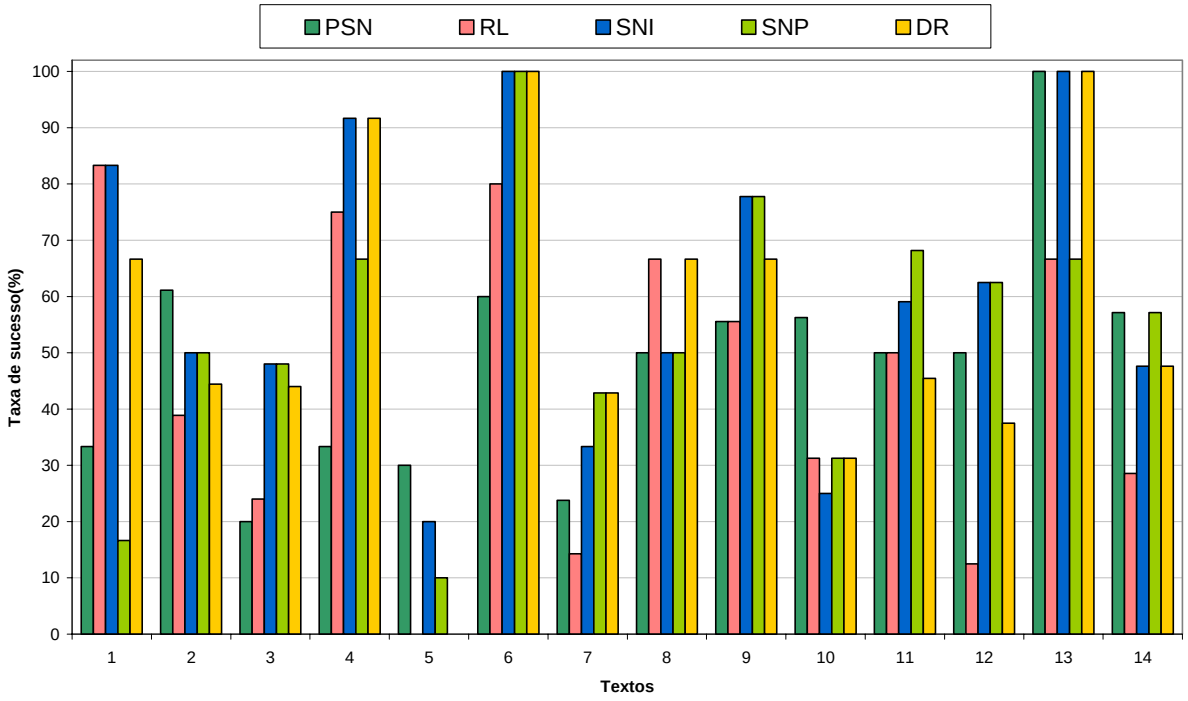



Figura 21: Taxa de sucesso de RA dos indicadores de antecedentes

A média geral de resolução anafórica, conforme ilustra a Tabela 17, demonstra que o melhor desempenho para pontuar os candidatos a antecedentes corretos é do indicador SNI (60,59%), seguido pelo indicador DR (56,06%), enquanto o pior desempenho é resultante do indicador reiteração lexical (44,77%). Conforme ilustra o gráfico da Figura 21, o desempenho geral dos indicadores foi melhor para os textos veja6 e veja13, ao passo que, para o texto veja5, obteve-se o pior desempenho. Inclusive, para esse texto, a utilização dos indicadores RL e DR como estratégias de RA não conseguiram resolver nenhuma anáfora.

O desempenho de cada indicador, ora excelente, como no caso do texto veja 13, em que os indicadores PSN, SNI e DR conseguiram resolver todas as anáforas, ora indesejável, como no texto veja5, em que o indicador RL não resolveu corretamente nenhuma anáfora, pode ser justificado por alguns problemas encontrados na anotação morfossintática dos textos, na extração de seus SNs e, algumas vezes, na própria pontuação do indicador.

O baixo desempenho dos indicadores quando aplicados ao texto veja5 deve-se à natureza de alguns antecedentes e a erros de pré-processamento. Das doze anáforas identificadas, duas não possuíam antecedentes e outras três foram etiquetas com informações morfológicas incorretas, o que impossibilitou a inclusão dos seus antecedentes na lista de candidatos. Além disso, o antecedente de uma das anáforas é uma oração. Uma vez que o sistema se resume a encontrar antecedentes que são SNs, essa anáfora também não pôde ser

resolvida. Ademais, duas outras anáforas possuíam como antecedente um SN complexo, que também não pôde ser identificado pelo sistema e nem mesmo foi indicado pela anotação manual. Esse caso é ilustrado no seguinte trecho do texto veja5.

> (5.1) **O menino** que sai do bueiro não estava brincando com amigos nem fazendo travessura. Morava com **seis outras crianças** debaixo da Avenida Vieira Souto, endereço de artistas, empresários e outros endinheirados da cidade. Havia dois meses *eles se* abrigavam (...).

No exemplo (5.1), o antecedente das anáforas 'eles' e 'se' é um SN complexo formado pela união de dois outros sintagmas, os termos em negrito, 'O menino' e 'seis outras crianças'. O extrator de SNs não consegue identificar o SN completo como 'O menino e seis outras crianças'. Além disso, a anotação manual de co-referência também não o indica como antecedente, mas sim aponta para tais anáforas o SN 'seis outras crianças'. Totalizando oito erros de pré-processamento, as anáforas do texto veja5 não puderam ser corretamente resolvidas.

A ausência de erros de pré-processamento comprovada para os textos veja6 e veja13 e o fato dos antecedentes serem SNs simples vêm comprovar a eficiência da aplicação dos indicadores de antecedentes. A isenção de erros nesses casos permitiu que todos os antecedentes fossem incluídos no conjunto de candidatos e que fossem pontuados pelos indicadores.

Algumas considerações são necessárias sobre os casos em que a taxa de sucesso de RA de alguns indicadores se aproxima, como nos textos veja4 e veja6, nos quais essa taxa ultrapassa 90% para os indicadores SNI e DR e, no entanto, o sucesso geral de RA do texto veja4 é pior que o do texto veja6. Essa inferioridade se deu porque, no texto veja4, para duas anáforas os seus antecedentes não foram incluídos no conjunto de candidatos, pois não foram identificados como SNs completos. Já a proximidade de desempenho dos indicadores SNI e DR deve-se ao fato dos antecedentes desses textos serem SNs definidos e estarem localizados na mesma sentença que a anáfora ou em uma sentença precedente.

Verifica-se também que o indicador SNP, para o texto veja9, obteve um desempenho melhor (77,78%) do que para o texto veja13 (66,67%). Isso ocorreu porque tal indicador puniu incorretamente um dos antecedentes do texto veja13, considerando-o preposicionado. Assim, com uma pontuação negativa, inferior a de outros candidatos, o antecedente não pôde ser escolhido como antecedente. Ilustramos esse caso com o seguinte trecho do texto veja13.

(5.2) Isso é importante porque a insulina é o hormônio responsável por retirar **as moléculas de açúcar** da circulação e jogá-l*as* para dentro das células (...).

Nesse exemplo, o antecedente do pronome oblíquo 'as' (l + as)[35] é o SN em negrito 'as moléculas de açúcar'. Este sintagma nominal não está incluído em um sintagma preposicional e, no entanto, ele foi punido pelo indicador SNP erradamente, pois a identificação de SNs preposicionados é realizada da seguinte maneira: ao encontrar um SN candidato a antecedente, neste caso 'as moléculas de açúcar', percorre-se o texto que antecede o candidato em busca das duas palavras que o precedem. Encontrando-as, é realizada a verificação de sua categoria gramatical. Sendo ela uma preposição, o SN candidato é considerado preposicionado, isto é, o SN é parte constituinte de um sintagma preposicional. Nesse exemplo, o erro ocorreu porque dentre as duas palavras encontradas, 'por' e 'retirar', uma delas é uma preposição. No entanto, essa preposição está ligada ao verbo 'retirar' e não ao SN 'as moléculas de açúcar', portanto o mesmo não pode ser preposicionado.

Descrevemos anteriormente os motivos de divergência entre os melhores e os piores resultados da RA quando se aplica cada indicador de antecedente individualmente como estratégia distinta de resolução anafórica. Uma análise detalhada de cada texto do corpus foi realizada, o que permitiu identificar os principais problemas que levaram à redução do desempenho, como seguem:

**Problema 1: extração de SNs incorreta ou incompleta**

Alguns SNs não foram identificados e sintagmas que não são nominais foram extraídos incorretamente como SNs. No texto veja1, por exemplo, o sintagma adverbial 'até o momento' foi considerado um SN, como mostra o exemplo (5.3):

(5.3) 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 (...).

Nesse exemplo, o pronome 'ele' se refere ao antecedente Jefferson, em negrito, encontrado na sentença anterior à da anáfora. Entretanto, como o sintagma adverbial 'Até o momento' foi extraído como sendo um SN e foi incluído no conjunto de candidatos da anáfora, quando aplicamos os indicadores PSN, SNI, SNP e DR a tal conjunto, ele é escolhido

[35] Os pronomes o, a, os, as, quando associados a verbos terminados em r,s ou z, assumem as formas: lo, la, los, las.

indevidamente como antecedente anafórico, pois foi considerado o primeiro SN da sentença, SN definido, não preposicionado e está mais próximo da anáfora.

Outros erros relacionados à extração de SNs estão relacionados com a não identificação de SNs completos e complexos, como, por exemplo, o SN completo 'o documento em branco', presente no texto veja2, foi particionado em dois SNs simples: 'o documento' e 'branco'. Já o SN complexo do exemplo 5.4, em negrito, representado por 'o "e foram felizes para sempre" ', foi extraído como 'o e'. E este definitivamente não é um SN. O mesmo erro se aplica ao exemplo 5.1, citado anteriormente.

> (5.4) Os problemas que atormentam os casais nesse período em nada lembram **o "e foram felizes para sempre"** dos contos de fada (...).

**Problema 2: o antecedente da anáfora não é um SN**

A ferramenta de RA proposta resolve somente a anáfora cujo antecedente é um SN. Para anáforas cujo antecedente é uma oração, como ocorre no exemplo (5.5), extraído do texto veja2, essa ferramenta certamente não conseguirá resolvê-la de forma correta.

> (5.5) No que se refere à devastação causada pela corrupção na Amazônia, o governo Lula não pode dizer que não teve chance de, ao menos, **contribuir para reduzi-la drasticamente**. Poderia tê-l*o* feito por meio de uma assinatura.

Nesse exemplo a anáfora indicada pelo pronome pessoal oblíquo 'o' tem como antecedente a oração 'contribuir para reduzi-la drasticamente'.

**Problema 3: anotação morfológica das anáforas e SNs**

Geralmente as anáforas representadas pelo pronome 'se' são etiquetadas com gênero masculino-feminino e número singular-plural, já que esse termo é um pronome de dois gêneros e invariante. Essa marcação faz com que o filtro morfológico escolha diversos candidatos a antecedente, o que acarreta a inclusão de antecedentes incorretos, como pode ser visto no exemplo (5.6):

> (5.6) O filho mais velho de Pelé nasceu dois meses depois da conquista da Copa do Mundo de 1970. Em 1975, **o craque** foi contratado pelo Cosmos, dos Estados Unidos, e mudou-*se* com (...).

Nesse exemplo o pronome ‘se’ refere-se ao SN ‘o craque’, que é incluído no conjunto de candidatos a antecedente dessa anáfora. Entretanto, como o filtro morfológico inclui todos os candidatos cujo número seja singular ou plural, os SNs ‘dois meses depois da conquista da Copa do Mundo de 1970’ e ‘os Estados Unidos’ também são incluídos no conjunto de candidatos. Com isso, ao serem aplicados os indicadores de antecedentes a tais candidatos, eles são promovidos e selecionados incorretamente como antecedentes da anáfora. Esses dois sintagmas não deveriam nem ser incluídos no conjunto de candidatos, mas como o *parser* não conseguiu determinar qual o número correto do pronome anafórico quando ele está embutido em um contexto textual, as restrições morfológicas utilizadas nesse trabalho não dão conta de descartar os antecedentes incorretos e, portanto, uma solução inválida do processo de RA pode ser gerada.

Além de anáforas anotadas indevidamente, muitos SNs que são nomes próprios foram etiquetados incorretamente, como por exemplo o nome ‘Victor’, etiquetado com gênero masculino-feminino, foi selecionado como candidato da anáfora ‘a’ (l+a), cujo gênero é feminino, conforme exemplo (5.7).

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

Ainda, há casos em que o antecedente não é incluído na lista de candidatos porque a anáfora está etiquetada incorretamente como singular e o antecedente está no plural ou vice-versa, como em:

(5.8) Para **os habitantes das áreas rurais** em países como Congo, (...) alimentar-*se* de animais (...).

Nesse exemplo o SN ‘os habitantes das áreas rurais’ não foi incluído na lista de candidatos a antecedente da anáfora ‘se’. Portanto, a RA para tal anáfora necessariamente estará incorreta.

**Problema 4: extração de pronomes**

Alguns pronomes foram extraídos como anafóricos e, no entanto, são catafóricos, como mostra o exemplo (5.9):

(5.9) “É melhor elas irem pra lá do que ficarem aqui pegando homem casado.”, diverte-*se* **Carmem Lucia Morais**, coordenadora do Colégio Nossa Senhora.

A catáfora não é foco desse trabalho, portanto, não é resolvida.

**Problema 5: escopo de busca**

Certos antecedentes não foram incluídos no conjunto de candidatos a antecedente porque estavam localizados fora do escopo de busca do sistema. Restringimos esse escopo a três sentenças anteriores à da anáfora; SNs antecedentes em sentenças com distância maior que três sentenças em relação à sentença que inclui a anáfora são descartados. Esse escopo poderia ser aumentado para cobrir tais antecedentes; contudo, o número de candidatos acresceria consideravelmente, aumentando, portanto, a possibilidade do sistema errar ao apontar o antecedente.

Os problemas acima relatados indicam que a taxa de sucesso das estratégias de RA propostas é, geralmente, reduzida devido aos erros de pré-processamento. Acredita-se que aperfeiçoar as ferramentas de pré-processamento ou fazer uma pós-edição manual antes da RA possa melhorar significativamente o desempenho de tais estratégias. Ressalta-se que erros de pré-processamento não foram incluídos na avaliação de Mitkov para o inglês, pois os erros foram corrigidos manualmente pelo autor, o que justifica parte do alto desempenho obtido na avaliação da mesma.

A próxima seção descreve o experimento E3, no qual é avaliada, dentre outras estratégias, a aplicação de todos os indicadores conjuntamente para a RA, aqui denominada *Baseline* Mitkov. Os problemas acima descritos, necessariamente, se reproduzem nessa avaliação e não serão mais mencionados.

.

### *5.1.5 - Experimento E3: o uso dos indicadores de forma conjunta e a resolução anafórica de estratégias baseline*

O experimento E3 consiste em avaliar o sucesso de RA do RAPM quando este utiliza como cerne da resolução três estratégias distintas, como descritas na introdução: a estratégia *Baseline* SN, que escolhe como antecedente da anáfora o SN que estiver mais próximo da mesma; a *Baseline* Sujeito, que identifica como antecedente SNs que são sujeito em suas sentenças e que estejam mais próximos da anáfora; e a estratégia *Baseline* Mitkov,

que aplica conjuntamente todos os indicadores de antecedentes (escolhidos através da análise de corpus descrita na Seção 5.1.2) aos candidatos a antecedente da anáfora. Cada indicador atribui um peso aos candidatos. A soma dos pesos dos indicadores indicará a contribuição total dos mesmos para a RA e, portanto, a significância do candidato em foco, como antecedente da anáfora. O candidato com maior peso é identificado como antecedente. Para os casos de candidatos significativos coincidentes, o SN mais próximo da anáfora é sempre o escolhido.

As estratégias *Baseline* SN e *Baseline* Sujeito foram avaliadas com o intuito de verificar a eficiência e a superioridade da estratégia *Baseline* Mitkov frente às mesmas, como o fez Mitkov (Seção 4.1.2), já que a estratégia de Mitkov representa o foco deste trabalho. Da mesma maneira que no experimento E2, a avaliação dessas três estratégias consistiu em comparar a solução gerada automaticamente com a anotação manual de co-referência.

Na Tabela 18, a síntese dos resultados desse experimento pode ser visualizada; o gráfico da Figura 22 sintetiza-os, evidenciando o melhor desempenho da estratégia *Baseline* Mitkov (média de 60,52%).

Tabela 18: Taxa de sucesso das estratégias *baseline*

| | | ***Baseline* SN** | | ***Baseline* Sujeito** | | ***Baseline* Mitkov (cinco Indicadores)** | |
|---|---|---|---|---|---|---|---|
| **Texto** | **Anáforas válidas** | **AR[36]** | **TS[37] %** | **AR** | **TS %** | **AR** | **TS %** |
| veja 1 | 6 | 4 | 66,67 | 1 | 16,67 | 2 | 33,33 |
| veja 2 | 18 | 8 | 44,44 | 9 | 50,00 | 12 | 66,67 |
| veja 3 | 25 | 11 | 44,00 | 13 | 52,00 | 11 | 44,00 |
| veja 4 | 12 | 11 | 91,67 | 4 | 33,33 | 7 | 58,33 |
| veja 5 | 10 | 0 | 0,00 | 3 | 30,00 | 2 | 20,00 |
| veja 6 | 5 | 5 | 100,00 | 2 | 40,00 | 4 | 80,00 |
| veja 7 | 21 | 8 | 38,10 | 6 | 28,57 | 6 | 28,57 |
| veja 8 | 6 | 4 | 66,67 | 3 | 50,00 | 5 | 83,33 |
| veja 9 | 9 | 6 | 66,67 | 3 | 33,33 | 9 | 100,00 |
| veja 10 | 16 | 5 | 31,25 | 5 | 31,25 | 10 | 62,50 |
| veja 11 | 22 | 11 | 50,00 | 12 | 54,55 | 15 | 68,18 |
| veja 12 | 8 | 2 | 25,00 | 4 | 50,00 | 4 | 50,00 |
| veja 13 | 3 | 3 | 100,00 | 2 | 66,67 | 3 | 100,00 |
| veja 14 | 21 | 10 | 47,62 | 9 | 42,86 | 11 | 52,38 |
| **Média Total** | | | **55,15** | | **41,37** | | **60,52** |

[36] AR: Anáforas Resolvidas corretamente
[37] Taxa de Sucesso: Anáforas resolvidas corretamente / anáforas válidas

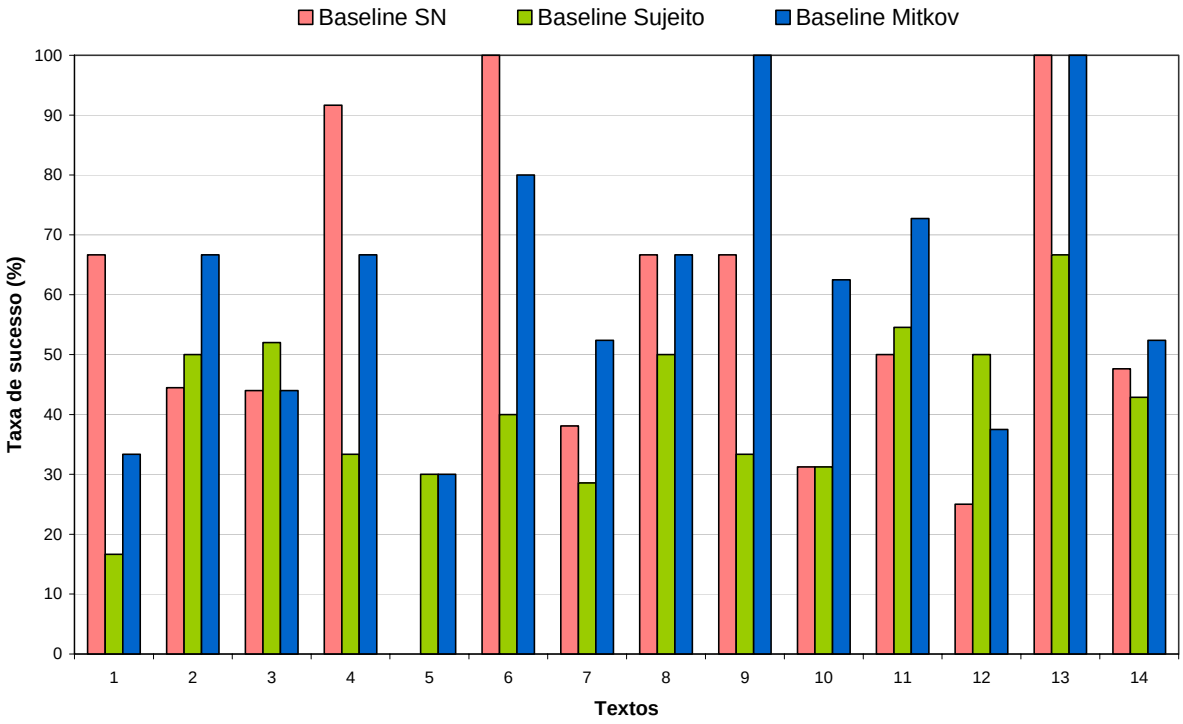



Figura 22: Taxa de sucesso das estratégias *baseline*

Por esse gráfico, verifica-se que todas as estratégias de RA para o texto veja13, como evidenciado também no experimento E2, obtiveram o melhor índice de desempenho, enquanto para o texto veja5 as estratégias *Baseline* SN e *Baseline* Mitkov obtiveram o pior resultado. A estratégia *Baseline* SN, por exemplo, não conseguiu resolver nenhuma anáfora para tal texto. Como já foi visto na Seção 5.1.4, para o texto veja13 não houve erros de pré-processamento, o que justifica o alto índice de resolução de suas anáforas, enquanto, para o texto veja5, das 12 anáforas identificadas, apenas 4 puderam ser bem processadas. Esse baixo desempenho foi causado pelo alto número de erros (8) gerados pelas ferramentas de pré-processamento.

Pela Tabela 18, observa-se que a estratégia *Baseline* Sujeito atingiu a menor taxa de sucesso de RA (41,37%). Um dos motivos do baixo desempenho dessa estratégia é o fato de muitos dos SNs que passaram pelo filtro morfológico não serem os sujeitos em suas orações. Não sendo sujeito, são impedidos de serem escolhidos como antecedentes da anáfora. Dessa forma, as anáforas cujos antecedentes não são sujeitos não são resolvidas.

## 5.2 - Considerações finais

Neste capítulo foi apresentada a arquitetura do sistema de RA proposta neste trabalho, bem como um estudo de caso no qual foi realizada uma análise de corpus e dos indicadores de antecedentes propostos por Mitkov que levou à escolha de alguns indicadores em detrimento de outros. Essa análise teve como propósito verificar quais indicadores poderiam melhor contribuir para a RA do português. Os três experimentos realizados permitiram também identificar erros gerados por ferramentas de pré-processamento bem como avaliar o sistema RAPM, quando esse utiliza estratégias de RA distintas. Ressaltamos que os resultados obtidos são dependentes do corpus jornalístico escolhido, assim como da língua natural, o português.

A avaliação final realizada no experimento E3 demonstra que a estratégia *Baseline* Mitkov, a qual utiliza os cinco indicadores de antecedentes escolhidos no estudo de caso (Seção 5.1.2) e conjuntamente aplicados a um corpus jornalístico, é superior às outras duas estratégias *baseline* também avaliadas. Conforme visto na Seção 4.1.2, Mitkov também utiliza os métodos *baseline* com o intuito de evidenciar a superioridade da sua estratégia de RA.

Os experimentos realizados apontam algumas modificações que, possivelmente, melhorariam o desempenho da RAPM, no que tange à estratégia *Baseline* Mitkov como método de resolução anafórica. São essas:

- A inclusão de um dicionário de nomes próprios (onomástico) a ser utilizado pelo filtro morfológico. Este, ao encontrar um SN que é nome próprio, busca no dicionário onomástico pelo gênero e número corretos deste nome e os compara com os da anáfora. Com isso, pretendemos resolver o problema de etiquetagem incorreta dos nomes próprios.
- Inclusão de um dicionário de sinônimos para o indicador de antecedente 'reiteração lexical', permitindo que esse identifique também as reiterações sinonímias.
- Inclusão de três novos indicadores de antecedentes promocionais: Nomes Próprios, Paralelismo Sintático e SN mais Próximo. Um *score* positivo '+1' é atribuído por esses indicadores aos SNs candidatos que os satisfazem.

A sugestão de inclusão desses novos indicadores é devida aos seguintes fatores: a) a análise dos textos do corpus mostrou que os nomes próprios ocorrem com freqüência como antecedentes de anáforas; portanto, a promoção dos mesmos poderia

melhorar a taxa de sucesso da RA; b) os arquivos utilizados como entrada do ambiente onde o RAPM foi implementado contêm anotações sintáticas, o que justificaria o uso do indicador paralelismo sintático, já que tais informações estão disponíveis; e c) finalmente, SNs que se encontram mais próximos do pronome anafórico tendem a ser o seu antecedente e, portanto, o indicador SN mais Próximo promoveria tais candidatos. Esse último indicador também é utilizado como estratégia de RA, a *Baseline* SN. O próximo capítulo visa apresentar o sistema no qual foi implementada a estratégia RAPM e sua avaliação.

# Capítulo 6 - Implementação e avaliação da proposta

Para facilitar a implementação do algoritmo de Mitkov para o português, foi construído um ambiente que permite acompanhar a análise dos textos pré-processados, acionar o RAPM, além de avaliá-lo automaticamente. Esse ambiente contempla uma interface gráfica amigável[38], cuja descrição completa estará disponível no manual de instruções do mesmo, ainda em desenvolvimento. Ele é composto por quatro módulos distintos:

Módulo 1: é utilizado para análise de corpora. Ele facilita a visualização de alguns dos arquivos que compõem os pacotes do corpus jornalístico, descrito no capítulo anterior, especialmente os arquivos com extensão '.np' (sintagmas nominais), '.words' (arquivo de palavras), '.pron' (pronomes anafóricos) e '.markables' (informações de co-referência).

Módulo 2: Filtro Morfológico. Este módulo é utilizado para restringir, automaticamente, o conjunto de SNs candidatos a antecedentes para cada anáfora a ser resolvida.

Esse filtro é aplicado a todos os SNs presentes no arquivo de sintagmas que estejam dentro do escopo de busca da anáfora. Esse escopo se limita a 4 sentenças, dentre elas a que contém a anáfora e suas três sentenças precedentes, conforme decisão de projeto antes indicada. O filtro verifica, para cada SN do escopo, se o mesmo concorda em gênero e número com a anáfora para, então, incluí-lo no conjunto de candidatos possíveis a antecedentes da anáfora. As informações morfológicas pesquisadas por tal filtro se encontram no arquivo 'pos' já descrito anteriormente. Caso o SN seja um nome próprio, a consulta por gênero e número é realizada no arquivo gerado (se disponível) pelo dicionário onomástico, que contém a informação morfológica correta para os nomes próprios do corpus.

Módulo 3: Resolução anafórica. Esse módulo realiza a resolução anafórica propriamente dita e pode ser subdividido em dois sub-módulos: em um deles é realizada a implementação dos indicadores de antecedentes e no outro, a implementação das estratégias de resolução anafórica. As estratégias podem ser de dois tipos: *baseline*, que utiliza uma heurística para RA e não envolve pontuação de candidatos e a estratégia com base no

[38] Algumas interfaces do ambiente de RA desenvolvido podem ser visualizadas no apêndice desta dissertação.

algoritmo de Mitkov adaptado para o português, o RAPM, que envolve o 'rankeamento' dos candidatos, isto é, utiliza os indicadores de antecedente para pontuá-los.

As estratégias *baseline* são as mesmas que foram utilizadas por Mitkov e descritas na Seção 4.1.2. Elas podem ser de dois tipos: *Baseline* SN, que determina como antecedente o SN que estiver mais próximo da anáfora e *Baseline* Sujeito, que determina como antecedente o SN que for sujeito em sua oração e que estiver mais próximo da anáfora e, caso os SNs que passaram pelo filtro morfológico não sejam sujeitos em suas orações, a anáfora não é resolvida. Essas estratégias, sendo simples, foram utilizadas com o intuito de verificar a eficiência da proposta RAPM frente às mesmas. Os indicadores utilizados pelo RAPM são: PSN, RL, PS, SNMP, NP, SNI, SNP e DR.

Módulo 4: Avaliação da RA. Esse módulo é utilizado para avaliar automaticamente as estratégias de resolução anafóricas empregadas no módulo 3. Essa avaliação consiste em comparar o arquivo anotado manualmente contendo informações de co-referência com o arquivo de resultado gerado automaticamente pelo módulo 3.

Nesse contexto, uma anáfora é considerada corretamente resolvida caso a solução gerada automaticamente seja idêntica à anotada manualmente, ou caso ela seja um SN que é o núcleo ou faz parte do núcleo do SN da anotação manual. A avaliação dessas estratégias, que será exibida na Seção 6.3, utiliza esse módulo como instrumento auxiliar de avaliação, pois as soluções geradas automaticamente que são SNs co-referentes do antecedente anotado manualmente, mas que não são recuperados pela avaliação automática, também foram consideradas corretas; porém, esse módulo da ferramenta não consegue recuperar esse tipo de informação. Portanto, soluções co-referentes foram conferidas.

## 6.1 - Arquivos gerados pelo sistema

O resultado da última etapa da resolução anafórica, a qual identifica o antecedente da anáfora, é salvo em um arquivo XML semelhante ao da Figura 23. Esse arquivo contém todos os pronomes que o algoritmo tentou resolver, acompanhados de seus referentes. Essas informações são armazenadas da seguinte maneira: cada elemento 'RAPM' representa um pronome resolvido juntamente com seu antecedente. Esses elementos são mapeados pelas *tags* 'IdAnafora', 'Anafora', 'IdAntecedente' e 'Antecedente'. As tags 'IdAnafora' e 'IdAntecedente' representam, respectivamente, o chunk que identifica univocamente a anáfora no arquivo de pronomes e o SN no arquivo de SNs, enquanto as *tags*

‘Anafora’ e ‘Antecedente’ representam a anáfora e o antecedente como eles aparecem no texto-fonte.

```
<?xml version="1.0" ?>
- <RA>
  - <RAPM>
      <IdAnafora>chunk_171</IdAnafora>
      <Anafora>ele</Anafora>
      <IdAntecedente>chunk_59</IdAntecedente>
      <Antecedente>Jefferson</Antecedente>
    </RAPM>
  - <RAPM>
```

Figura 23: Arquivo resultante da resolução anafórica

Para efeito de avaliação do acerto dos indicadores, como relatado no capítulo anterior (experimento E1), o filtro morfológico gera um arquivo texto contendo a lista das anáforas resolvidas juntamente com o número de candidatos selecionados para cada uma delas, conforme mostra a Figura 24.

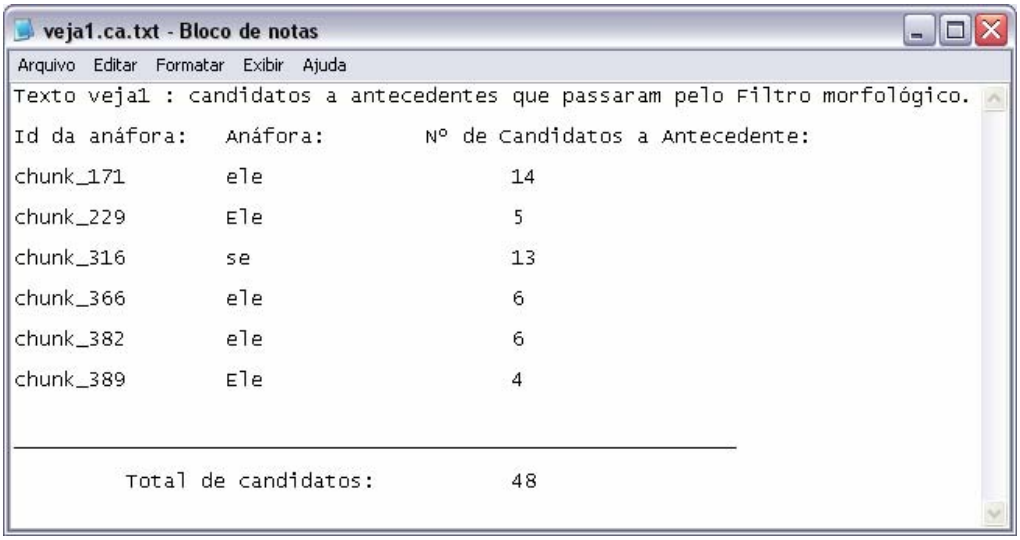



Figura 24: Arquivo gerado pelo filtro morfológico

## 6.2 - Indicadores de antecedentes utilizados

O estudo de caso realizado no capítulo anterior nos levou à escolha de cinco indicadores de antecedentes dentre os propostos por Mitkov para o inglês e à inclusão de três novos, totalizando oito heurísticas, que foram utilizadas como proposta para a resolução das anáforas pronominais do português. Elas foram implementadas obedecendo à descrição das mesmas já relatadas na Seção 4.1.1 e às modificações sintetizadas na Seção 5.2. Os indicadores promocionais utilizados são: Primeiro Sintagma Nominal (PSN), Reiteração

Lexical (RL), Sintagma Nominal mais próximo (SNMP) e Nome Próprio (NP). Os impeditivos são: Sintagma Nominal Indefinido (SNI) e Sintagma Nominal Preposicionado (SNP). O oitavo indicador utilizado é a Distância Referencial (DR) que, conforme visto no capítulo anterior, pode punir ou promover os candidatos a antecedentes, pois atribui um *score* que varia de '-1' a '+2' aos SNs candidatos de acordo com a sua posição em relação à anáfora. SNs mais próximos da anáfora são promovidos enquanto os mais distantes são punidos.

Esses indicadores de antecedentes foram combinados, de maneira *ad-hoc*, como estratégia de RA de diferentes maneiras, sendo, posteriormente, avaliadas. Essas múltiplas combinações foram nomeadas de RAPM_n ($x_1$, $x_2$,...$x_8$), onde n representa a quantidade de indicadores utilizados no cômputo dos candidatos e x representa quais foram os indicadores utilizados para tal cômputo. Foram estas as estratégias combinadas:

- RAPM_2 (SNI e DR).
- RAPM_3 (SNI, SNP e DR).
- RAPM_4 (SNI, SNP, DR e SNMP).
- RAPM_5 (PSN, RL, SNI, SNP e DR).
- RAPM_6_PS (PSN, RL, SNI, SNP, DR e PS).
- RAPM_6_SNMP (PSN, RL, SNI, SNP, DR e SNMP).
- RAPM_6_NP (PSN, RL, SNI, SNP, DR e NP).
- RAPM_8 (PSN, RL, SNI, SNP, DR, PS, SNMP e NP).

A estratégia RAPM_8 utiliza todos os indicadores de antecedentes como estratégia de RA e é considerada a solução final deste trabalho, pois como será visto na Seção 6.3, ao ser avaliada, ela obteve o melhor desempenho frente a todas as combinações realizadas.

A estratégia RAPM_2 foi proposta porque o estudo de caso demonstrou que os indicadores SNI e DR tiveram o melhor desempenho quando aplicados individualmente. Esta estratégia pretende verificar se a combinação desses indicadores é também representativa. Ademais, o indicador SNP obteve o terceiro melhor resultado dentre os cinco indicadores primeiramente avaliados no estudo de caso, por isso foi feita a proposta de RA combinando também esses três indicadores através da estratégia RAPM_3.

O estudo de caso também permitiu que fossem identificadas mais 3 heurísticas utilizadas como indicadores de antecedentes. Elas foram incorporadas à solução final desse trabalho, a RAPM_8. Com o objetivo de verificar a contribuição de cada uma dessas

heurísticas na proposta inicial (avaliada no experimento E3[39] do Capítulo 5), a qual utiliza apenas 5 indicadores (RAPM_5), foi feita uma combinação da RAPM_5, acrescendo à mesma cada um desses três indicadores, a saber: PS, SNMP e NP, gerando as combinações RAPM_6_PS, RAPM_6_SNMP e RAPM_6_NP. A RAPM_4 também foi sugerida, pois desses três novos indicadores o SNMP se mostrou o de melhor desempenho. Assim, foi utilizada uma estratégia que englobasse os 4 melhores indicadores de antecedentes (SNI, SNP, DR e SNMP).

O corpus empregado na avaliação dessas estratégias foi o corpus jornalístico já detalhado no estudo de caso, o qual possibilitou a derivação de todas essas soluções. A seguir apresentamos os resultados obtidos na avaliação dessas estratégias de RA.

## 6.3 - Resultados obtidos

A avaliação da estratégia de RA proposta, a qual utiliza 8 indicadores de antecedentes para pontuar os candidatos, consistiu na determinação da taxa de sucesso de RA da RAPM_8 frente aos modelos *baseline* e às sete outras estratégias de combinações dos indicadores de antecedentes exibidas anteriormente.

O gráfico da Figura 25 apresenta os resultados globais obtidos com o processamento do corpus jornalístico para os 14 textos do corpus.

[39] O filtro morfológico utilizado nesse experimento não acessava os arquivos do dicionário onomástico. Portanto, a única diferença entre o *Baseline* Mitkov e a RAPM_5 é uso do onomástico.

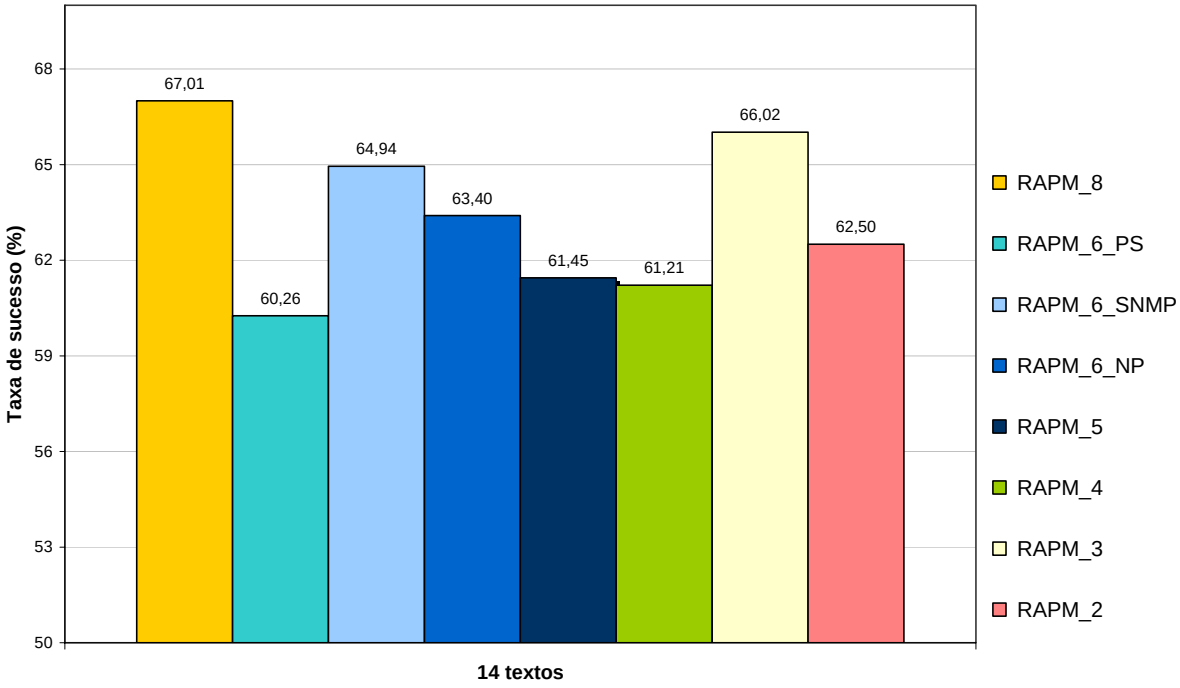



Figura 25: RAPM – Avaliação geral

Como se pode observar, a RAPM_8 obteve o melhor desempenho (67%). No entanto, estratégias mais simples como RAPM_3 e RAPM_2 obtiveram resultado aproximado (66% e 62,5%, respectivamente). Esses resultados sugerem que o uso de alguns indicadores, mais especificamente os indicadores impeditivos, como é o caso do SNI e SNP, quando aplicados para a língua portuguesa, conseguem resolver bem as anáforas de um texto jornalístico. Além disso, essas duas estratégias são mais simples e menos custosas que a RAPM_8.

A proposta RAPM_6_SNMP obteve o terceiro melhor desempenho (64,94%), também muito próximo da RAPM_8. Dentre as três estratégias RAPM_6, essa foi comprovadamente a melhor, o que demonstra que os antecedentes de anáforas pronominais, de fato, encontram-se bem próximos das mesmas e que esse indicador, dentre os três novos propostos (PS, SNMP e NP), é o que mais contribui para o aumento da RA total da estratégia RAPM_8.

Vê-se também que os resultados obtidos por todas essas estratégias, comparados com a abordagem de Coelho (2005), que adapta o algoritmo de Lappin e Leass para o português e também utiliza esse corpus jornalístico em sua avaliação, são superiores. A taxa de sucesso global obtida por Coelho foi de 43,56%, enquanto a estratégia de menor desempenho deste trabalho, a RAPM_5, obteve 61,45% de resolução. Esses resultados

demonstram a superioridade dessa proposta para a língua portuguesa. Para reforçar essa validação foi realizada uma avaliação da RAPM_8[40] com os outros dois corpora utilizados por Coelho. Foram esses: o corpus literário, contendo 573 anáforas válidas e o corpus jurídico, com 260. A Tabela 19 apresenta os resultados globais obtidos com os três corpora avaliados, que confirmam o melhor desempenho da RAPM_8 frente a corpora distintos.

Tabela 19: Comparação da RAPM_8 com a abordagem de Coelho

| | **Taxa de sucesso (%)** | |
|---|---|---|
| **Corpus** | **RAMP_8** | **Coelho (2005)** |
| Jornalístico | 67,01 | 43,56 |
| Literário | 38 | 31,32 |
| Jurídico | 54 | 35,15 |

Uma terceira avaliação foi realizada, com o intuito de comparar a RAPM_8 com as estratégias *baseline* utilizadas também por Mitkov, a fim de demonstrar a sua superioridade frente a estratégias simples. Como resultados obtivemos uma taxa de sucesso de 55,49% para o modelo *Baseline* SN e para o *Baseline* Sujeito 42,27%. Percebe-se que o algoritmo de Mitkov, adaptado para o português, também é superior a estratégias *baselines*.

A adaptação do algoritmo de Mitkov aqui apresentada teve um resultado superior à proposta de Coelho, também utilizada para a RA da língua portuguesa. No entanto, ela demonstrou ser bem inferior em relação às abordagens apresentadas no Capítulo 2, para o inglês. Contudo, é importante frisar que os resultados inferiores gerados pela adaptação de tal abordagem se justificam, principalmente, pela quantidade de erros inserida pelas ferramentas de pré-processamento, como já detalhada na seção 5.1.4, que apresenta os problemas encontrados no processamento do corpus. Dentre os erros destacam-se informações morfológicas incorretas das anáforas e SNs, identificação incorreta de pronomes e de SNs, a existência de antecedentes que não são SNs, caso não tratado pelo módulo de RA RAPM, etc. Acredita-se que a retirada de tais erros possa contribuir para o aumento da resolução das anáforas do corpus. Além disso, a maioria das propostas para o inglês tiveram os dados de entrada do algoritmo de RA corrigidos manualmente, o que resultou em uma entrada perfeita, sem erros, permitindo assim que se avaliasse a eficiência de fato de tais propostas.

Apesar do desempenho dessa abordagem ter sido inferior às abordagens apresentadas para a resolução anafórica do inglês, ela superou a proposta de Coelho, para o

[40] A RAPM_8 foi escolhida para realizar essa comparação porque foi a proposta que obteve o melhor desempenho.

português, devendo, portanto, ser considerada útil no âmbito do processamento de língua natural, como módulo adicional de garantia de coesão para aplicações de PLN como sumarização automática, recuperação de informação, tradução automática, etc.. Esse módulo poderia ser incluído para melhorar os resultados gerados por tais ferramentas.

# Capítulo 7 – Considerações finais

Este trabalho apresentou uma implementação do algoritmo de Mitkov adaptado para a língua portuguesa, que resolve anáforas pronominais de terceira pessoa. Este algoritmo foi avaliado com um corpus jornalístico constituído de 14 textos, contendo um total de 182 anáforas válidas, resultando em uma taxa de sucesso média de 67%. Foi verificado que a taxa de sucesso obtida para a abordagem original desse algoritmo foi bem superior. Esse alto índice de desempenho pode ser decorrente da entrada perfeita utilizada por Mitkov, a qual consistiu de arquivos pré-processados e corrigidos manualmente para remoção dos erros inseridos pelas ferramentas de pré-processamento. Essa correção, talvez, levaria o sistema aqui proposto a obter um desempenho superior ao já reportado. Vale ressaltar que, como já enumerados na Seção 5.4.1, vários foram os erros inseridos pelo pré-processamento. Dentre eles destacamos a extração de SNs incorreta ou incompleta, anotação morfológica de anáforas e SNs incorreta, bem como a extração de pronomes catafóricos. Além disso, algumas anáforas não eram nominais ou o seu antecedente se encontrava fora do escopo de busca considerado nesta implementação, impossibilitando, portanto, que elas fossem resolvidas automaticamente.

Ressalta-se que, dentre as abordagens de RA pronominal desenvolvidas para a língua portuguesa, a proposta de Paraboni (1997) obteve uma taxa de sucesso superior a 85%. Contudo, a entrada do algoritmo proposto por ele consistiu de anotações morfossintáticas manuais do corpus avaliado, dispensando o uso de ferramentas de pré-processamento e, conseqüentemente, resultando em uma entrada ideal, ou seja, livre de erros. Ademais, ele se restringiu à resolução de pronomes possessivos, portanto, não pudemos realizar uma comparação da nossa abordagem com a dele. Já em comparação à proposta de RA desenvolvida por Coelho (2005), o RAPM demonstrou várias melhorias, dentre elas, a maior taxa de sucesso para os três corpora avaliados, bem como, a utilização de heurísticas que permitiram averiguar a relação de algumas estruturações lingüísticas com o fenômeno da referenciarão pronominal, como por exemplo, o posicionamento do SN antecedente na sentença.

As próximas seções apresentam as contribuições desse trabalho, suas limitações e a possibilidade de trabalhos futuros.

## 7.1 - Contribuições

Nesta seção são apresentadas as contribuições obtidas com este trabalho. São elas:

**1) Softwares**

✓ Ambiente automático de RA para o português, composto de quatro módulos: análise de corpus, filtro morfológico, RA e avaliação automática da RA. Esse ambiente é importante para a realização de testes de RA e combinação das várias heurísticas propostas, a fim de identificar as que melhor se aplicam no processo de RA do português, além de permitir, através da análise de corpus, a identificação de novas heurísticas.

✓ Construção de um sistema que apóia a criação de um dicionário onomástico.

**2) Corpora**

Vários corpora foram gerados automaticamente. Dentre eles, os que contêm anotações de co-referência e os que possuem o tempo de processamento de 10 estratégias de resolução anafórica. Além desses, um corpus contendo o número de candidatos a antecedente gerados pelo filtro morfológico para os três corpora utilizados no experimento. São esses:

✓ Corpus_FM: contém 34 arquivos com anotações sobre o número de candidatos por anáfora, que passaram pelo filtro morfológico, e o total geral de candidatos por texto processado.

✓ Corpus_Proc.: contém 174 arquivos com anotações sobre o tempo de processamento de todas as avaliações realizadas com os três corpora: jornalístico, literário e jurídico, utilizando as diversas estratégias de RA. Das estratégias *baseline* às estratégias combinadas do RAPM.

✓ Corpus_NP.: contém 14 arquivos com anotações morfológicas corretas sobre os nomes próprios presentes no corpus jornalístico.

✓ Corpus_J_Baseline: contém 42 arquivos com anotações de co-referência gerados pelas estratégias de RA *Baseline suejito*, *Baseline SN* e *Baseline Mitkov*, obtidos a partir do processamento do corpus jornalístico.

Os corpora denominados Corpus_J_RAPM_n, listados a seguir, contêm, individualmente, 14 arquivos com anotações de co-referência, gerados pela estratégia de RA RAPM_n, obtidos a partir do processamento do corpus jornalístico. Nesse contexto, n representa o número de indicadores de antecedentes utilizados para rankear os candidatos a antecedente. Esses corpora totalizam 112 arquivos anotados. São esses os corpora:

- ✓ Corpus_J_RAPM_2: resultante da utilização dos indicadores de antecedente SNI e DR.
- ✓ Corpus_J_RAPM_3: resultante da utilização dos indicadores de antecedente SNI, SNP e DR.
- ✓ Corpus_J_RAPM_4: resultante da utilização dos indicadores de antecedente SNI, SNP, DR e SNMP.
- ✓ Corpus_J_RAPM_5: resultante da utilização dos indicadores de antecedente SNP, RL, SNI, SNP e DR.
- ✓ Corpus_J_RAPM_6_SNMP: resultante da utilização dos indicadores de antecedente SNP, RL, SNI, SNP, DR e SNMP.
- ✓ Corpus_J_RAPM_6_PS: resultante da utilização dos indicadores de antecedente SNP, RL, SNI, SNP, DR e PS.
- ✓ Corpus_J_RAPM_6_NP: resultante da utilização dos indicadores de antecedente SNP, RL, SNI, SNP, DR e NP.
- ✓ Corpus_J_RAPM_8: resultante da utilização dos indicadores de antecedente SNP, RL, SNI, SNP, DR , SNMP, PS e NP.

Os dois últimos corpora contêm anotações de co-referência, gerados pela estratégia de RA RAPM_8, obtidos a partir do processamento dos corpora literário e jornalístico. Eles foram denominados, respectivamente, Corpus_L e Corpus_Ju. Ambos somam 20 arquivos anotados. A soma total dos arquivos gerados pelo ambiente de RA desenvolvido, agrupados em corpora distintos, totalizam 396 arquivos.

**3) Outras contribuições**

- ✓ Identificação dos diversos erros de pré-processamento que contribuíram para o decréscimo da taxa de sucesso do RAPM.
- ✓ Investigação, pela primeira vez para o português, de vários indicadores de antecedentes, dentre os propostos por Mitkov.
- ✓ Diferentes formas de relacionar esses indicadores, buscando descobrir a combinação mais representativa para a RA do português. Embora essas

combinações não determinem resultados definitivos, como primeira proposta pesquisada, ela demonstrou um olhar curioso sobre as possíveis relações anafóricas entre os pronomes e os indicadores de antecedentes.

✓ Olhar abrangente sobre a RA pronominal, já que este trabalho levou em conta diferentes aspectos relevantes do fenômeno de RA.

✓ O uso de uma metodologia de avaliação diversificada, envolvendo várias estratégias de RA, mesmo que considerando somente a taxa de sucesso como medida de resolução.

✓ A utilização de uma metodologia de avaliação comparativa do cômputo dos índices de acerto e erros dos indicadores. Embora essa avaliação tenha sido manual, esse julgamento humano permitiu um diagnóstico preciso sobre os resultados automáticos.

## 7.2 - Limitações deste trabalho

Este trabalho apresenta algumas limitações, a saber:

1) O ambiente desenvolvido não inclui um módulo de pré-processamento, o que impossibilita a utilização do mesmo para resolver anáforas de textos que não tenham sido anteriormente pré-processados.
2) Ele também não possui um módulo para identificação das anáforas e nem dos SNs presentes nos textos, pois foi implementado para utilizar os arquivos já processados com tais informações, gerados pelos módulos desenvolvidos por Coelho (2005): o extrator de SNs e o extrator de pronomes. A inexistência desses módulos dificulta a replicação dos experimentos para corpora ainda não processados por tais módulos.
3) Falta de sistematização na escolha dos indicadores de antecedentes.
4) Os resultados obtidos através da avaliação dos índices de acerto e erro dos indicadores são limitados, já que se restringiram à avaliação determinada por apenas um juiz.

## 7.3 - Trabalhos futuros

Como aperfeiçoamento deste trabalho, destaca-se a necessidade de se fazer uma avaliação do impacto dos erros introduzidos pelas ferramentas de pré-processamento

utilizadas, quantificando-os, a fim de verificar se é necessário modificar os pesos atribuídos pelos indicadores ou encontrar novos indicadores que possam ser aplicados para melhoria dos resultados. Além disso, é desejável realizar uma avaliação da ferramenta com um corpus maior, inclusive com o corpus SUMMIT[41], este já disponível e anotado com informações co-referenciais; evidenciar a dificuldade de RA do pronome 'se', demonstrando a sua influência nos problemas de RA pronominal e, através de pesquisa sobre as teorias lingüísticas que envolvem tal pronome, tentar diminuir o conjunto de candidatos gerados pelo filtro morfológico ao processar esse pronome, além de determinar a influência de cada pronome nos problemas abordados.

Ademais, seria interessante realizar a inclusão de outros indicadores de antecedentes, como, por exemplo, centro de sentença, além de acrescentar novas restrições, como restrição c-comando e filtros sintáticos; e utilizar um algoritmo genético, como no MARS, para determinar automaticamente a pontuação adequada a ser atribuída por cada indicador de antecedente. Esse algoritmo genético objetiva elevar ao máximo a taxa de sucesso da RA; e, o uso de uma metodologia de combinação de heurísticas, como o fizeram Leite & Rino (2006), poderia contribuir para a descoberta da combinação de indicadores mais adequada e representativa para a RA do português.

Em relação à avaliação, sugere-se que ela seja mais sistemática, principalmente em relação à avaliação manual, que deve incluir mais juizes, a fim de se obter uma concordância satisfatória. Propusemos, também, adotar a mesma estratégia de outros autores como Paraboni e Mitkov: de corrigir os dados de entrada do algoritmo de RA, a fim de medir o acréscimo na taxa de sucesso. Desse modo, estaríamos de fato avaliando a qualidade de nossa proposta e o desempenho do RAPM.

Quanto à ferramenta, como proposta de continuidade deste trabalho, pretende-se incluir na mesma os módulos de identificação de pronomes e de SNs e o *parser* PALAVRAS para que qualquer texto possa ter suas anáforas resolvidas, já que assim teríamos um resolvedor anafórico completo e totalmente automático. Além de utilizar informações onomásticas e semânticas diretamente obtidas do PALAVRAS, com o intuito de investigar se essas informações melhorariam a RA.

[41] Disponível em <http://www.inf.unisinos.br/~renata/laboratorio/corpora.htm>.

# Referências bibliográficas

Allen, J. (1995). *Natural Language Understanding*. Benjamim Commings Publ. Co. Inc..

Bick, E. (2000) *The parsing system PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. Thesis, Århus University, Århus.

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10, 137-167.

Brennan, S. E.; Friedman, M. W.; Pollard, C. J. (1987). *A centering approach to pronouns*. In: Proceedings of the 25th ACL.

Coelho, J.C.B.; Collovini, S.; Vieira, R. (2005). Estudo de corpus para classificação de expressões anafóricas da língua portuguesa. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação* (III Workshop em Tecnologia da Informação e da Linguagem Humana – TIL 2005, pp. 2168-2177. São Leopoldo, RS.

Coelho, J.C.B.; Muller, V.M.; Collovini, S.; Vieira, R.; Rino, L.H.M. (2006) Resolving Portuguese Nominal Anaphora. In: Renata Vieira and Paulo Quaresma (eds.), *Proceedings of the 7th Workshop on Computational Processing of the Portuguese Language - Written and Spoken (PROPOR'2006)*, pp. 160-169. Itatiaia, RJ.

Coelho, T.T. (2005) *Resolução de anáfora pronominal em português utilizando o algoritmo de Lappin e Leass*. Dissertação de Mestrado. Unicamp, SP.

Coelho, T.T. & Carvalho, A.M.B.R. (2005) Uma adaptação de Lappin e Leass para resolução de anáforas em português. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação* (III Workshop em Tecnologia a Informação e da Linguagem Humana – TIL 2005), pp. 2069-2078. São Leopoldo, RS.

Collovini, S.; Coelho, J.C.B.; Vieira, R. (2005) Classificação automática de expressões anafóricas em textos da língua portuguesa. In *Anais do XXV Congresso da Sociedade Brasileira de Computação* (V Encontro Nacional de Inteligência Artificial – ENIA 2005), pp. 942-951. São Leopoldo, RS.

Dagan, I. & Itai, A. (1991) A statistical filter for resolving pronoun references. In: Fedman, Y. A. and Bruckstein, A. (eds.), *Artificial intelligence and computer vision*, pp. 125-135. Elsevier Science Publishers (North-Hollan).

Elbourne, P. D. (2006). *Split antecedents in ellipsis*. Invited talk in the University College London Linguistics Department colloquium series, October 2006.

London Linguistics Department colloquium series, October 2006.

Evans, R. (2001) Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1): 45-57. Oxford, UK.

Fernández, A., Palomar, M.; Moreno L. (1997) Slot unification grammar and anaphora resolution. *Proceeding of the International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, pp. 294-299. Tzigov Chark, Bulgária.

Gasperin, C.V.; Vieira, R.; Goulart, R.R.V.; Quaresma, P. (2003) Extracting xml chunks from portuguese corpora. *Proceedings of the Workshop on Traitement automaticque dês langues ninoritaries (TALN 2003)*. Batz-sur-Mer, France.

Grosz, B. J.; Joshi, A.K.; Weinstein, S. (1995) Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2): 203-225.

Grosz, B.J.; Joshi, A.K.; Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 21(2): 203-225.

Haag, C.R. & Othero, G.A. (2003) Anáforas associativas nas análises das descrições definidas. *Revista Virtual de Estudos da Linguagem – ReVEL*. Ano 1, n.1. Disponível em <http://paginas.terra.com.br/educacao/revel/edicoes/num_1>. Acesso em 13 de jun. de 2006.

Halliday, M.A.K. & Hasan, R. (1976) *Cohesion in English*. London: Longman UK group Limited.

Hobbs, J. R. (1978) *Resolving pronoun references*. Lingua, vol. 44, pp. 311-338.

Jensen, K. (1986) *PEG 1986: a broad-coverage computational syntax of English*. Technical Report, IBM T.J. Watson Research Center.

Kameyama, M. (1997) Recognizing referential links: in information extraction perspective. *Proceedings of the ACL'97 Workshop on Operational Factors in Practical*, *Robust Anaphora Resolution*, pp. 46-53. Madrid, Spain.

Kennedy, C.; Boguraev, B. (1996) Anaphora for everyone: pronominal anaphora resolution without parser. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)I*, pp. 113-118. Copenhagen, Denmark.

Koch, I.G.V. & Travaglia, L.C. (1996) *A coerência textual***.** 7ª ed. São Paulo: Contexto. 94 p.

Koch, I.G.V. (1994) *A coesão textual***. 7**ª ed. São Paulo: Contexto. 75 p.

Lappin, S. & Leass, H.J. (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4): 535-561.

Lappin, S. & McCord, M. (1990a) Anaphora resolution in slot grammar. *Computational Linguistics*, 16(4): 197-212.

Lappin, S. & McCord, M. (1990b) A syntactic filter on pronominal anaphora resolution for slot grammar. In: *28th Annual Meeting of the Association for Computational Linguistics*, pp. 135-142. Morristown, NJ, USA.

Lefffa, V.J. (2001) *A resolução da anáfora no processamento da língua natural*. Relatório final de pesquisa do Núcleo de Pesquisa Lingüística e Literatura da Universidade Católica de Pelotas. Disponível em <http://www.leffa.pro.br/anafor_rel.htm>. Acesso em 15 de jun. de 2006.

Leite, D. S. & Rino, L.H.M. (2006) *SuPor: extensões e acoplamento a um ambiente para mineração de dados*. NILC-TR-06-03, 18 p.

McCord, M. (1990) Slot grammar: a system for simpler construction of practical natural language grammars. In: Studer, R(eds.), *Natural language an logic: international scientific symposium*, pp. 118-145. Lecture Notes in Computer Science. Berlin: Springer Verlag.

Meyer, J. & Dale, R. (2002a) Learning selectional preferences for use in resolving associative anaphora. *Proceedings of the 2002 Australasian Natural Language Processing Workshop*. Canberra, Australia.

Meyer, J. & Dale, R. (2002b) Mining a corpus to support associative anaphora resolution. *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*. Lisbon, Portugal.

Miller, G. A. & Fellbaum, C. (1992) Semantic networks of English. In: B. Levin and S. Pinker (eds.), *Lexical and Conceptual Semantics*, pp. 197-229. Blackwell, Cambridge and Oxford, England.

Mitkov, R. (2002) *Anaphora Resolution*. Longman, UK.

Mitkov, R. (1998) Robust pronoun resolution with limited knowledge. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869-875. Montreal, Canada.

Mitkov, R. (1997) Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. *Proceedings of the ACL97/EACL97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pp. 14-21. Madrid, Spain.

Müller, C. & Strube, M. (2001) *MMAX*: a tool for the annotation of multi-modal corpora. In *the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 45-50. Washington, USA.

Muñoz, R. (2001) *Tratamiento y resolución de las descripciones definidas y su aplicación en sistemas de extracción de información*. PhD. Thesis. University of Alicante.

Orasan, C. & Evans, R. (2000) Experiments in optimizing the task of anaphora resolution. *Proceedings of ICEIS 2000*, pp. 191-195. Stanford, UK.

Palomar, M., Moreno, L., Peral, J., Muñoz, R., Fernández, A., Martínez-Barco, P., and Saiz-Noeda, M. (2001) An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*. 27: (4) (Dec. 2001), 545-567. Cambridge, MA, USA.

Paraboni, I. (1997) *Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa*. Dissertação de Mestrado. PUC, RS.

Paumier, S. (2006) *Unitex 1.2: user manual*. Université Marne-la-Valée. Disponível em: <http://www-igm.univ-mlv.fr/~unitex>. Acesso em 20 de dez. de 2006.

Poesio, M.; Alexandrov-Kabadjov, M.; Vieira, R.; Goulart, R.; Uryupina. O. (2005) Do discourse-new detectors help definite description resolution? *Proceedings of IWCS*. Tilburg, The Netherlands.

Reinhart, T. (1983) *Anaphora and semantic interpretation*. London: Croom Helm.

Rino, L.H.M & Seno, E.R.M. (2006) A importância do tratamento co-referencial para a sumarização automática de textos.In: *Estudos Lingüísticos*, v. 35, p. 1179-1188. São Paulo-SP.

Rocha Lima, C.H. da. (1978) *Gramática normativa da língua portuguesa*. 19ª edição. Rio de Janeiro: Livraria José Olympio Editora.

Rossi, D.; Pinheiro, C.; Feier, N.B.; Vieira, R. (2001) Resolução automática de co-referência em textos da língua portuguesa. *Revista Eletrônica de Iniciação Científica da SBC REIC*, ano I, vol. 1, n.2.

Russell, B. (1905) On denoting. *Mind. Reprinted in 1985, Logic and Knowledge (eds. R. C. Marsh)*, vol. 14, pp. 479-493. London: George Allen and Unwin.

Santos, D. N. A. & Carvalho, A. M. B. R. (2007) *Hobbs' Algorithm for Pronoun Resolution in Portuguese*. Trabalho em andamento na Unicamp (disponibilizado pelos autores). Campinas, SP.

Sidner, C. L. (1983). Focusing in the Comprehension of Definite Anaphora. In: *Brady, M. & Berwick, R. C. (eds.) Computational Models of Discourse*. MITPress, London, England.

Tapanainen, P. & Järvinen, T. (1997) A non-projective dependency parser. *Proceedings of the 5th Conference of Applied Natural Language Processing* (ANLP-5), pp. 64-71. Washington, DC, USA.

Ventura, C.S.M. &.Lima-Lopes, R.E. (2002) O Tema: caracterização e realização em português. In: *DIRECT Papers*, v. 47, p. 1-18. São Paulo – SP.

Vieira, R. (2001) Resolução automática de co-referência textual. *I Congresso e IV Colóquio da Associação Latino-americana de Estudos do Discurso ALED*, 23-28 de setembro. Recife, PE.

Vieira, R. (1998) *Definite description processing in unrestricted text*. PhD thesis. University of Edinburgh, Edinburgh.

Vieira, R.; Gorziza, F.; Rossi, D.; Chishman, R.; Rossoni, R.; Pinheiro, C. (2000) Extração de sintagmas nominais para o processamento de co-referência. *Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR*, 19-22 Novembro. Atibaia, SP.

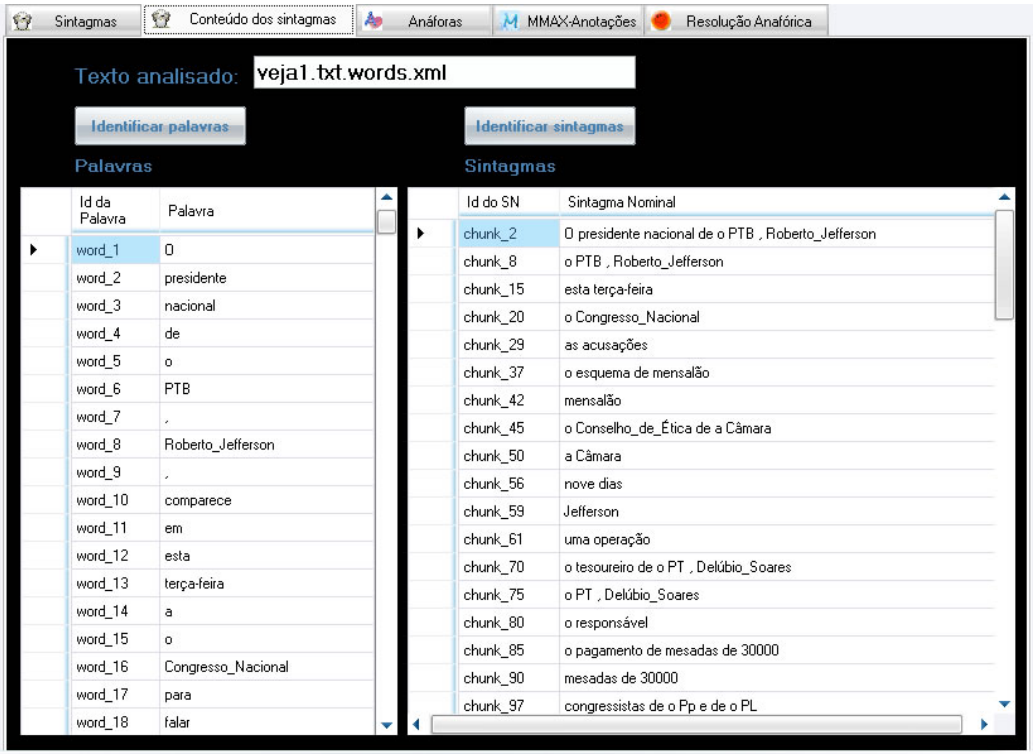
Vieira, R. & Lima, V.L.S. de. (2001) Lingüística computacional: princípios e aplicações. In: Luciana Nedel (eds.), *IX Escola de Informática da SBC-Sul*, pp. 27-58. Passo Fundo, RS.

Vieira, R. & Poesio, M. (2000) An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4): 525-579.

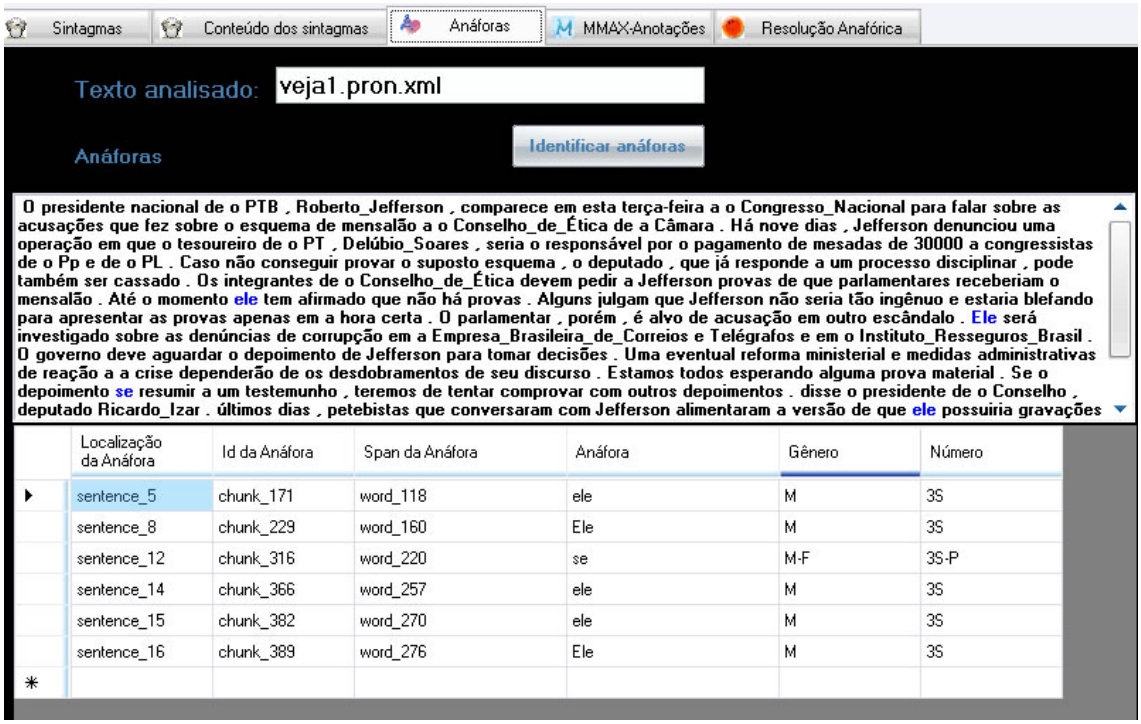
# Apêndice - Interfaces do ambiente desenvolvido para resolução anafórica

## 1- Interface inicial

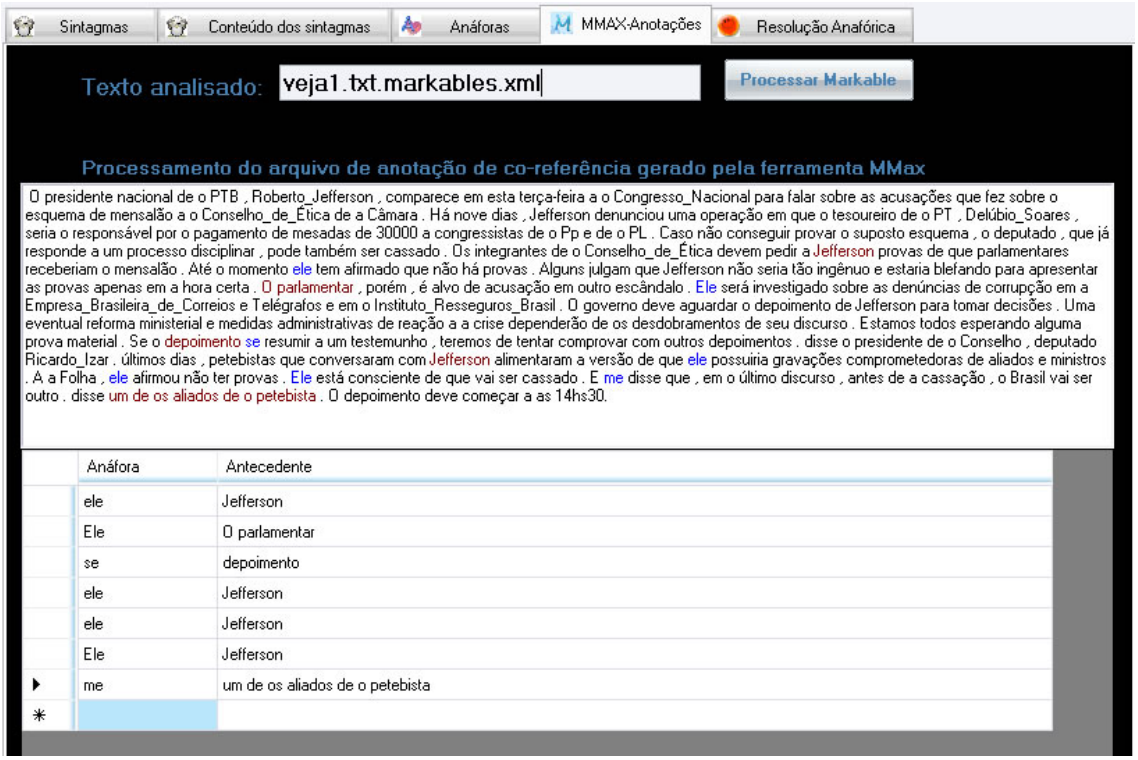

## 2- Modulo1: Interface para visualização dos arquivos de palavras e SNs.



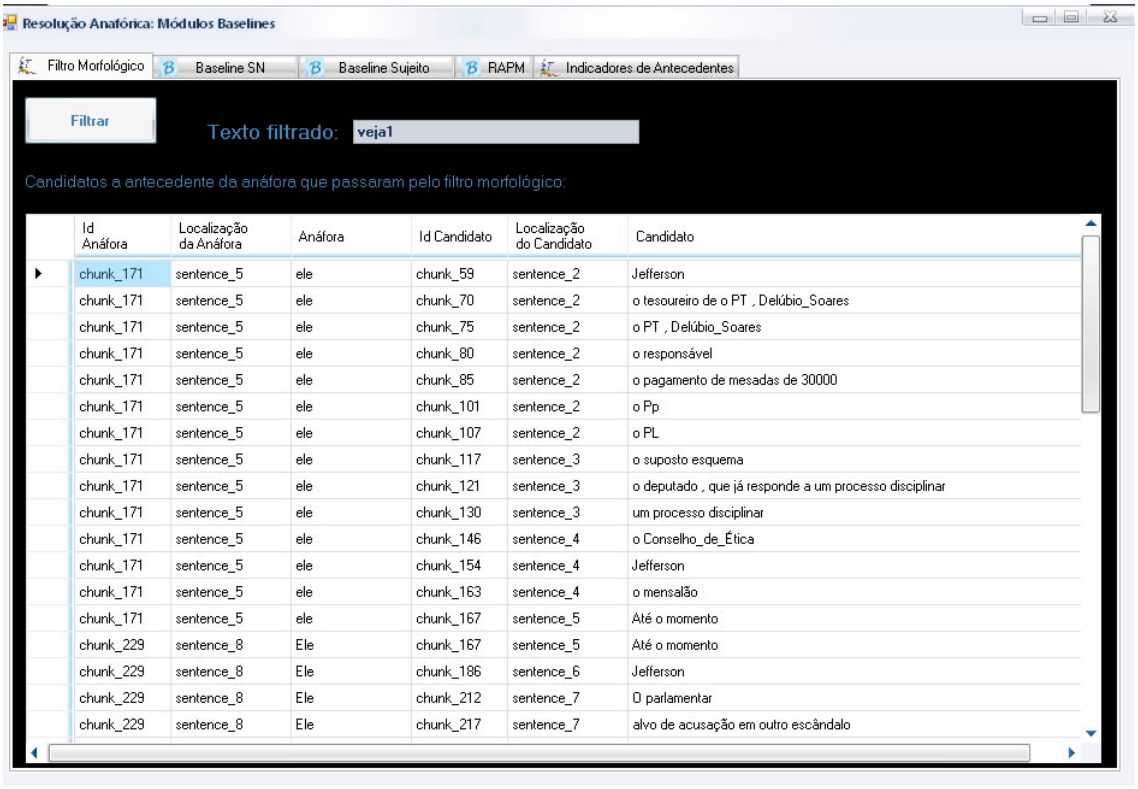
**3- Modulo1: Interface para visualização do arquivo de pronomes e de seu contexto.**



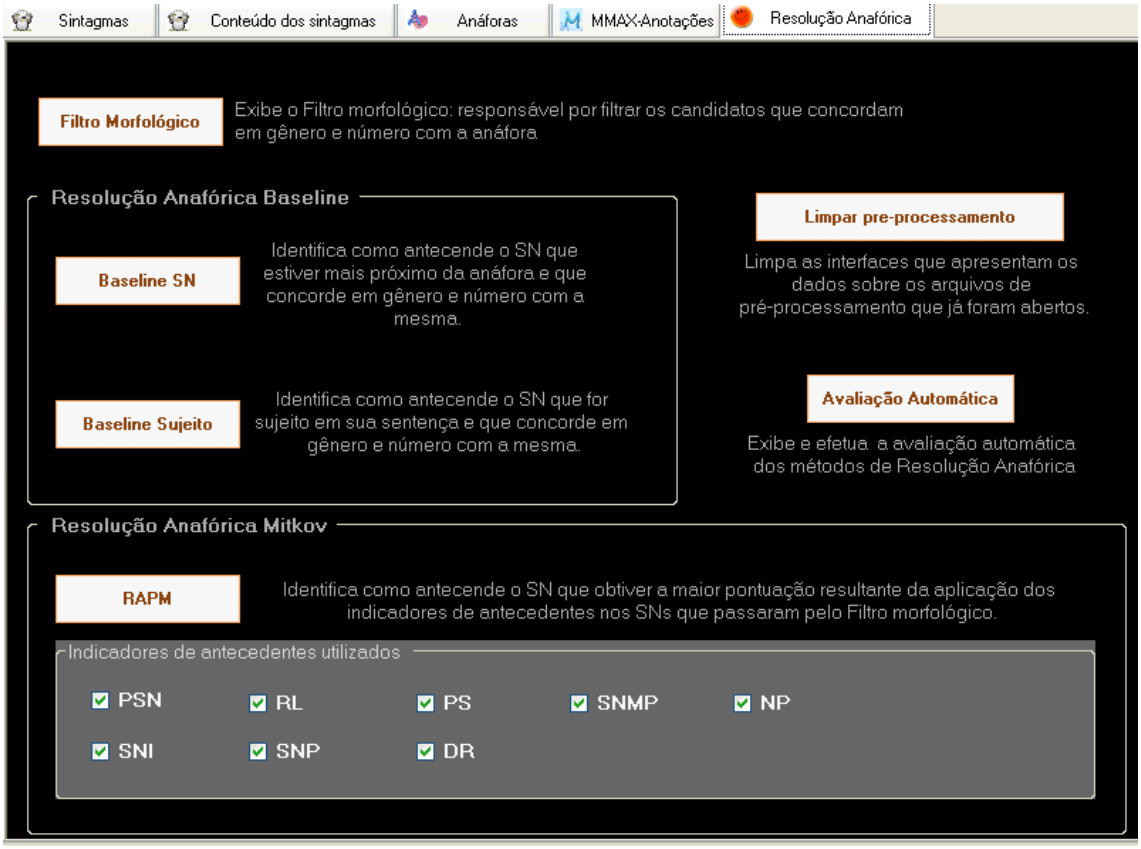

**4- Modulo1: Interface para visualização do arquivo com anotação de co-referência e de seu contexto.**



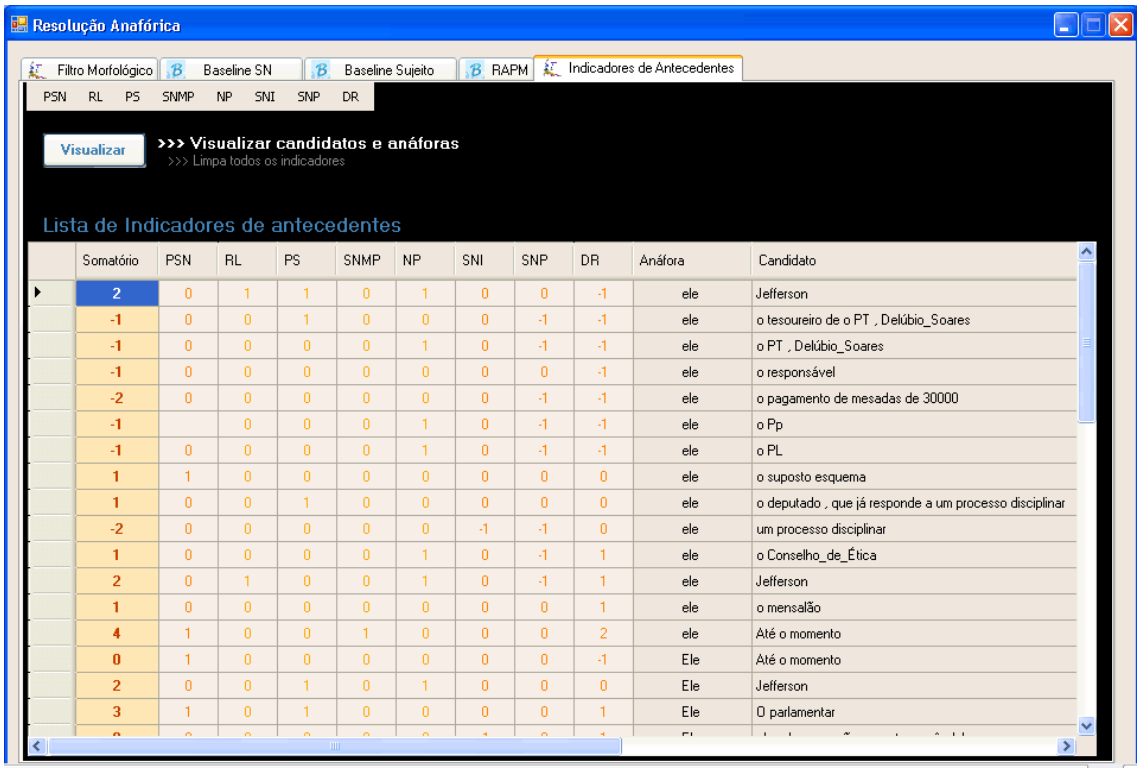
## 5- Modulo2: Filtro morfológico




## 6- Modulo3: Escolha das estratégias de resolução anafórica e de avaliação

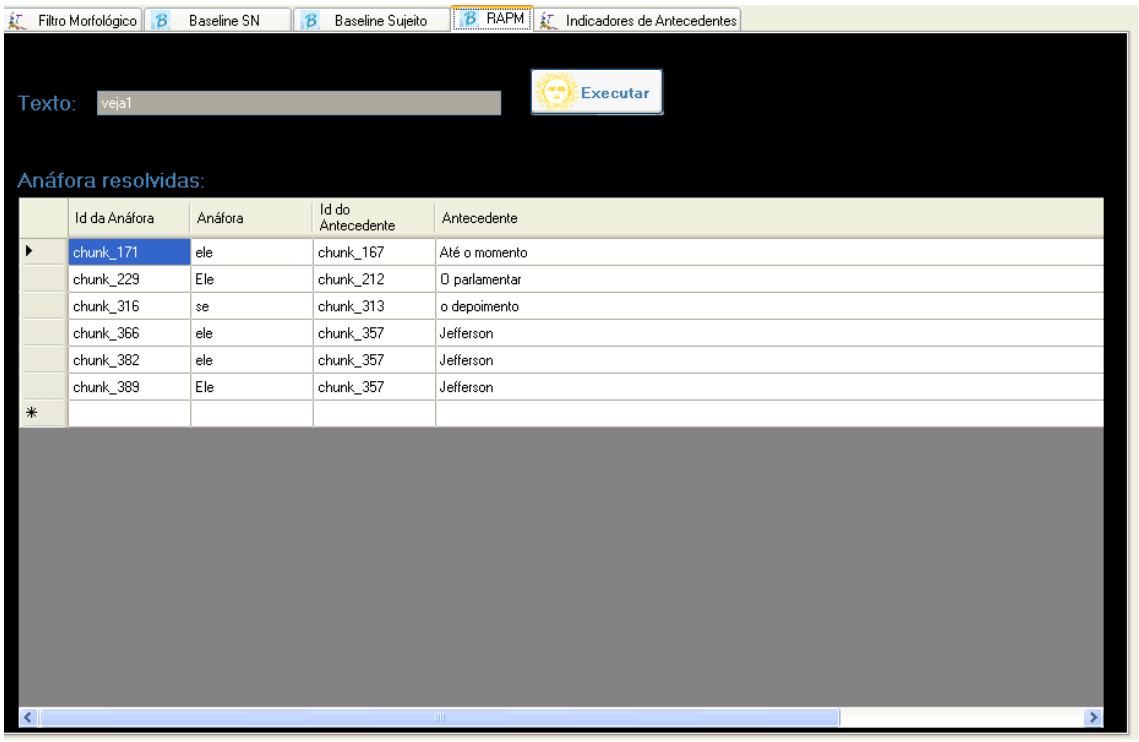


## 7- Modulo3: Aplicação dos indicadores de antecedentes




| Somatório | PSN | RL | PS | SNMP | NP | SNI | SNP | DR | Anáfora | Candidato |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | -1 | ele | Jefferson |
| -1 | 0 | 0 | 1 | 0 | 0 | 0 | -1 | -1 | ele | o tesoureiro de o PT , Delúbio_Soares |
| -1 | 0 | 0 | 0 | 0 | 1 | 0 | -1 | -1 | ele | o PT , Delúbio_Soares |
| -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | ele | o responsável |
| -2 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | ele | o pagamento de mesadas de 30000 |
| -1 | | 0 | 0 | 0 | 1 | 0 | -1 | -1 | ele | o Pp |
| -1 | 0 | 0 | 0 | 0 | 1 | 0 | -1 | -1 | ele | o PL |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ele | o suposto esquema |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ele | o deputado , que já responde a um processo disciplinar |
| -2 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | 0 | ele | um processo disciplinar |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | -1 | 1 | ele | o Conselho_de_Ética |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | -1 | 1 | ele | Jefferson |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ele | o mensalão |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | ele | Até o momento |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | Ele | Até o momento |
| 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | Ele | Jefferson |
| 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | Ele | O parlamentar |

## 8- Modulo3: Resolução anafórica




| Id da Anáfora | Anáfora | Id do Antecedente | Antecedente |
|---|---|---|---|
| chunk_171 | ele | chunk_167 | Até o momento |
| chunk_229 | Ele | chunk_212 | O parlamentar |
| chunk_316 | se | chunk_313 | o depoimento |
| chunk_366 | ele | chunk_357 | Jefferson |
| chunk_382 | ele | chunk_357 | Jefferson |
| chunk_389 | Ele | chunk_357 | Jefferson |

## 9- Modulo4: Avaliação automática da resolução anafórica