Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente

Amanda Rocha Chaves
Lucia Helena Machado Rino

NILC-TR-09-07

Agosto, 2007

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Apresentamos neste relatório um estudo sobre a aplicabilidade dos indicadores de antecedentes propostos por Mitkov (2002), originalmente aplicados para o inglês, para a língua portuguesa. Esses indicadores foram utilizados no processo de resolução de anáforas pronominais cujos antecedentes são sintagmas nominais. Este estudo baseou-se em um corpus pré-processado com informações morfológicas, sintáticas e co-referenciais que foram utilizados como entrada para um sistema de resolução anafórica, no qual alguns desses indicadores de antecedentes foram implementados. Esse sistema foi implementado em um ambiente que inclui quatro módulos: o primeiro para análise do corpus, o segundo para aplicação de um filtro morfológico que seleciona os sintagmas nominais candidatos a antecedentes da anáfora. Os sintagmas nominais devem concordar em gênero e número com a anáfora para que possam fazer parte do conjunto de candidatos no qual, possivelmente, estão presentes os antecedentes das anáforas. O terceiro módulo do ambiente construído é responsável pela resolução anafórica e o quarto pela avaliação dessa resolução.

Índice

Índice de Figuras

Índice de Tabelas

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 os constituintes do texto, 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, 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.

Trabalhos como o de Mitkov (2002) propõem a resolução automática de anáforas pronominais em três passos: 1) identificar a anáfora, 2) identificar o conjunto de possíveis antecedentes e 3) identificar e selecionar o antecedente da anáfora.

Mitkov utiliza como estratégia de resolução anafórica (RA) um conjunto de heurísticas que ele denomina 'indicadores de antecedentes' e avalia essa estratégia sobre um corpus de manuais técnicos. Esses indicadores são descritos e detalhados neste relatório, com o intuito de verificar se os mesmos podem ser aplicados para a língua portuguesa da mesma maneira que foram aplicados para a língua inglesa, se devem ser modificados ou mesmo se devemos criar novos indicadores.

A estratégia de RA desenvolvida neste trabalho consistiu na construção de um ambiente que desse suporte à análise da aplicabilidade de tais indicadores. Esse ambiente foi construído em linguagem C# e utiliza como entrada arquivos de um corpus selecionado. Esse corpus está constituído de 14 textos do gênero jornalístico anotados com informações morfossintáticas e co-referenciais, além de conter arquivos com notação XML representando os pronomes e sintagmas nominais presentes nesses textos.

Foram realizados três experimentos com o intuito de escolher os indicadores de

antecedentes a serem aplicados para o português e validar as estratégias de RA do ambiente construído.

Nas próximas seções apresentamos o contexto dos experimentos realizados e os resultados dos mesmos. Na seção 2 apresentamos o algoritmo de Mitkov e os indicadores de antecedentes utilizados por ele em seu processo de resolução anafórica. Na seção 3, o corpus utilizado, a ferramenta de RA desenvolvida e os experimentos realizados são detalhados e na seção 4, as conclusões desses experimentos são relatadas.

2 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 (SNs). 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'.

Sobre um texto pré-processado por um *parser* e por um extrator de SNs, o algoritmo proposto por Mitkov 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 são pontuados por cada indicador 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 mesma soma resultante, escolhe-se como antecedente o que estiver mais próximo da anáfora. A Figura 1 ilustra esse processo de RA.

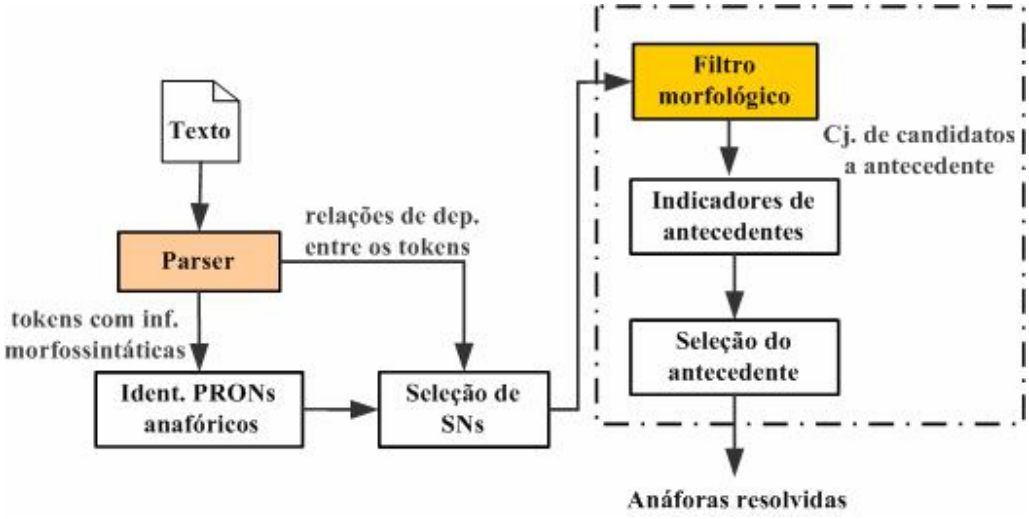


Figura 1: Arquitetura de RA com base no algoritmo de Mitkov

2.1 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 o quão semelhante a um pronome pode ser um possível antecedente e os negativos refletem a pouca probabilidade de um SN ser o antecedente do pronome anafórico.

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 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)[1].

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

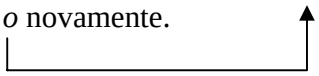
Preferência por SNs em título de seção (PSNTS): um *score* '+1' é atribuído aos SNs que ocorrem no título da seção na qual o pronome anafórico aparece. Esse *score* serve

[1] 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>, ou se o verbo for ‘ser/estar’, o seguinte padrão também é aceito: <SN/pronome, verbo, adjetivo/particípio>. Vejamos um exemplo:

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



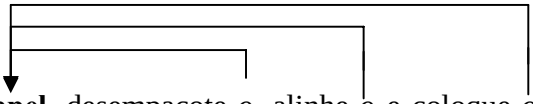
No exemplo (2.1), o padrão de ocorrência dos termos ‘o botão’ e ‘o’ é igual nas duas sentenças: <verbo, SN >, <verbo, PRON>, o que permite fazermos uma redução para o padrão <verbo, SN/PRON >. Para este caso, podemos considerar a premissa de que, se um SN possui o mesmo padrão de ocorrência do pronome, 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 (2.2) ilustra esse caso:



(2.2) 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:

(2.3) 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 (2.4):

(2.4) **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). A pontuação negativa atribuída a esse indicador pode ser explicada em termos de saliência, com base na teoria *centering* (Sidner, 1983). 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 (2.5).

(2.5) 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 este é pontuado negativamente pelo indicador SNP por fazer parte de um sintagma preposicionado, 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, neste trabalho o mesmo 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 1, para o segmento de texto jornalístico (2.6).

(2.6) 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

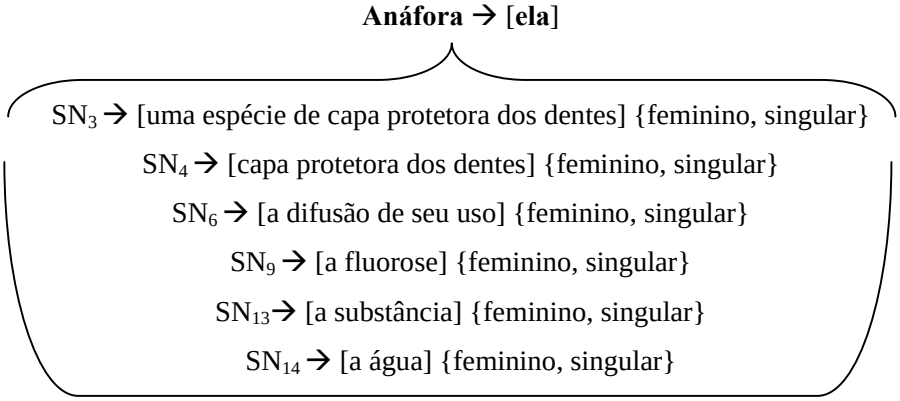
no organismo. Afinal, **a substância** não se encontra apenas na água e cremes dentais: *ela* também está presente em diversos alimentos, (...).

Para encontrar o antecedente do pronome ‘ela’ o sistema recebe como entrada o texto (2.6) já etiquetado com informações morfológicas e sintáticas, além de receber um arquivo contendo todos os seus sintagmas. Todos os SNs presentes no texto são extraídos na ordem em que os mesmos aparecem, conforme ilustra a Figura 2. 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. 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 2: SNs do texto 2.6

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:



A última etapa de RA, representada na Figura 1, 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 1 são apresentados os pesos associados aos 6 SNs anteriores, organizados de forma descendente por seus pesos.

Tabela 1: 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. Provavelmente, esses indicadores poderiam ser descartados para a resolução anafórica desse gênero textual.

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.

A próxima seção esclarece o objetivo desse estudo e descreve em detalhes quais indicadores de antecedentes foram escolhidos para serem implementados, levando em consideração a língua portuguesa e o corpus utilizado.

3 Proposta de trabalho

O objetivo geral desse 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 foram utilizadas. Outras foram desenvolvidas, especialmente, o ambiente que inclui nosso sistema de RA. 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 descritos o corpus utilizado e sua análise, o ambiente desenvolvido que inclui o processo de RA, bem como os três experimentos, seus resultados e as contribuições de cada experimento para a resolução de anáforas pronominais do português.

3.1 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 de co-referência 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 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 2 apresenta todos os arquivos contidos em ambos os pacotes com suas respectivas extensões e conteúdos.

Tabela 2: 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 (anáforas e antecedentes).
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 Pronomes	.pron	Pronomes anafóricos do texto.
Gerados pelo Extrator de Sintagmas Nominais	.np	Sintagmas Nominais do texto.

Na Figura 3 é mostrado um exemplo do conteúdo de um desses arquivos listados na Tabela 2, o arquivo '.pos'.

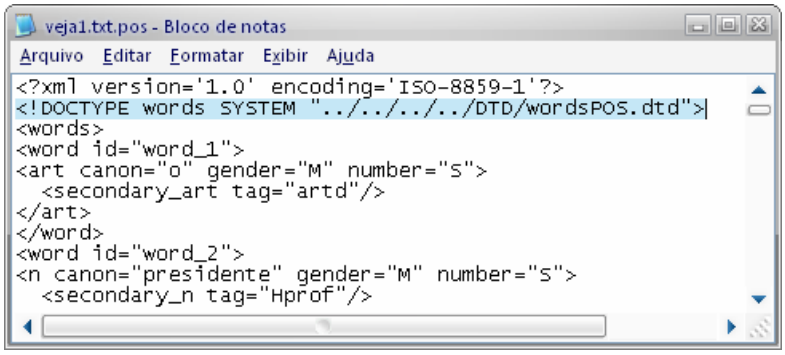


Figura 3: Exemplo de um arquivo .pos

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 3, 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 4, sem essa linha de código. Ademais, esses arquivos devem ser mantidos dentro de um mesmo diretório de trabalho.

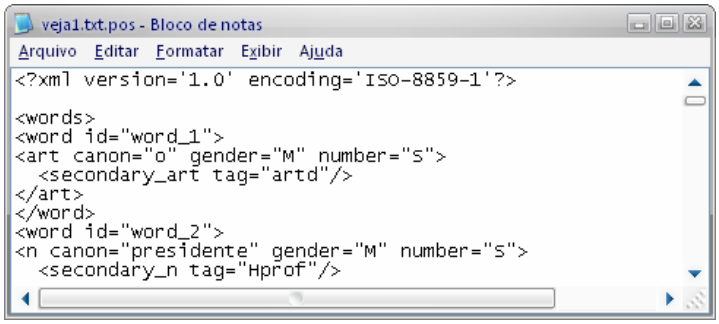


Figura 4: Arquivo .pos modificado

3.2 Análise do corpus

A análise do corpus jornalístico consistiu em averiguar se os indicadores de antecedentes de Mitkov (vide Seção 2.1) se aplicavam aos textos em português. Como resultado, foram descartados seis indicadores e selecionados cinco (Tabela 3). A escolha desses indicadores se deu pela independência de gênero textual e pertinência dos mesmos para a RA de textos em português, como será visto nas próximas seções.

Tabela 3: 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 os 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 está 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, através do qual são colhidos padrões de ocorrência de SNs e verbos. Isso o leva 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 descreve o ambiente desenvolvido para a RA, no qual são implementados os indicadores de antecedentes escolhidos nesta seção.

3.3 RAPM: características

A realização da análise do corpus e os experimentos E1, E2 e E3 foram subsidiados pelo uso de um ambiente que permite acompanhar a análise dos textos pré-processados, acionar um módulo de RA, o RAPM (**R**esolução **A**nafórica do **P**ortuguês baseada no algoritmo de **M**itkov), além de avaliá-lo automaticamente. Esse ambiente contempla uma interface gráfica amigável e 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 na Seção 3.1, 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. 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.

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

As estratégias *baseline* são as mesmas que também foram utilizadas por Mitkov (2002). 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 pela RAPM nos experimentos realizados são: PSN, RL, 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 no experimento E3, 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 manualmente.

A próxima seção relata o primeiro experimento executado com o intuito de avaliar os indicadores de antecedentes implementados nesse ambiente.

3.4 Experimento E1: índice de acerto e erro dos indicadores de antecedentes escolhidos

O experimento E1 teve 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 3 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 3: 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 5 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.

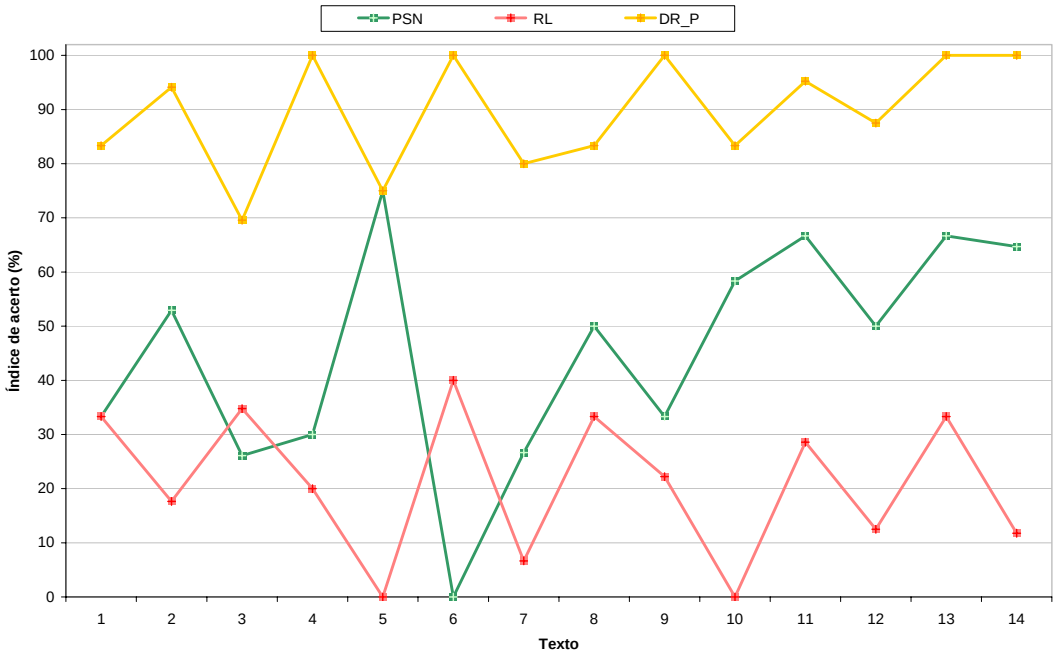


Figura 5: Índice de acerto dos indicadores promocionais

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 4. 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 4: 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 3), 156, o que equivale a uma média de 9,3 candidatos a antecedente por anáfora com antecedente.

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 3) quanto no erro (Tabela 4) 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 Figura 6 sintetiza bem a relação entre o número de candidatos pontuados incorretamente pelos indicadores de antecedentes promocionais.

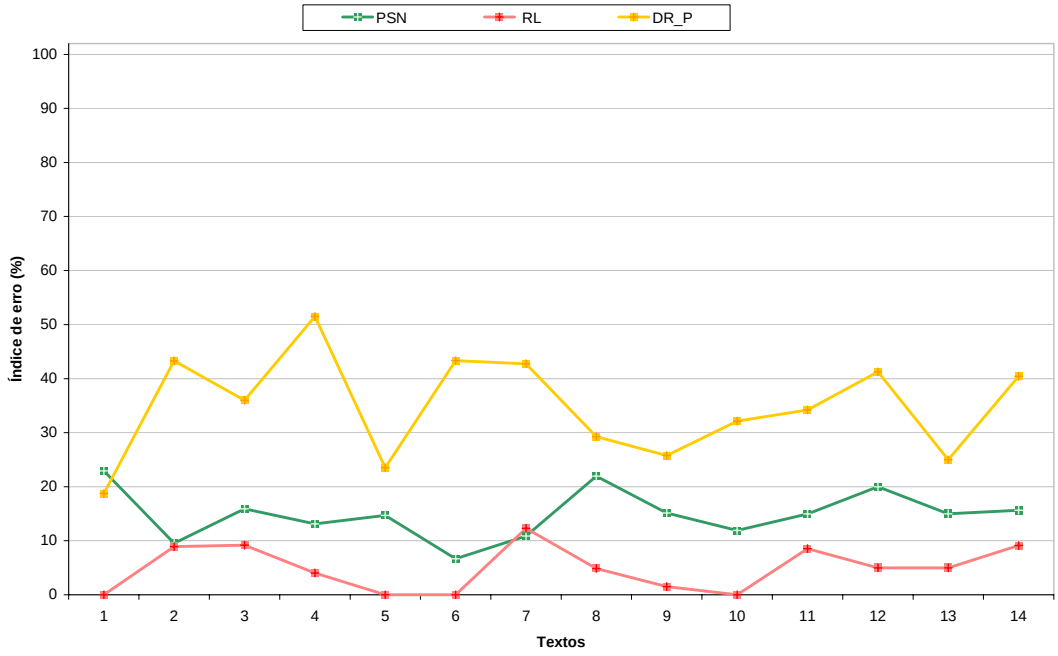


Figura 6: Índice de erro dos indicadores promocionais

Os índices de acerto e erro para os indicadores impeditivos são distintos dos já descritos para os indicadores promocionais. Na Tabela 5, 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 6, 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 7.

Acertos (Figura 7) e erros E (Figura 8) 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 9, 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 5: 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**

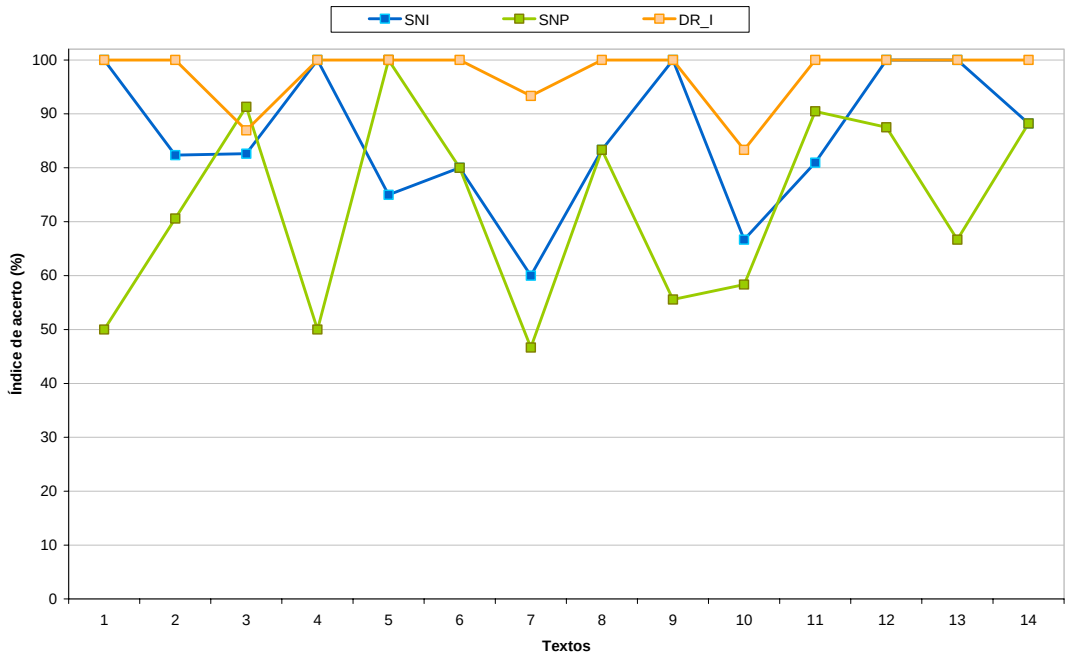


Figura 7: Índice de acerto dos indicadores impeditivos

Tabela 6: 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**

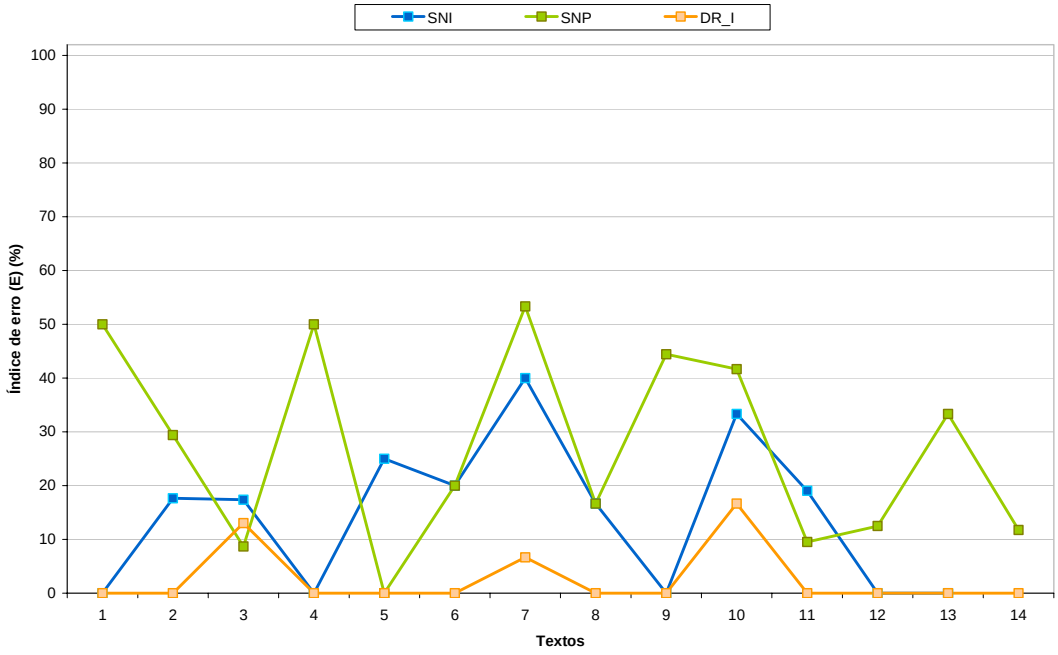


Figura 8: Índice de erro E dos indicadores impeditivos

Tabela 7: 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**

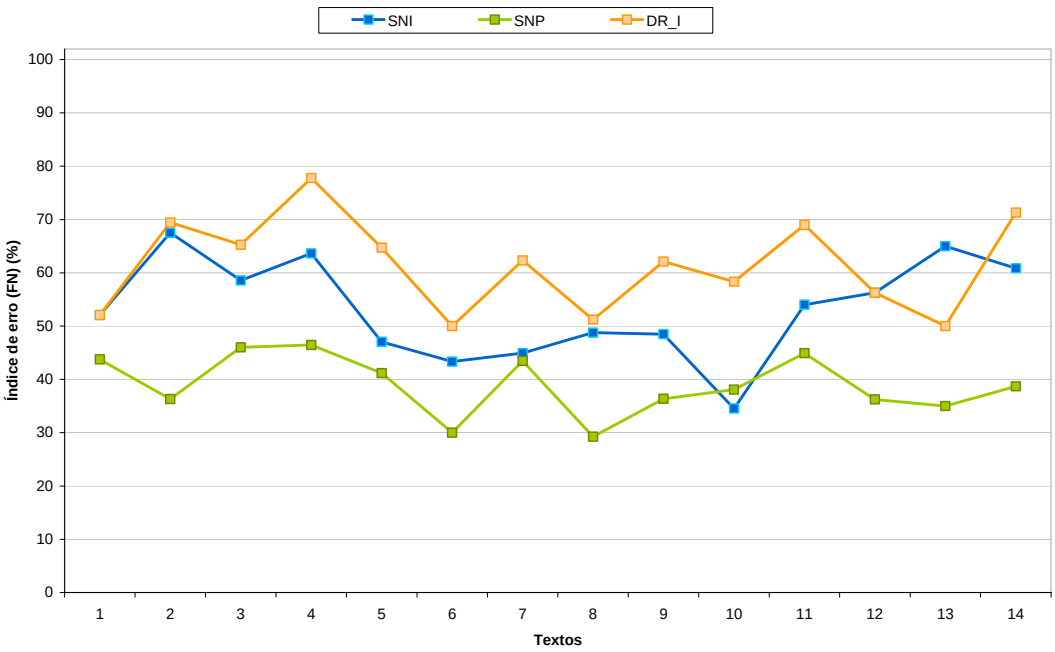


Figura 9: Índice de erro FN dos indicadores impeditivos

A média de acertos, exibida na última linha da Tabela 5 e ilustrada na Figura 7, 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 8) 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, um especialista fez a comparação manual dos resultados.

3.4 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[2].

Na Tabela 8 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 8: 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,

[2] 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 consideradas neste trabalho.

unicamente, como estratégia de RA. Através do gráfico da Figura 10 visualiza-se claramente essa medida para todos os indicadores avaliados para cada texto individualmente.

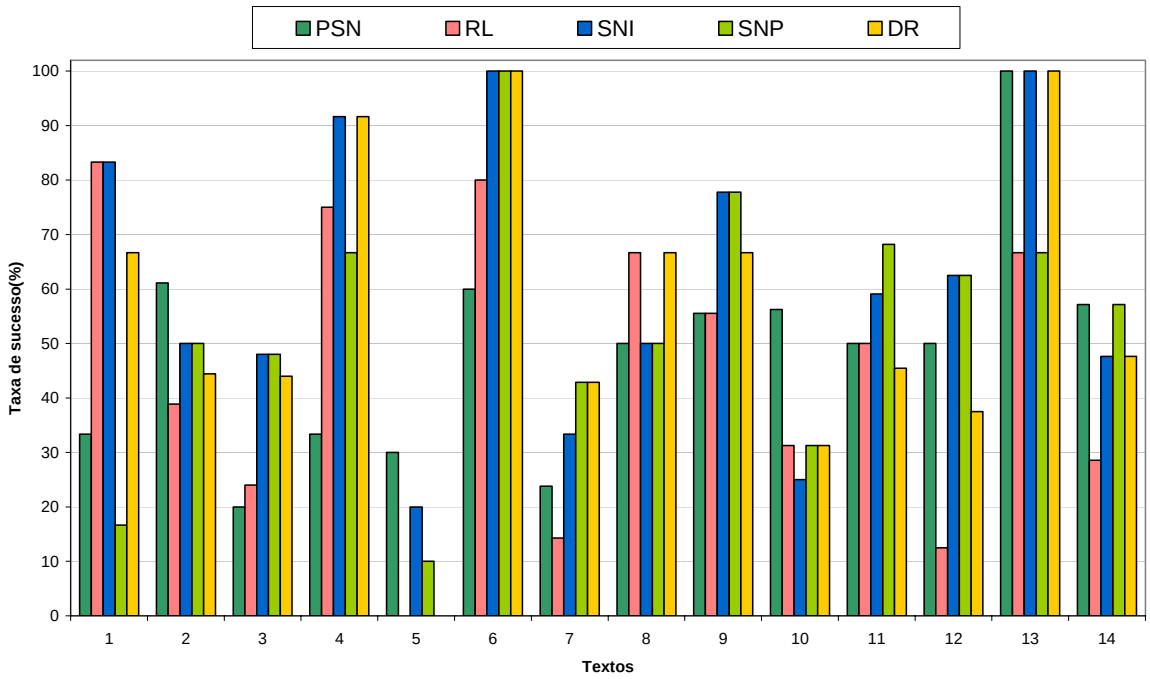


Figura 10: Taxa de sucesso de RA dos indicadores de antecedentes

A média geral de resolução anafórica, conforme ilustra a Tabela 8, 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 10, 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.

> (3.1) **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 (3.1), 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.

> (3.2) 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)[3] 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

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

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 (3.3):

> (3.3) 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 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 3.4, 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 3.1, citado anteriormente.

> (3.4) 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 (3.5), extraído do texto veja2, essa ferramenta certamente não conseguirá resolvê-la de forma correta.

(3.5) 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 (3.6):

(3.6) 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 (3.7).

(3.7) 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:

(3.8) 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 (3.9):

(3.9) "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 (89,7% de sucesso).

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.

3.5 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 da RAPM quando esta utiliza como cerne da resolução três estratégias distintas, como já dito anteriormente: 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 3.1) 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. 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 9, a síntese dos resultados desse experimento pode ser visualizada; o gráfico da Figura 11 sintetiza-os, evidenciando o melhor desempenho da estratégia *Baseline* Mitkov (média de 60,52%).

Tabela 9: Taxa de sucesso das estratégias *baseline*

		***Baseline* SN**		***Baseline* Sujeito**		***Baseline* Mitkov (cinco Indicadores)**	
Texto	**Anáforas válidas**	**AR[4]**	**TS[5] %**	**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**

[4] AR: Anáforas Resolvidas corretamente
[5] Taxa de Sucesso: Anáforas resolvidas corretamente / anáforas válidas

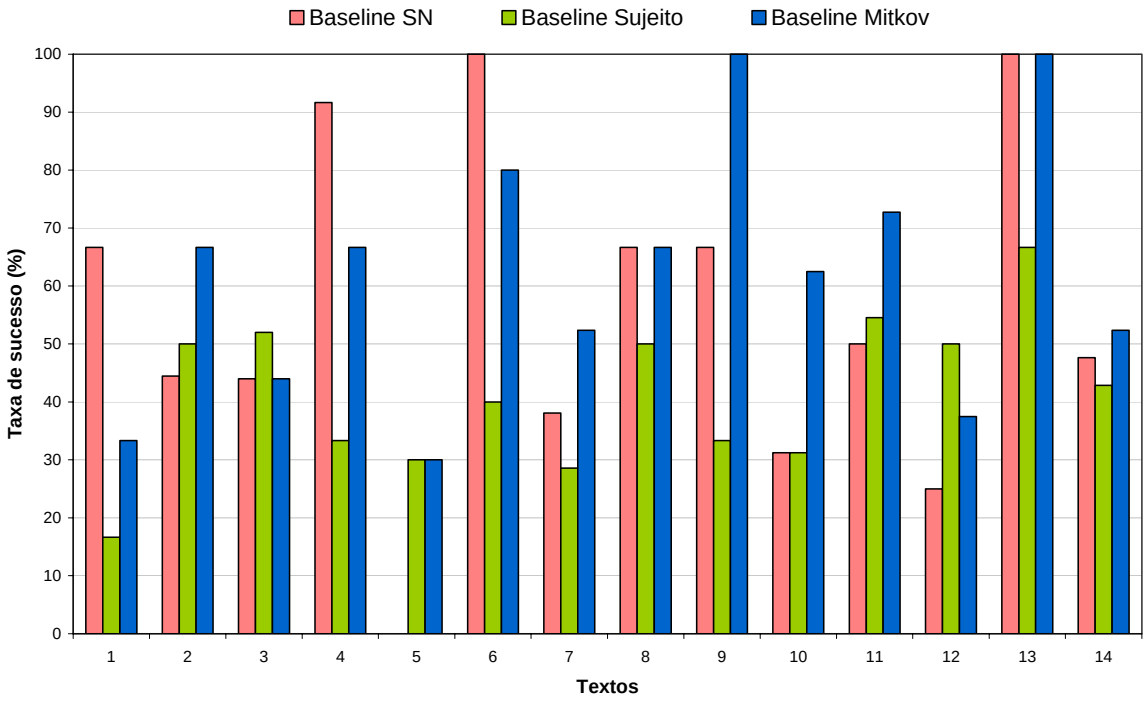


Figura 11: Taxa de sucesso das estratégias *baseline*

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 3.4, 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 9, 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.

4 Considerações finais

Esse relatório descreveu 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 3.2) e conjuntamente aplicados a um corpus jornalístico, é superior às outras duas estratégias *baseline* também avaliadas. Mitkov (2002) 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 seja nome próprio, busca no dicionário onomástico pelo gênero e número corretos deste nome e compara-os 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 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; os arquivos utilizados como entrada para a ferramenta RAPM contêm

anotações sintáticas, o que justificaria o uso do indicador paralelismo sintático, já que tais informações estão disponíveis; e finalmente, SNs que se encontram mais próximos do pronome anafórico tendem a ser o seu antecedente, o indicador SN mais Próximo promoveria tais candidatos. Esse último indicador também é utilizado como estratégia de RA, a *Baseline* SN.

Referências Bibliográficas

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

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.

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.

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

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.

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.

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.

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

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.

Bibliografia complementar

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

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.

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).

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.

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.

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.

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. (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.

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.

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.

Rezende, M. R. B.(2007) O papel dos artigos no discurso.In: *Revista Saberes*. Disponível em: <http://www.estacio.br/graduacao/letras/revista/amaria.htm>. Acesso em 02 de março de 2007.

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.

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.

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.