Uma adaptação do algoritmo de Lappin e Leass para resolução de anáforas em português.*

Thiago Thomes Coelho[1] **, Ariadne Maria Brito Rizzoni Carvalho**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas
Caixa Postal 6176 – 13084-971 Campinas, SP

{thiago.coelho, ariadne}@ic.unicamp.br

Abstract. *This paper presents the Lappin and Leass anaphora resolution algorithm, adapted to Portuguese; the algorithm solves third person pronominal anaphora and reflexive/reciprocal pronouns, and relies on salience measures derived from the syntactic structure of the sentence, and from a simple discourse representation model. The algorithm, as well as its evaluation with legal corpus, are presented.*

Resumo. *Este artigo apresenta uma versão adaptada para o português do algoritmo de Lappin e Leass, que visa resolver anáforas pronominais em terceira pessoa e pronomes reflexivos/recíprocos. O algoritmo é baseado em um sistema de pesos, atribuídos de acordo com a estrutura sintática da sentença, e em uma representação simples do modelo do discurso. O algoritmo desenvolvido, assim como o resultado de sua avaliação sobre um corpus jurídico, são apresentados.*

1. Introdução

O fenômeno da correferência, ou anáfora, que ocorre na língua natural, consiste no mecanismo do discurso no qual é feita uma referência abreviada à alguma entidade (ou entidades) na expectativa de que o ouvinte do discurso seja capaz de determinar a identidade da entidade abreviada. A referência abreviada é denominada anáfora, e a entidade referenciada chama-se referente ou antecedente. O processo de determinação do referente da anáfora é chamado resolução [Hirst 1981].

A resolução de anáforas pode melhorar a qualidade do resultado de diversas aplicações de processamento de língua natural, como por exemplo a recuperação e extração de informações, geração automática de resumos e tradução automática, entre outras. A referência anafórica é um fenômeno habitualmente associado aos pronomes, especialmente pronomes pessoais. A dificuldade no tratamento de anáforas reside na identificação do elemento correferenciado nos casos em que múltiplos antecedentes são possíveis para uma certa referência. Diversos algoritmos foram propostos para fazer a identificação do antecedente anafórico de pronomes, como o algoritmo de Lappin e Leass [Lappin and Leass 1994], o algoritmo de Hobbs [Hobbs 1978] e o algoritmo de Centering [Grosz et al. 1995].

O algoritmo de Lappin e Leass, ou RAP (*Resolution of Anaphora Procedure*), como é mais conhecido, visa resolver anáforas inter e intra-sentenciais pronominais de terceira pessoa, na língua inglesa. Este artigo apresenta uma adaptação do algoritmo de

*Este trabalho foi parcialmente financiado pelo CNPq.

Lappin e Leass para a resolução de anáforas pronominais em português, assim como sua avaliação através de um corpus jurídico.

O restante do artigo está organizado da seguinte forma: na próxima seção é apresentado o algoritmo original de Lappin e Leass; na seção 3 o algoritmo adaptado para o português é descrito e um exemplo de seu funcionamento é apresentado; na seção 4 é mostrado o resultado da avaliação do algoritmo; finalmente, na seção 5 são apresentadas as conclusões e apontados os trabalhos futuros.

2. O algoritmo de Lappin e Leass

Os componentes essenciais do algoritmo de Lappin e Leass são [Lappin and Leass 1994]:

- Um filtro intra-sentencial sintático [Lappin and McCord 1990b, Lappin and McCord 1990a] para descartar dependências anafóricas de um pronome a um sintagma nominal, com base em fundamentos sintáticos;
- Um filtro morfológico para excluir dependências anafóricas de um pronome a um sintagma nominal devido a discordâncias de gênero, número ou pessoa;
- Um algoritmo de ligação [Lappin and McCord 1990a] para identificar os possíveis antecedentes de um pronome reflexivo na mesma sentença;
- Uma função para atribuir os fatores de saliência adequados a cada sintagma nominal como paralelismo sintático, papel gramatical, entre outros; e
- Uma função de decisão para selecionar o antecedente de um pronome de uma lista de candidatos.

O algoritmo de ligação identifica os candidatos intra-sentenciais à antecedente dos pronomes reflexivos/recíprocos; já o filtro sintático elimina os candidatos intra-sentenciais à antecedentes de pronomes de terceira pessoa não reflexivos/recíprocos, com os quais correferência não é permitida. Para os candidatos remanescentes são calculados os valores de saliência como descrito na seção 2.1. Assim, o candidato escolhido como antecedente do pronome será aquele que possuir maior valor de saliência e, em caso de empate, será escolhido o candidato mais próximo ao pronome. Tanto o filtro sintático como o algoritmo de ligação analisam a estrutura sintática da sentença onde ocorre o pronome para determinar se a correferência é permitida. O RAP utiliza a representação gramatical gerada pelo analisador sintático desenvolvido por McCord [McCord 1990, McCord 1993].

2.1. Fatores de saliência

O RAP utiliza um sistema de pesos baseado em características sintáticas; nenhuma informação semântica é utilizada no processo de resolução. O algoritmo possui basicamente dois tipos de operação: atualização do modelo de discurso e resolução do pronome. Quando um sintagma nominal, que introduz uma nova entidade no discurso, é encontrado, uma representação para ele é criada e seu valor de saliência é calculado. O valor de saliência é dado pela soma de todos os fatores de saliência que se aplicam ao sintagma nominal. Os valores iniciais dos fatores de saliência são apresentados na tabela 1.

Os fatores de saliência retratam a preferência na escolha do antecedente de acordo com seu papel gramatical segundo a seguinte hierarquia [Jurafsky and H. Martin 2000]:

sujeito > *construção existencial* > *objeto direto* > *objeto indireto* > *locução adverbial preposicionada demarcada*

Tabela 1. Valores dos fatores de saliência iniciais [Lappin and Leass 1994].

Fatores de Saliência	Valores
Sentença Atual	100,0
Sujeito	80,0
Construção Existencial	70,0
Objeto Direto	50,0
Objeto Indireto	40,0
Ênfase Não Adverbial	50,0
Sintagma Nominal Não Contido	80,0

As sentenças a seguir ilustram a atribuição dos fatores de saliência às expressões em negrito:

- *Sujeito:* (1) **Os fiscais** procederam à prova com atraso.
- *Construção Existencial:* (2) Havia **uma casa azul** ao lado da padaria.
- *Objeto Direto:* (3) Mariana transformava **a minha vida**.
- *Objeto Indireto:* (4) Duvidava **da riqueza da terra**.
- *Ênfase Não Adverbial:* (5) Não saímos por causa **da chuva**.
- *Sintagma Nominal Não Contido:* (6) **Pedro** comprou **um carro**.

A atribuição dos quatro primeiros fatores é bastante clara, mas a atribuição dos dois últimos fatores merece uma explicação adicional. O fator de saliência *Ênfase Não Adverbial* é atribuído à entidades do discurso que não estejam contidas numa locução adverbial preposicionada demarcada. Assim, na sentença (5), o sintagma nominal "chuva" recebe esse fator pois, apesar de estar contido numa locução adverbial preposicionada, ela não é demarcada. Por outro lado, na sentença (7), abaixo, o fator de saliência *Ênfase Não Adverbial* não é atribuído ao sintagma nominal "da padaria", pois ele está contido na locução adverbial preposicionada demarcada "Ao lado da padaria".

(7) Ao lado da padaria, Pedro foi roubado.

O fator de saliência *Sintagma Nominal Não Contido* é atribuído à entidades do discurso que não estejam contidas em outro sintagma nominal. Assim, na sentença (6), os sintagmas nominais "Pedro" e "um carro" recebem esse fator. Por outro lado, na sentença (8), abaixo, os sintagmas nominais "usuário do carro" e "carro" não recebem o fator de saliência *Sintagma Nominal Não Contido*, pois estão contidos nos sintagmas nominais "O manual de usuário do carro" e "usuário do carro", respectivamente. Entretanto, esse fator é atribuído ao sintagma nominal "O manual de usuário do carro" como um todo, pois ele não está contido em nenhum outro sintagma nominal.

(8) **O manual de usuário do carro** está no porta-luvas.

O fator de saliência *Sentença Atual* é atribuído a todos os sintagmas nominais presentes na sentença que está sendo processada. O valor de saliência atribuído a toda entidade do modelo do discurso é dividido pela metade no início do processamento de cada sentença. O mecanismo de degradação dos valores de saliência e o fator de saliência *Sentença Atual* visam priorizar candidatos à correferência mencionados recentemente, já que os candidatos que ocorrem na sentença atual, e nas sentenças mais próximas a ela, tenderão a possuir valores de saliência maiores. No processo de escolha do antecedente, são considerados mais dois fatores de saliência: o fator de saliência *Paralelismo Sintático* e *Catáfora*. O primeiro prioriza candidatos que apresentam paralelismo sintático em relação ao pronome que está sendo resolvido, como no texto (9), abaixo:

(9) **Pedro** foi à concessionária com Bruno. Ele comprou um carro.

O sintagma nominal "Pedro" receberá o fator de saliência *Paralelismo Sintático* na resolução do pronome "Ele", pois ambos exercem a função sintática de sujeito.

O segundo fator penaliza catáforas, como na sentença (10), abaixo:

(10) Ela estava preparando o almoço quando **Maria** chegou.

O sintagma nominal "Maria" será penalizado na resolução do pronome "Ela" pois ocorre após o pronome. A atribuição desses fatores de saliência adicionais depende do pronome que está sendo resolvido e é temporária, já que seu peso é adicionado ao valor de saliência da entidade do discurso somente durante a resolução do pronome. Os valores desses fatores de saliência são apresentados na tabela 2, abaixo. Os valores de todos os fatores de saliência foram obtidos empiricamente por Lappin e Leass visando otimizar o desempenho do algoritmo no corpus utilizado para avaliá-lo. Esse corpus era composto por manuais de computadores na língua inglesa.

É importante ressaltar que uma entidade do discurso poderá receber mais de um fator de saliência, atribuídos de acordo com a estrutura da sentença e com sua função sintática. Detalhes sobre essas atribuições serão ilustrados através de um exemplo, apresentado na seção 3.3.

Tabela 2. Fatores de saliência adicionais [Lappin and Leass 1994].

Fatores de Saliência	Valores
Paralelismo Sintático	35,0
Catáfora	-175,0

2.2. Classes de equivalência

As classes de equivalência agrupam os pronomes que possuem um mesmo referente. O fator de saliência de uma classe de equivalência é dado pela soma dos fatores de todos os seus membros.

3. O algoritmo de Lappin e Leass adaptado para o português

O algoritmo desenvolvido e avaliado neste trabalho é baseado no RAP. O algoritmo aqui proposto implementou todos os principais componentes do algoritmo original, com as seguintes diferenças:

- O filtro sintático e o algoritmo de ligação foram substituídos pelas restrições de correferência propostas por Reinhart [Reinhart 1983]. Uma análise dos exemplos apresentados em [Lappin and Leass 1994, Lappin and McCord 1990b, Lappin and McCord 1990a] mostrou que as restrições de Reinhart seriam eficientes para resolver os casos apresentados pelos autores do algoritmo original;
- O analisador sintático utilizado foi o PALAVRAS [Bick 2000];
- O PALAVRAS Xtractor [Gasperin et al. 2003] foi empregado para converter a saída do analisador sintático em XML (*Extensible Markup Language*[1]). Essa ferramenta foi utilizada visando facilitar a extração das informações lingüísticas do corpus analisado pelo PALAVRAS;
- Não foi implementado um tratamento para catáforas.

[1]Disponível em http://www.w3.org/XML/

O PALAVRAS consta de um analisador sintático para o português que efetua, mesmo em sentenças incompletas ou incorretas, a análise morfossintática. Devido a falta de padronização da análise gerada pelo PALAVRAS, uma segunda ferramenta foi empregada, o PALAVRAS Xtractor. Essa ferramenta converte a saída do analisador em três arquivos XML: o arquivo *words*, que contém uma lista das palavras do texto e seus respectivos identificadores; o arquivo *pos*, contendo informações morfossintáticas sobre as palavras do texto; e o arquivo *chunks*, que contém a estrutura do texto.

3.1. Restrições de Reinhart

Algumas propriedades estruturais da sentença impõem restrições à relação de correferência [Carvalho 1989]. As restrições de Reinhart, utilizadas neste trabalho, são fundamentadas em torno do conceito de c-comando (*constituent-command*), que é assim definido [Reinhart 1983]:

- *O nó A c-comanda o nó B se e somente se o primeiro nó com ramificação α que domina* [2] *o nó A também domina o nó B, ou α é imediatamente dominado por um nó β que possui a mesma categoria sintática do nó α.*

As restrições relevantes para a resolução de anáforas pronominais que foram implementadas são as seguintes [Reinhart 1983]:

1. A correferência é proibida caso o pronome c-comande o sintagma nominal;
2. A correferência é permitida caso o pronome seja reflexivo/recíproco, o sintagma nominal c-comande o pronome e esteja dentro da categoria de governo mínima (*Minimal Governing Category* - MGC[3]) do pronome.
3. A correferência é proibida caso o pronome não seja reflexivo/recíproco, o sintagma nominal c-comande o pronome e esteja dentro da categoria de governo mínima do pronome.

Considere as sentenças abaixo:

(11) Ele sentou perto de Pedro.
(12) Maria gosta de si.
(13) Maria gosta dela.

As arvores sintáticas das sentenças (11), (12) e (13) são apresentadas nas figuras 1(a), 1(b) e 1(c), respectivamente[4].

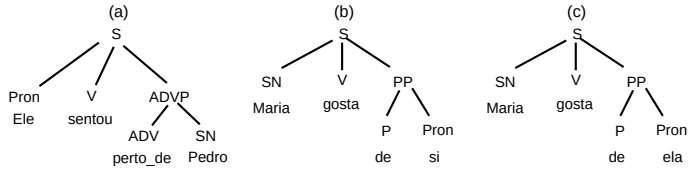

Figura 1. Arvores sintáticas das sentenças (11), (12) e (13).

[2]Um nó α é dominado por todos os seus ancestrais [Allen 1995].

[3]A categoria de governo mínima do nó α é definida como o nó raiz da sentença ou sintagma nominal ancestral do nó α que domina o sujeito da sentença.

[4]Legenda: ADVP: locução adverbial; ADV: advérbio; PP: locução preposicionada; P: preposição; PRON: pronome; SN: sintagma nominal; V: verbo.

Na sentença (11), de acordo com a restrição 1, a correferência entre o sintagma nominal “Pedro” e o pronome “Ele” é proibida pois o pronome c-comanda o sintagma nominal, ou seja, o primeiro nó que se ramifica e domina “Pedro”, o nó *S*, também domina “Ele”. Já na sentença (12), a restrição 2 permite que a correferência ocorra entre o sintagma nominal “Maria” e o pronome “si”, pois o sintagma nominal c-comanda o pronome, o pronome é reflexivo, e o sintagma nominal se encontra dentro da MGC do pronome. Já na sentença (13), a restrição 3 não permite a correferência entre o sintagma nominal “Maria” e o pronome “ela”, pois o sintagma nominal c-comanda o pronome, o pronome não é reflexivo, e o sintagma nominal se encontra dentro da MGC do pronome.

3.2. Processo de resolução

Os passos abaixo descrevem o processo de resolução de correferência de acordo com o algoritmo desenvolvido:

- No início do processamento de uma nova sentença, degradar o fator de saliência de todas as classes de equivalência, ou seja, dividir o fator de saliência de todas as classes por dois;
- Extrair todos os possíveis candidatos na sentença;
- Criar classes de equivalência para todos os candidatos que não pertencem a nenhuma das classes existentes. Nesses casos, os fatores de saliência apropriados serão atribuídos ao candidato segundo a tabela 1. O valor de saliência do candidato já incluído numa classe de equivalência será o fator de equivalência da classe a qual ele pertence;
- Para cada pronome da sentença:
  - Calcular o fator de saliência do pronome utilizando os valores da tabela 1;
  - Gerar lista com possíveis candidatos para pronomes de terceira pessoa:
    - Extrair candidatos utilizando uma janela de 4 sentenças que concordem em gênero e número com o pronome;
  - Gerar lista com possíveis candidatos para pronomes reflexivos/recíprocos:
    - Extrair somente candidatos intra-sentenciais que concordem em gênero e número com o pronome;
  - Aplicar as restrições de Reinhart aos candidatos intra-sentenciais remanescentes. Caso o candidato seja rejeitado pelas restrições, todas as entidades da sua classe de equivalência são excluídas da lista de candidatos;
  - Atribuir pesos adicionais aos candidatos de acordo com a tabela 2;
  - Selecionar o candidato com o maior fator de saliência. Caso ocorra empate, o candidato mais próximo ao pronome será escolhido;
  - Incluir o pronome na classe de equivalência do melhor candidato selecionado.

3.3. Exemplo de execução do algoritmo

A execução do algoritmo é ilustrada a seguir. Considere as seguintes sentenças:

(14) Pedro bebeu vinho de jabuticaba.
(15) Ele comprou-o no supermercado.

As árvores sintáticas das sentenças (14) e (15), geradas pelo PALAVRAS, são apresentadas na figura 2. No processamento da sentença (14), primeiro são coletados todos os

candidatos a referente, isto é, “Pedro” e “vinho de jabuticaba”; a seguir, seus valores de saliência são calculados e novas classes de equivalência são criadas para representá-los. Ao primeiro é atribuído o valor de saliência 310, pois “Pedro” é sujeito (80), é sintagma nominal não contido (80), não está contido numa locução adverbial preposicionada demarcada (50) e está presente na sentença que está sendo processada (100). Ao segundo sintagma nominal, isto é, “vinho de jabuticaba”, é atribuído o valor de saliência 280, pois ele é objeto direto (50), é sintagma nominal não contido (80), não está contido numa locução adverbial preposicionada demarcada (50) e está na sentença que está sendo processada (100). As classes de equivalência criadas para a sentença (14) são ilustradas na figura 3(a).

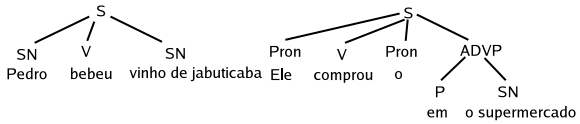

Figura 2. Arvores sintáticas simplificadas das sentenças (14) e (15)[4].

(a)

Referente	Anáforas	valor
Pedro		310
vinho de jabuticaba		280

(b)

Referente	Anáforas	valor
Pedro		155
vinho de jabuticaba		140
supermercado		230

(c)

Referente	Anáforas	valor
Pedro	Ele	465
vinho de jabuticaba		140
supermercado		230

(d)

Referente	Anáforas	valor
Pedro	Ele	465
vinho de jabuticaba	o	420
supermercado		230

Figura 3. Evolução do estado do modelo do discurso no processamento das sentenças (14) e (15).

Ao iniciar o processamento da sentença (15), todas as classes de equivalência tem seus valores de saliência divididos por dois. O possível candidato extraído nessa sentença é “o supermercado”, cujo valor de saliência é 230, pois é sintagma nominal não contido (80), não está contido numa locução adverbial preposicionada demarcada (50) e está na sentença que está sendo processada (100). A figura 3(b) mostra o estado atual do modelo do discurso. A seguir os pronomes da sentença são resolvidos. Na resolução do pronome “Ele”, não há candidatos intra-sentenciais já que a escolha do sintagma nominal “o supermercado” resultaria em catáfora. O possíveis candidatos inter-sentenciais são “Pedro” e “vinho de jabuticaba”, pois ambos concordam em gênero e número com o pronome que está sendo resolvido, e estão contidos na janela de 4 sentenças. Os valores da tabela 2 são então aplicados para calcular os valores de saliência final dos candidatos. O candidato “Pedro” tem adicionado 35 ao seu valor de saliência, pois ele exerce a mesma função sintática do pronome, isto é, ambos são sujeitos; já o candidato “vinho de jabuticaba” não tem nenhum acréscimo ao seu valor de saliência. Note que a atribuição dos fatores de saliência da tabela 2 é temporária; assim o acréscimo de 35 ao valor de saliência do candidato “Pedro” somente será considerado durante a resolução do pronome “Ele”. Assim, o candidato “Pedro” é escolhido como antecedente para o pronome “Ele”, pois possui o maior valor de saliência. O modelo do discurso é agora atualizado, e o fator de saliência do pronome é calculado. O pronome “Ele” tem o valor de saliência 310 pois é sujeito (80), é sintagma nominal não contido (80), não está contido numa locução adverbial preposicionada demarcada (50) e está na sentença que está sendo processada (100).

Posteriormente, o pronome é adicionado à classe de equivalência do candidato e o valor de saliência do pronome é adicionado ao valor de saliência da classe de equivalência do candidato escolhido. A figura 3(c) mostra o modelo do discurso após essa atualização. O próximo pronome a ser resolvido é o pronome "o". O possível candidato a antecedente intra-sentencial é "Ele". Note que o pronome "Ele" é considerado candidato à correferência, pois foi resolvido anteriormente e, portanto, é tratado como o sintagma nominal ao qual ele faz referência. Assim, a correferência com esse candidato é impedida pela restrição 3 de Reinhart. Os possíveis candidatos inter-sentenciais são "vinho de jabuticaba" e "Pedro", pois concordam em gênero e número com o pronome. O candidato "Pedro" não é considerado no próximo passo, pois pertence a classe de equivalência do pronome "Ele", que foi eliminado após a verificação das restrições de Reinhart. Assim, o sintagma nominal "vinho de jabuticaba" é o único candidato a referente e é, portanto, escolhido. Novamente, o modelo do discurso é atualizado e o valor de saliência do pronome é calculado. O pronome "o" recebe o valor de saliência 280, pois é objeto direto (50), é sintagma nominal não contido (80), não está contido numa locução adverbial preposicionada demarcada (50) e está na sentença que está sendo processada (100). O processo de atualização do valor de saliência da classe de equivalência do candidato escolhido é o mesmo descrito anteriormente. A figura 3(d) mostra o estado final do modelo do discurso após o término da execução do algoritmo.

4. Resultados

Os resultados apresentados são produto de testes sobre um corpus constituído por diversos pareceres da Procuradoria Geral da República de Portugal. O corpus foi anotado automaticamente pelo PALAVRAS com informações morfossintáticas. A anotação dos pronomes pessoais foi feita manualmente, utilizando uma ferramenta de anotação de discurso, o MMAX (*Multi-Modal Annotation in XML*) [Müller and Strube 2001]. A tabela 3 apresenta o resultado global da avaliação, e a tabela 4 mostra o resultado da avaliação individual de cada arquivo.

A anotação manual do corpus englobou quase todos os pronomes, com exceção dos pronomes reflexivos/recíprocos. Isso impossibilitou a avaliação do algoritmo desenvolvido com relação a esse tipo de pronome e também gerou uma disparidade entre o número total de pronomes anotados e o número de pronomes que o algoritmo tentou resolver, já que o algoritmo resolve somente pronomes pessoais de terceira pessoa.

Tabela 3. Análise global.

Total anáforas anotadas[5]	297
Anáforas mal resolvidas	190
Anáforas resolvidas corretamente	103
Anáforas mal identificadas[6]	4
Percentagem de sucesso	35,15 %

Os resultados obtidos foram bastante inferiores ao percentual de acerto de 86% obtido por Lappin e Leass, utilizando corpus de manuais de computadores na língua inglesa. A baixa taxa de acerto se deve, principalmente, à natureza complexa do corpus utilizado, do qual faz parte, por exemplo, a seguinte sentença:

[5]Número total de anáforas pronominais de terceira pessoa anotadas manualmente; os pronomes recíprocos/reflexivos não foram anotados.

[6]Número de pronomes anotados manualmente que não foram identificados pelo algoritmo.

Tabela 4. Avaliação em cada parecer jurídico.

	2.txt	3.txt	4.txt	7.txt	10.txt	11.txt	12.txt	13.txt	14.txt	15.txt
Total anáforas anotadas	17	21	23	13	5	41	21	32	10	2
Anáforas resolvidas corretamente	5	7	9	6	2	16	6	10	5	1
Anáforas mal identificadas	0	1	0	0	1	1	0	1	0	0
Percentagem de sucesso	29,41%	35%	39,13%	46,15%	50%	40%	28,57%	32,26%	50%	50%

	24.txt	25.txt	28.txt	29.txt	30.txt	36.txt
Total anáforas anotadas	36	4	21	22	13	16
Anáforas resolvidas corretamente	16	0	3	7	4	6
Anáforas mal identificadas	0	0	0	0	0	0
Percentagem de sucesso	44,44%	0%	14,29%	31,82%	30,77%	37,50%

(16) O casamento não irá afectar as legítimas expectativas dos filhos já existentes, visto que quer eles, quer os eventuais e futuros irmãos germanos, serão herdeiros de ambos os progenitores, quaisquer que sejam os bens, próprios ou comuns, destes, qualquer que seja o regime de bens convencionado.

Além disso, foram encontrados problemas na anotação do PALAVRAS e do PALAVRAS Xtractor, tais como árvores sintáticas geradas parcialmente, identificação incorreta de pronomes reflexivos/recíprocos e anomalias no XML gerado pelo PALAVRAS Xtractor (como, por exemplo, nós na estrutura sintática da sentença duplicados ou ausentes e extração incorreta de informações morfossintáticas). Outros fatores que podem ter influenciado no desempenho são o analisador sintático escolhido, de acordo com os estudos apresentados em [Preiss 2002], e o peso dos fatores de saliência aplicados, que foram os mesmos utilizados no algoritmo original, otimizados para o corpus em inglês.

5. Conclusão e trabalhos Futuros

Neste trabalho foi desenvolvida e avaliada uma versão modificada do algoritmo de Lappin e Leass para a língua portuguesa. Esse algoritmo é capaz de resolver anáforas inter-sentenciais e intra-sentenciais pronominais de terceira pessoa, e pronomes reflexivos/recíprocos. A avaliação preliminar desse algoritmo num corpus jurídico mostrou um desempenho inferior ao do algoritmo original para o inglês. É importante ressaltar que a avaliação do algoritmo original foi feita num corpus composto por manuais de computadores, ou seja, o corpus utilizado na avaliação era mais simples. Os resultados aqui apresentados nos parecem promissores, visto que foram obtidos sobre um corpus cujas sentenças são estruturalmente complexas e extremamente longas. A avaliação do desempenho em pronomes reflexivos/recíprocos não foi possível, pois esses pronomes não foram anotados no corpus utilizado. Como trabalho futuro, será feita outra avaliação em um corpus composto por textos literários e científicos, com pronome reflexivos/recíprocos anotados. Além disso, será feita uma comparação com o algoritmo de Centering, também adaptado para o português, e avaliado em [Aires et al. 2004], com o mesmo corpus jurídico utilizado neste trabalho.

6. Agradecimentos

Agradecemos a Ana Margarida Aires, Paulo Quaresma, Renata Vieira e Sandra Collovini, por terem prontamente sanado as inúmeras dúvidas que surgiram durante a execução deste trabalho, e aos revisores cujas sugestões muito contribuíram para a melhoria do artigo.

Referências

Aires, A. M., Coelho, J. C. B., Collovini, S., Quaresma, P., and Vieira, R. (2004). Avaliação de centering em resolução pronominal da língua portuguesa. In *Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués (Iberamia)*.

Allen, J. (1995). *Natural Language Understanding*. The Benjamin/Cummings Publishing Company.

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Frammework*. PhD thesis, Århus University.

Carvalho, A. M. B. R. (1989). *Logic Grammars and Pronominal Anaphora*. PhD thesis, University of Reading.

Gasperin, C. V., Vieira, R., Goulart, R. R. V., and Quaresma, P. (2003). Extracting xml chunks from Portuguese corpora. In *Proceedings of the Workshop on Traitement automatique des langues minoritaires*, Batz-sur-Mer.

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

Hirst, G. (1981). *Anaphora in natural language understanding: a survey*. Lecture notes in computer science : 119. Springer-Verlag.

Hobbs, J. (1978). Resolving pronoun references. *Lingua*, 44:311–338.

Jurafsky, D. and H. Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 1 edition.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronomial anaphora resolution. *Computational Linguistics*, 20(4):535–561.

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

Lappin, S. and McCord, M. (1990b). A syntatic filter on pronominal anaphora in slot grammar. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 135–142.

McCord, M. (1990). Slot grammar: A system for simpler construction of practical natural language grammars. In Studer, R., editor, *Natural Language and Logic: International Scientific Symposium*, pages 118–145.

McCord, M. (1993). Heuristics for broad-coverage natural language parsing. In *ARPA Human Language Technology Workshop*, University of Pennsylvania.

Müller, C. and 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*, pages 45–50, Washington, USA.

Preiss, J. (2002). Choosing a parser for anaphora resolution. In *the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002)*, pages 175–180.

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. Croom Helm Ltd.