UNIVERSIDADE CATÓLICA DE PELOTAS

ESCOLA DE EDUCAÇÃO

NÚCLEO DE PESQUISA LINGÜÍSTICA E LITERATURA

A resolução da anáfora no processamento da língua natural

Vilson J. Leffa (Coordenador)

Relatório Final de Pesquisa

Setembro de 2001

1. DADOS DE IDENTIFICAÇÃO

TÍTULO:	A resolução da anáfora no processamento da língua natural.
RESPONSÁVEL:	Prof. Dr. Vilson J. Leffa
ÁREAS:	Lingüística Aplicada (CNPq:8.01.00.00 - 7) Ciência da Computação (CNPq: 1.03.00.00 -7)
SUBÁREAS:	Lingüística Aplicada (CNPq: 8.01.06.00 - 5) Sistemas de Computação (CNPq: 1.03.04.00 -2)
INSTITUIÇÃO:	Universidade Católica de Pelotas
ESCOLA/NÚCLEO:	Escola de Educação — Curso de Mestrado em Letras Núcleo de Pesquisa em Lingüística e Literatura Rua Félix da Cunha, 412 96010-000 — Pelotas, RS Fone: (0532)848-282 — Fax: (0532)253105
FINANCIAMENTO:	CNPq (Bolsa PIBIC)
VIGÊNCIA:	De agosto de 1999 a julho de 2001
ENDEREÇO DO PESQUISADOR:	Caixa Postal 166 94400-970 - Viamão - RS Fone: (051)485-1380 Email: leffa@atlas.ucpel.tche.br

Introdução

Pode-se dizer que uma palavra tem duas partes : forma e conteúdo. Em termos muito simples, isso quer dizer que para cada forma lingüística há pelo menos um objeto correspondente no mundo real.

A forma "árvore", por exemplo, como uma seqüência de sons ou uma seqüência de letras,

pronunciada por alguém ou impressa numa página, corresponde ao conceito usual que nós temos de que árvore são feitas de tronco, galhos e folhas. A relação entre forma e conteúdo - significante e significado nos termos de Saussure - é muito próxima, como os dois lados de uma moeda. Significante e significado são unificados em uma unidade maior, geralmente definida como signo lingüístico, e não podem ser separados.

Obviamente, quando a linguagem é usada por pessoas em situações da vida real a dicotomia de Saussure, tão límpida na teoria, torna-se confusa na prática. Primeiro, o problema está na ambigüidade, onde uma forma lingüistica pode referir-se a vários objetos no mundo e vice versa. Segundo, há o complicado problema da anáfora, onde a forma lingüistica não se refere diretamente a um conceito, mas a outra forma lingüistica que finalmente se relacionará a um conceito.

A anáfora pode ser descrita como um processo que acarreta a volta no texto. O processo começa quando o anaforizante é conhecido (por exemplo, o pronome) e concluído quando o anaforizado é encontrado (a palavra a qual o pronome se refere). Descrever o que acontece entre esses dois momentos foi o propósito desta pesquisa. O objetivo é oferecer a descrição num nível explícito que possa ser usado para implementação em diferentes linguagens computacionais, incluindo Prolog, C ou Basic.

Rastreando o antecedente

O seguinte segmento ilustra muitos detalhes envolvidos na resolução da anáfora e serve para demonstrar como o segmento abaixo, escolhido por sua simplicidade e pela ausência de ambigüidade, é usado para ilustrar os conceitos básicos subjacentes a esse processo.

Segmento 01: *Houses* ^{i} are bought because *they* ^{i} offer comfort.

O pronome *they* não se relaciona diretamente com um objeto no mundo mas com palavras mencionada antes. A tarefa mental desempenhada pelo leitor ao processar a sentença, é voltar no texto e encontra a palavra a qual ele se refere. No Segmento 1, há quatro palavras (*because, bought, are* e *houses*) mas somente uma é um sério candidato (*houses*). O pronome *they* só pode ser substituído por

um substantivo plural e a única palavra que preenche este requisito é *houses*.

Exemplo da vida real nem sempre são tão simples. Um problema que pode surgir é a possibilidade de haver mais de um candidato legítimo para o antecedente, como está demonstrado no caso seguinte:

Segmento 02: *Houses* ^{i} are bought by people because *they* offer ^{i} comfort.

Agora não há um mas dois candidatos para *they*, que são as palavras *houses* e *people* (ambas substantivos plurais). Como resolver este problema? Uma hipótese é resolvê-lo aplicando as restrições sintáticas. Pode ser argumentado que há um paralelismo sintático entre o substantivo *houses* e o pronome *they*, isto é : *houses* e *they* estão na posição do sujeito em suas próprias orações. A palavra *people* por outro lado, embora substantivo plural não compartilha deste paralelismo com *they*. Assim, entre os dois candidatos, escolhemos o substantivo *houses.*

Restrições sintáticas baseadas no paralelismo, no entanto, parecem funcionar bem apenas quando os exemplos são cuidadosamente escolhidos.

No Segmento 2, por exemplo uma simples mudança num item lexical pode reverter totalmente a relação entre anaforizante e o referente. Isso pode ser observado no próximo segmento onde *offer* foi substituído por *like*.

Segmento 03: *Houses* ^{i} are bought by *people* ^{j} because *they* ^{j} like comfort.

Mais uma vez, existem dois candidatos exatamente como nos exemplos anteriores. Mas se aplicarmos a restrição sintática, como fizemos antes, escolhendo o sintagma nominal que está na posição do primeiro, chegaríamos a palavras *houses*, que obviamente é a escolha errada (**Houses like confort*.). O paralelismo sintático, que tão eficazmente facilitou a escolha entre os 2 candidatos no segmento anterior, parece não funcionar mais.

O único candidato que pode ocupar legitimamente a posição ocupada por *they* é *people*. O outro candidato (*houses*) viola a restrição semântica: *"houses não gostam de coisas, somente people gostam de coisas"*.

Assim, o paralelismo sintático é superado pelas restrições semânticas. Não basta o antecedente possuir a mesma função sintática do anaforizante. Tanto o anaforizado como o anaforizante devem compartilhar o mesmo traço semântico.

As restrições sintáticas e semânticas, desse modo, não bastam para resolver os problemas associados à resolução da anáfora, como pode ser também observado nos seguintes casos:

Segmento 04: The *campanies*^{i} sold their cars^{j} to the sheiks^{k} because *they*^{i} offered long-term guarantee.

Segmento 05: The *companies*^{i} sold their *cars*^{j} to the *sheiks*^{k} because *they*^{j} were bulletproof models.

Segmento 06: The companies^{i} sold their cars^{j} to the *sheiks*^{k} because *they*^{k} offered more money.

Os segmentos 4-6 aparentemente podem ser resolvidos apenas pela recorrência das representações de mundo em que compradores, vendedores e mercadorias trocam de mãos: dinheiro de compradores para vendedores e carros de vendedores para compradores. Precisamos saber também que carros podem ser a prova de bala, que as companhias oferecem garantias sobre o que elas vendem e que os xeques podem ser muito ricos.

Todo esse conhecimento de mundo precisa estar disponível para que o antecedente de *they* possa ser identificado corretamente em cada um dos segmentos.

O problema, no entanto, é o alto custo computacional que o uso do conhecimento de mundo acarreta. São tantas as variáveis que uma explosão combinatória se torna inevitável. Cada variável pode interagir com muitas outras variáveis, com muitas possibilidades diferentes de combinações e o sistema pode entrar num laço infinito – "endless loop" – e a combinação certa jamais poderá ser encontrada.

A solução para o problema do rastreamento do antecedente na anáfora parece estar entre a

simplicidade das restrições sintáticas e semânticas e a complexidade do conhecimento de mundo. Este foi o problema investigado neste projeto. Há duas questões para serem respondidas: (1) Quais são as limitações das restrições sintáticas e semânticas na resolução da anáfora? (2) Quais outras possíveis soluções podem ser encontradas entre as restrições e o conhecimento de mundo?

Discurso, cognição e Restrições Textuais

A anáfora pode ser estudada a partir de diferentes perspectivas; incluindo o discurso (e.g. McEnery and Botley, 1998; Indursky, 1997), cognição (e.g. Langacker, 1996; van Hoek, 1992) e restrições textuais (Dagani and Itai, 1990; Nasukawa, 1994; Mitkov and Belguith, 1998). Muitos desses estudos enfatizam a correlação entre alguns fatores discursivos/pragmáticos (ex. topicalidade) e uma determinada forma anafórica ("mecanismos de rastreamento" na terminologia de Du Bois, 1980). Fox (1996) resume essas correlações da seguinte maneira:

> (a) uso do pronome ou do pronome zero quando a anáfora está perto do tópico que está sendo desenvolvido, uso de sintagma completo quando a topicalidade é baixa;
> (b) uso do pronome ou do pronome zero quando a anáfora está na mesma seqüência discursiva do que foi mencionada antes, uso do sintagma completo quando isso não acontece;
> (c) uso do pronome ou pronome zero quando o falante pressupõe mais atenção do ouvinte, uso de sintagma completo quando o falante pressupõe um nível mais baixo de atenção;
> (d) uso do pronome ou do pronome zero quando o falante estiver envolvido emocionalmente;
> (e) uso dos sintagmas completos quando a atitude do falante for muito positiva ou negativa. (Fox, 1996, p. vii)

Os mecanismos de rastreamento, tanto pelo uso de pronomes, pronome zero ou de sintagmas nominais completos quando estiverem correlacionados com a topicalidade, seqüência discursiva e estado

emocional e cognitivo do falante não revelam muito em relação ao processo envolvido. Tudo se resume na probabilidade do uso do antecedente – numa escala que vai do pronome zero aos sintagmas nominais completos. Isso não é uma descrição do que realmente acontece na mente do leitor ou do ouvinte quando eles encontram o anaforizante e tentam rastrear o anaforizado, dentro ou fora do texto. A resolução da anáfora neste processamento de baixo nível, na maioria das vezes abaixo do controle consciente, provavelmente não seja uma área que interesse às pesquisas do paradigma discursivo/ pragmático, que talvez se concentre mais no quadro geral, vendo o processo em um nível mais abstrato. Uma perspectiva muito diferente, oferecida pelos estudos na lingüística aplicada computacional, é a implementação de um sistema de resolução da anáfora que traduza conceitos abstratos para um código legível pela máquina, usando dados que devem ser encontrados na superfície textual. Com o poder de processamento dos computadores modernos, essa variedade de dados, passíveis de análise, tem aumentado. Não estamos mais limitados a dados de baixo nível lingüístico, tais como informações sobre classes de palavras, mas podemos também incluir estruturas lingüísticas complexas de alto nível, relacionadas a possíveis configurações entre diferentes segmentos de texto. Podemos recursivamente encapsular segmentos da língua em unidades cada vez maiores, construindo grandes blocos, e abstraindo suas características. O ponto crucial, no entanto, é que a ligação entre o anaforizado e o anaforizante, não pode ser ambígua, levando a um total acordo entre diferentes leitores que consumindo o mesmo texto. Se surgir um desacordo, não devido às diferenças do texto, mas a diferentes interpretações dos leitores, o problema está além de uma solução pelas perspectivas da lingüistica computacional, que é basicamente algorítmica.

Tentativas para dotar os computadores com conhecimento de mundo necessário para atribuir sentido ao

texto, em vez de apenas extraí-lo, são teoricamente interessantes mas extremamente caras e, por enquanto, impraticáveis. A resolução da anáfora, em termos de lingüística computacional, não pode ser atribuída ao estado afetivo ou cognitivo do leitor; os dados devem estar presentes na superfície do texto.

Os dados lingüísticos que podem ser encontrados no texto, tais como concordância de gênero e número, restrições do c-comando, paralelismos semântico e sintático, repetições lexicais ou proximidade do antecedente são favorecidos no processo de resolução por que podem ser mais facilmente manipulados pelas ferramentas disponíveis na lingüística computacional. Essas ferramentas geralmente usam os conceitos de "restrição" e "preferência" - onde "restrição" é o mais poderoso dos dois instrumentos. Soluções baseadas em conhecimento de mundo restrito, usando metodologias baseadas em Corpus e modelos estatísticos/probabilísticos são preferidas.

Algumas abordagens estratégicas para rastear o antecedente, em oposição aos modelos estatísticos puros tem sido propostas. Essas abordagens podem ser formalizadas em termos de regras, geralmente baseadas nas restrições e preferências. As seguintes preferências, por exemplo, podem ser usadas na seleção do antecedente (baseado em Mitkov (1994,1996)):

- O SN é o objeto dos seguintes verbos: *discutir, presentear, ilustrar, resumir, examinar, descrever, definir, mostrar, checar, desenvolver, revisar, relatar, contornar, considerar, investigar, explorar, avaliar, analisar, sintetizar, estudar, negociar e cobrir;*
- O SN é modificado pelos seguintes adjetivos verbais: *definido, chamado, suposto;*
- O SN é modificado pelos seguintes advérbios: particularmente, especialmente, especificamente;
- O SN é o objeto dos seguintes substantivos: seção, tabela, figura, papel e relatório;
- O SN é repetido várias vezes no texto.

- O SN ocorre no cabeçalho da seção.

Paraboni (1997) adotou uma abordagem estratégica usando uma combinação de restrições e preferências nos seus estudos da anáfora na Língua Portuguesa sobre adjetivos possessivos. Esses adjetivos, quando pertencentes a terceira pessoa, são interessantes no português por não concordarem com o gênero e número do antecedente, como ocorre com o pronome *they* do inglês, mas sim com a coisa possuída, uma característica que torna ainda mais difícil localizar o antecedente. Por isso, os caminhos estratégicos para localização do antecedente possuem poucas restrições e preferências.

Na investigação de Paraboni pouquíssimas regras são oferecidas para rastrear o antecedente. Uma das mais produtivas é presença da conjunção coordenada entre anaforizado e anaforizante segmento 7.

Segmento 7: The law ^{i} and its ^{i} consequences

Paraboni, no entanto, é muito cauteloso ao apontar que exceções a essa regra podem ser facilmente encontradas, como aparece em casos com sintagmas complexos, em que a regra da conjunção coordenada é superada pela restrição semântica.(See also Baltazart & Kister, 1996).

Segmento 8: The book ^{i} on divorce ^{j} and its ^{j} consequences

Segmento 9: The book ^{i} on divorce ^{j} and its ^{i} author.

Nossa investigação optou por uma análise das restrições sintáticas, semânticas e textuais, sem usar o conhecimento de mundo ou metodologias mais abstratas como aquelas citadas pela análise do discurso. Pressupõe–se que uma descrição completa de aspectos restritos ofereça uma contribuição maior para o mapeamento de todo processo.

MÉTODOS

A resolução da anáfora é uma questão crucial no Processamento da Linguagem Natural. Muitos projetos na área da lingüística computacional, incluindo a recuperação de informações, processamento

de diálogos e tradução automática, têm que alocar uma parte do sistema para resolver o problema. Decidir em qual estágio do processo atacar o problema depende de muitos aspectos, incluindo as abordagens teóricas que estão sendo usadas. Para a abordagem proposta aqui, baseada em um projeto de tradução automática do inglês para o português, a anáfora é abordada depois que algumas análises preliminares já foram realizadas sobre o texto que está sendo processado, incluindo o seguinte:

Atribuição da classe gramatical: Cada palavra no texto já deve ter sido classificada numa das classes gramaticais básicas (substantivo, verbo adjetivo, etc.) e nas subclasses (verbo transitivo, verbo intransitivo, etc.).

Junção de atributos específicos: número (singular e plural), traços semânticos (+ humano, + animado, etc.) e especificações de gênero especificado se for necessário na tradução para o Português (masculino, feminino) também são acessórios do SN.

Segmentação dos sintagmas nominais: sintagmas complexos, envolvendo combinações de dois nomes (*stone houses*), adjetivos e substantivos (*the big house*) foram segmentados com a identificação do núcleo correspondente. A segmentação também inclui combinações de mais de um SN como *the president of the United States, Bill Clinton, and England's Prime Minister, Tony Blair*, que forma um SN plural complexo.

Atribuições de caso: a função sintática (nominativa, acusativa e dativa,etc.) da NP resultante já é conhecida.

Tabela 1 mostra como dois SN são classificados. Note que *a large house in the mountains* é

classificado como SN, singular já que todas as palavras que fazem parte do sintagma nominal são governadas pelo núcleo *house.*

Tabela 1 – Segmentação dos sintagmas nominais

Turistas	preferem	uma casa grande nas montanhas
SN Masculino Plural (+) Animado Nominativa		Substantivo Feminino Singular (-) Animado Acusativa

Para esta investigação dois pronomes diferentes foram escolhidos, usando duas linguagens diferentes. O pronome *they* em textos do Inglês e o pronome possessivo no Português. Há uma razão prática e teórica para essa escolha. Em termos teóricos espera-se que a análise explique as relações entre anáfora e o texto, de um ponto de vista estritamente lingüístico, independente da linguagem que esteja sendo usada. A questão que se procura responder é se é possível resolver a anáfora sem recorrer ao conhecimento de mundo – ou seja, até que ponto é possível uma solução usando apenas restrições sintáticas e semânticas. Em termos práticos, os resultados podem ser imediatamente aplicados a traduções automáticas da língua Inglesa para muitas línguas românicas, como o Francês, Espanhol ou Português. De um lado, há a ambigüidade do pronome *they*, que surge quando se passa de uma língua para outra; por outro lado, as dificuldades especiais dos pronomes possessivos no Português que não dependem das restrições, do gênero e do número. Acredita-se que essas dificuldades são a raiz de muitos problemas interligüísticos que uma vez resolvidos, podem levar a soluções práticas envolvendo

a anáfora.

PRIMEIRO ESTUDO

O primeiro estudo considerou a relação entre o pronome *they* e o antecedente. A metodologia básica envolveu a seleção de 1.400 ocorrências do pronome *they* extraídas de um *corpus* de 10 milhões de palavras de textos expositivos. Para essa a seleção foi usado um programa de concordância. Esse tipo de programa permite que uma palavra ou combinações de palavras sejam automaticamente extraídas do *corpus* e listados de acordo com a ordem de seleção (ordem alfabética pela primeira palavra da esquerda, pela primeira palavra da direita, segunda palavra, etc.), facilitando, assim as diferentes análises.

Depois que os 1.400 segmentos foram selecionados, o antecedente foi identificado e classificado de acordo com sua função sintática (sujeito, objeto direto, objeto indireto, etc.). No segmento 10 por exemplo o antecedente é “the Aztecs” e tem função de sujeito.

Segmento 10: Continually dislodged by the **small city-states**^{h} that fought one another in **shifting alliances**^{i}, **the Aztecs**^{j} finally found refuge on a small island in Lake Texcoco where, about 1345, ***they***^{j} founded the town of Tenochtitlan.

Os traços semânticos do verbo que seguia o pronome também foram analisadas em termos dos traços que exigiam para o sujeito. Isso pode ser visto no segmento 11 onde há 7 candidatos para antecedente de *they* (*economists, solutions, problems, economies, markets, prices* and *exports*), mas apenas o SN “economists” pode ser escolhido porque, embora esteja mais longe

do anaforizante, é o único que pode ser sujeito de "cite" sem produzir anomalia semântica.

Segmento 11: **Economists[g]** who disagree with imposed **solutions[h]** to Third World development **problems[i]** point to the excessive vulnerability of Southern **economies[j]**, which are largely dependent for their growth upon relatively open Northern **markets[k]** and reasonable international **prices[l]** for their **exports[m]**. ***They[g]*** cite the need to involve local populations (…).

A metodologia prática usada para encontrar as restrições semânticas foi simplesmente alinhar os possíveis candidatos do mais próximo ao mais distante, partindo do anaforizante, até que um antecedente viável seja encontrado. Isso é mostrado abaixo – exemplo retirado do segmento 11 – onde o sintagma nominal adequado só é encontrado na sétima tentativa.

They cite the need to involve local populations.

1. *** exports** cite the need to involve local populations.
2. *** prices** cite the need to involve local populations.
3. *** markets** cite the need to involve local populations.
4. *** economies** cite the need to involve local populations.
5. *** problems** cite the need to involve local populations.
6. *** solutions** cite the need to involve local populations.
7. **Economists** cite the need to involve local populations.

Uma heurística em forma de algoritmo foi usada para detectar as restrições sintática e semântica disponíveis no texto. O Quadro 1 resume o procedimento utilizado nesta investigação.

Quadro 1 – Procedimentos para resolução da anáfora

Fase de Testagem Sintática	Fase da Testagem Semântica	Resultados
Fase 1: Procure o sintagma nominal plural a esquerda de they, até 80 palavras no texto ou sujeito singular. Se o sintagma nominal foi encontrado passe para a etapa 2. Caso contrário vá para a etapa 4 (dentro do segmento de 80 palavras do texto).	**Passo 4: Procure o sintagma nominal a esquerda *they*, até 80 palavras no texto ou primeiro sujeito singular.** Essa etapa só é executada se o limite de 80 palavras for encontrado sem se chegar à condição do espaço 2 (Função Sintática). O procedimento começa de novo; desta vez considerando apenas as restrições semânticas. Assim, se o SN for encontrado, vá para a etapa 5 . Caso contrario (dentro do segmento de 80 palavras do texto) vá para etapa 6.	**Passo 6: Solução não encontrada.** Se o SN plural é encontrado no limite das 80 palavras adotar o procedimento?. (exemplo traduzir o *they* para masculino ou feminino ou pronome zero). Vá para a etapa 7. **Passo 7: Procedimento**

Passo 2: O sintagma nominal tem a mesma função sintática de they? Se a resposta for sim vá para a etapa 3 se for não, volte para a etapa 1. **Passo 3: O sintagma nominal pode substituir "they" sem produzir anomalia semântica?** Se a resposta for sim, vá para a etapa 7, se for não volte para a 1.	**Passo 5: O sintagma nominal pode substituir "they" sem produzir anomalia semântica?** Se resposta for sim vá para a etapa 7; se for não volte para a etapa 4.	**Final.** Procurar outras ocorrências de anáfora.

O procedimento é devidido em duas fases de testagem, cada uma levando a uma solução se o candidato a antecedente passar nos testes semântico e sintático. Usando o exemplo 11 para demonstrar a fase sintática, podemos ver que todos os candidatos do texto, com exceção de *economistas* não vão além do passo 2, o que significa que são descartados no nível sintático (Não possuem paralelismo sintático por não compartilhar da mesma função de sujeito que possui o anaforizante). Somente o sintagma nominal *economistas* chega à etapa 3. Ao passar o teste as etapas 4 e5 são ignoradas e neste caso a anáfora é resolvida.

Deve ser observado que no procedimento proposto aqui o paralelismo sintático por ele mesmo, (sujeito/ sujeito) não é qualificado para decidir se um SN pode ser classificado ou não como antecedente para o anaforizante. O paralelismo sintático está sujeito, portanto, a restrições semânticas. A etapa3 é o primeiro ponto de decisão: se a solução é encontrada, o procedimento é finalizado, se não, o procedimento recomeça, voltando ao passo1. O procedimento é repetido até a octogésima palavra à esquerda ou um substantivo na posição do sujeito for encontrado.

Descobriu-se que era necessário prosseguir com a busca quando o sujeito, tanto singular como plural, fosse um pronome. Isso não apenas evitaria problemas com expletivos (exemplo: ***It*** *is raining*.) mas também com os outros pronomes incluindo os pronomes possessivos e indefinidos. No Segmento 12, por exemplo, a procura pelo antecedente para no SN "The amnesty", porque ele é sujeito e é substantivo.

A fase de testagem semântica é ativada apenas se o SN passa pelo passo 3. Já que não foi encontrada solução considerando as restrições sintáticas e semânticas, uma segunda rodada começa agora, ignorando as restrições sintáticas. Isso pode ser demonstrado no segmento 12.

Segmento 12: An amnesty is an exemption from prosecution for criminal acts, usually issued by a government after a time of crisis such as a war or revolution. **The amnesty** may be for acts such as rebellion, treason, desertion, or draft evasion. It is usually granted to groups of **citizens**^{j} on condition that ***they***^{j} abide by the law in the future.

A primeira rodada termina, neste caso quando o primeiro sujeito singular expresso por um substantivo, for encontrado (*The amnesty*). A segunda rodada começa e chega a *citizens* como primeiro SN plural. Não é nem sujeito, mas como as restrições sintáticas não contam mias, o SN é somente testado pelas anomalias semânticas e passa no teste.

No caso do SN não ser um substantivo, mas um pronome plural, o procedimento continua, procurando por um substantivo, até que o limite das 80 palavras ou um sujeito singular seja encontrado. Isso pode ser visto no segmento 13, onde o processo começando pelo último *they*, passa pelo pronome they (in *They tried*) e pára em (*Mongol bands raided*).

Segmento 13: Following Kublai Khan's eventual overthrow of China's Song dynasty in 1279, **Mongol bands**^{i} raided much of Eastern Asia outside of China. ***They***^{j} tried in vain to invade Japan in 1274 and 1281, captured Burma's Pagan in 1287, and penetrated Champa and Annam in 1285-88. ***They***^{j} even attempted to invade Java in 1292-93.

Quando o procedimento descrito acima é incapaz de encontrar o antecedente do anaforizante, é marcado como não resolvido e um valor default pode ser usado. Isso pode ser observado no Segmento 14, por exemplo. O processo pararia em *Poseidon*, na sentença anterior por que ele é um sujeito singular, mas sem encontrar o SN plural – que neste caso acontece uma combinação do sujeito (*Perseus*) com objeto (*Andromeda*).

Segmento 14: When Cassiopeia boasted that Andromeda was more beautiful than the sea-goddesses called Nereids, Poseidon, god of the sea and father of the Nereids, sent a sea monster to ravage Ethiopia. Only the sacrifice of Andromeda could persuade Poseidon to call off the monster, so Andromeda was chained naked to a sea cliff. The hero Perseus saw her plight, rescued her, and killed the monster. Thereupon, Poseidon turned the dead monster into the sea's first coral. **Perseus**^{j} married **Andromeda**^{j}, and **they**^{j} eventually became king and queen of the Greek city of Tiryns.

SEGUNDO ESTUDO

Para o segundo estudo, os pronomes possessivos *seu, sua, seus, suas* da Língua Portuguesa foram escolhidos. Esses pronomes apresentam algumas características que os tornam a interessantes para uma investigação sobre a resolução da anáfora.

A característica mais importante é que no Português esses pronomes não concordam com o possuidor, como acontece no Inglês, mas com a coisa possuída. Assim, enquanto em Inglês temos "Mary^{1} her^{1} father, and her mother^{1}" – onde a escolha do pronome possessivo depende

do que há à esquerda – em Português temos "Maria[1] seu[1] pai e sua[1] mãe." – onde a escolha dos pronomes depende do que há à direita. Isso quer dizer que esses pronomes só podem ser traduzidos do Inglês para o Português colocacionalmente. Mais do que isso, eles não apenas dependem das palavras que os estão rodeando, mas também das palavras que se colocam em direções opostas.

A principal conseqüência desta concordância pelo lado direito (com a coisa possuída) é que o rastreamento do antecedente se torna mais difícil. Já que não se podem usar pistas sintáticas importantes, como gênero e número. No segmento 15 o leitor usando apenas regras baseadas no gênero, não deve ter dificuldade em escolher entre Bill e Mary como antecedente legitimo para his ou her. Quando esses dois segmentos são traduzidos para o Português, os dois pronomes possessivos his e her são unificados em um só e tornam: *seu* (Segmento 16), tornando impossível o rastreamento do antecedente apenas pelo uso de pistas de gênero. Essa redução de pistas apresenta um desafio para a resolução da Anáfora baseada nas restrições textuais, um desafio que merece ser investigado.

Segmento 15: **Bill**^{i} told **Mary**^{i} that he wanted **his**^{i} car

Bill^{i} told **Mary**^{i} that he wanted **her**^{i} car.

Segmento 16:

Bill^{i} disse a **Mary**^{i} ele queria **seu**^{i} carro.

Bill^{i} disse a **Mary**^{i} ele queria **seu**^{i} carro.

A terceira característica no que diz respeito aos pronomes possessivos é que diferentemente de they, que só ocorre em posição de sujeito (**They** visited Bill" mas nunca "Bill visited **they**"), os pronomes possessivos podem ocupar posições diferentes na sentença("**Her** car arrived",

"He drove **her** car", "He was arrested in **her** car",etc.). O paralelismo sintático com pronomes possessivos parece assim não ser muito útil.

Pressupõe-se que todas essas dificuldades, surgidas com a redução das pistas textuais, sobreacarregam a hipótese de que apenas as restrições sintáticas e semânticas, sem conhecimento de mundo, possam levar a uma resolução da anáfora. Resultados positivos, obtidos em condições tão desfavoráveis, serão provavelmente mais robustos e confiáveis.

A metodologia usada para testar a hipótese consistiu de um levantamento de um corpus de 1.300 ocorrências de pronomes possessivos em textos jornalísticos. A base para a obtenção desses dados foi o CD-ROM da *Folha de São Paulo* e Internet (jornais e revistas).

O procedimento usado para a obtenção dos dados baseou-se nos recursos disponíveis no próprio programa que gerencia o texto eletrônico da *Folha*, e que oferece um sistema de busca através de qualquer palavra do texto. Na Internet, usou-se um procedimento semelhante. Dessa maneira foi feita a montagem do corpus. A Figura 1 mostra a tela correspondente à busca das palavras *seu, sua, seus, suas*. Na Figura 2, vê-se uma tela com parte dos resultados.

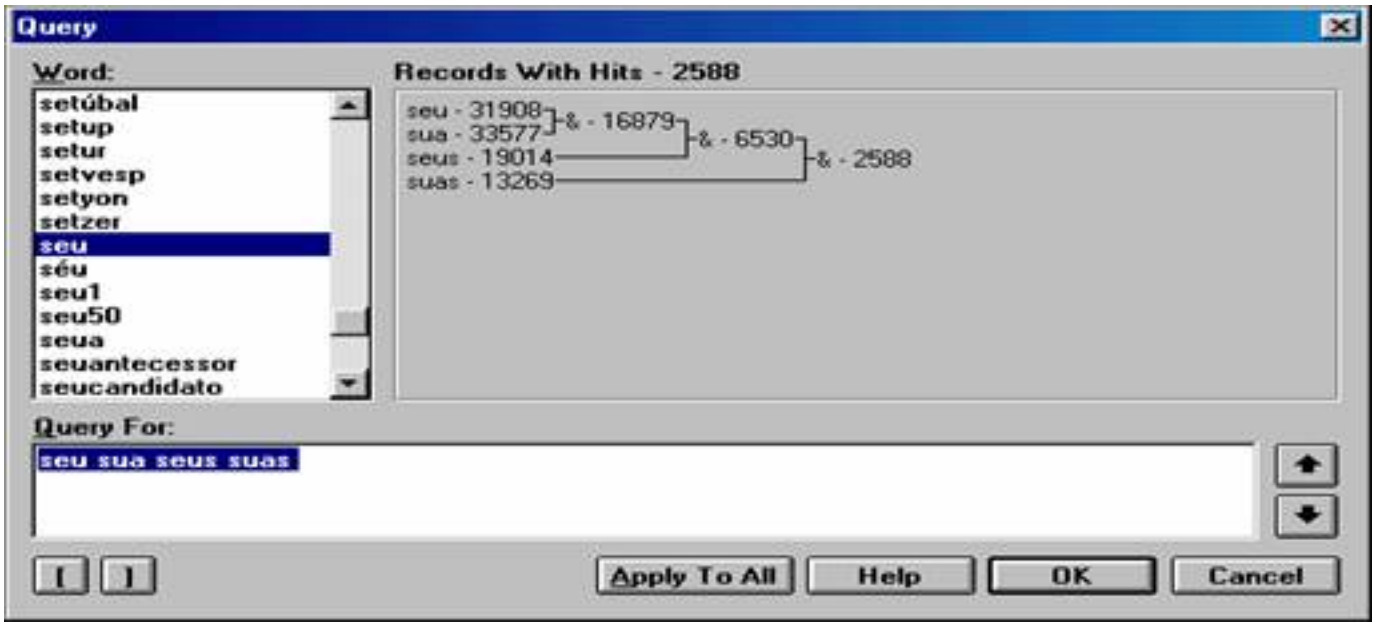

Figura 1 – Tela da etapa de busca.

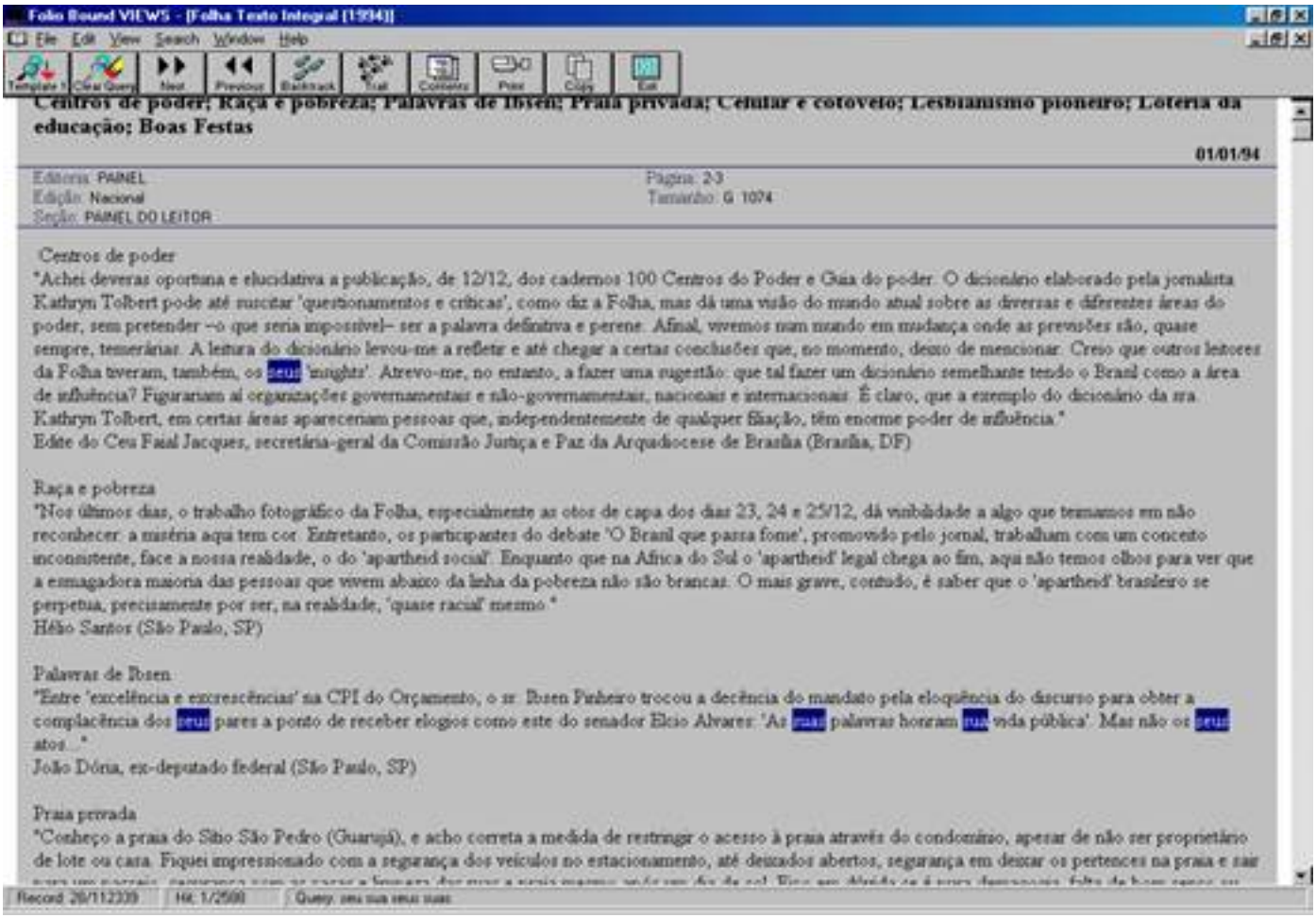

Figura 2 – Tela com seleção de exemplos

Um algoritmo simplificado foi usado para o levantamento do anaforizado, dentro de cada ocorrência, e que constou das etapas mostradas no Quadro 2.

Quadro 1 – Algoritmo para localização do anaforizado: possessivo.

1. Etapa 1: Procure “e” ou “ou” à esquerda do pronome possessivo. Se a resposta for sim, procure o primeiro Sintagma Nominal (SN) sem passar por verbo. Encontrando o SN, ele é o antecedente; caso contrário, passe para a etapa 2.

2. Etapa 2: Procure, à esquerda do pronome possessivo, um verbo que enfatize o objeto. Se a resposta for sim, procure o objeto. Encontrando o objeto, este será o antecedente; caso contrario passe para etapa 3.

3. Etapa 3: Procure o primeiro SN sujeito à esquerda do pronome possessivo. Se ele for encontrado e for semanticamente compatível com a coisa possuída, o SN é o antecedente. Caso contrário passe para etapa 4.

4. Etapa 4: Procure uma expressão indicadora de tópico à esquerda do pronome possessivo. Se ele for encontrado, procure o primeiro SN à sua direita. Encontrando-o, este será o antecedente. Caso contrário, passe para etapa 5.

5. Etapa 5: Procure o primeiro SN à esquerda do pronome possessivo. Se ele for encontrado, será o antecedente. Caso contrário, a anáfora não foi resolvida.

RESULTADOS E DISCUSSÃO

Esta investigação tentou responder três perguntas: (1) qual o percentual de resolução de anáfora que pode ser resolvido aplicando-se as restrições sintáticas e semânticas; (2) qual o percentual de acerto ao se aplicar apenas restrições semânticas, ignorando-se, portanto, o paralelismo sintático; e finalmente (3) qual o percentual de casos não resolvidos. A Tabela 3 mostra esses resultados, incorporando o pronome *they* e os possessivos *seu, sua, seus, suas*:

Tabela 3 – Nível de êxito com restrições sintáticas e semânticas.

Mecanismo de rastreamento	%
Paralelismo sintático	86%
Paralelismo semântico	12%
Não resolvido	2%

O paralelismo sintático é o fator mais significativo, resolvendo sozinho, 86% dos casos. Isso significa que simplesmente buscando um SN na posição de sujeito, ignorando as restrições semânticas, deixa apenas 14% dos casos não resolvidos.

Se as restrições semânticas, porém, forem consideradas, mais 12% dos casos são resolvidos, elevando o percentual para 98%. Uma revisão da bibliografia com relatos de investigações que usaram restrições sintáticas e semânticas, combinadas com abordagens estatísticas mostra que este é o percentual mais alto obtido até o momento. A Tabela 4 resume os resultados de alguns desses estudos com a resolução da anáfora pronominal em diversas línguas, incluindo inglês, polonês e árabe.

Tabela 4 –Índice de sucesso na resolução da anáfora

Estudo	%
Baldwin (1997)	75%
Mitkov (1998) (English)	89.7%
Mitkov (1998) (Polish)	93.3%
Mitkov (1996)	94.7%
Mitkov & Belguith (1998)	95.2%
Mitkov (1998) (Arabic)	95.2%
Mitkov & Stys (1997)	95.8%

O percentual de 98% obtido em nossa investigação surpreende, especialmente se considerarmos que o procedimento usado aqui foi muito mais simples do que aqueles usados em outros estudos, às vezes combinando escalas complexas de preferências e abordagens estatísticas com restrições sintáticas e semânticas.

Uma possível explicação é que o pronome “they” pode ser fácil em termos de rastreamento do

anaforizado, quando comparado com outros pronomes. Por outro lado, já que o anaforizado tende a ser o foco do parágrafo, é possível que essa condição facilitaria a resolução.

Na verdade, a condição de sujeito para o anaforizado pareceu ser bastante poderosa, mesmo em frases com elevando nível de subordinação, como no Segmento 15, onde o SN *founders*, embora numa oração subordinada, é o antecedente do pronome – e seria corretamente selecionado pelo algoritmo, já que é o primeiro SN a ocupar a posição de sujeito.

Segmento 15: **Historians^{i}** continue to debate what the nation's **founders^{j}** meant to include when ***they^{j}*** wrote that there shall be "no law" abridging the freedom of speech or press,

Gostaria de argumentar, no entanto, que o alto índice de resolução é devido a uma combinação e ordenamento de preferências e restrições sintáticas e semânticas, como foi usada nos dois algoritmos. De fato, se as restrições semânticas não tivessem sido aplicadas no exato momento em que o SN sujeito fosse encontrado, os resultados seriam bem diferentes.

Considerando apenas o paralelismo sintático, 94% (não 86%) dos segmentos investigados satisfariam a condição, mas iriam produzir uma margem de erro de 14% (em vez de 2%). Isso pode ser demonstrado no Segmento 16: aplicando-se apenas paralelismo sintático, o antecedente selecionado seria *farmers*, porque, tal qual *they*, está na posição de sujeito. A escolha de *farmers,* no entanto, seria incorreta porque o antecedente certo é *chickens*, ainda que sem paralelismo sintático, por estar na posição de objeto. Mas as restrições semânticas, baseadas no verbo *purchase*, favoreceria *chickens* mais do que

farmers – já que mercadorias têm mais probabilidade de serem adquiridas do que pessoas – e assim, de acordo com o logaritmo proposto, *chickens* seria corretamente selecionado.

CONCLUSÃO

A resolução da anáfora, usando apenas restrições sintáticas e semânticas, sem recorrer ao conhecimento enciclopédico ou de mundo, tem um lado bonito e um lado feio. O lado bonito é o alto índice de acerto, que, ao alcançar percentuais acima de 95%, fica próximo do nível de falantes fluentes da língua. Quantitativamente, os resultados podem ser interpretados como excelentes. O lado feio é a qualidade dos erros produzidos, muitas vezes ridículos de uma perspectiva de conhecimento baseada no senso comum e na intuição humana.

A tentação é concluir que existe na resolução da anáfora muito mais do que aparece na superfície textual e que o conhecimento de mundo parecer ser no fim a única fonte confiável. Recorrer ao conhecimento de mundo, no entanto, significa apenas transferir o problema para um nível mais alto de abstração sem conseguir resolvê-lo. O senso comum, a intuição, as variáveis sócio-históricas, e outros componentes do conhecimento de mundo são muito vagos para serem adequadamente tratados pela Lingüística Computacional.

Uma solução para evitar a ocorrência de erros ridículos tem que ficar além das restrições morfológicas baseadas na concordância de gênero e número ou outros paralelismos sintáticos entre anaforizantes e anaforizados – tais como as simetrias sujeito com sujeito, objetivo direto com objeto direto, etc. – mas não pode ir tão longe até chegar ao que tradicionalmente se define como conhecimento de mundo; as

restrições são incontroláveis nesse nível. Possíveis caminhos que poderiam ser explorados aqui incluem o conceito de colocação – iniciando com a idéia de Firth de que uma palavra é conhecida pela companhia com que anda, e incluindo a contribuição de Hoey (1991) sobre os padrões de repetição do léxico, onde a ênfase está mais nas relações lexicais do que gramaticais. As metarregras de Charolles (1988), explorando a necessidade de ordens combinatórias e conexão lógica entre os itens lexicais do texto, poderiam também ser úteis.

Qualquer solução encontrada na anáfora pode contribuir para outras áreas do estudo da língua como a resolução da ambigüidade, coesão textual e, eventualmente, a compreensão de leitura e produção textual. A relação entre anáfora e ambigüidade, por exemplo, está tão próxima que é provavelmente impossível fazer referência a uma sem usar a outra, sendo a anáfora por si mesma um tipo de ambigüidade. Isso vale também para a coesão textual, considerando que o discurso é uma seqüência lógica de idéias costuradas entre si de acordo com certas preferências e restrições. Em termos mais práticos, podemos também argumentar que as descobertas realizadas pelos estudos sobre a anáfora eventualmente contribuirão para a instrução em leitura e escrita, mostrando aos alunos quais são os mecanismos usados para ligar as diferentes partes do texto.

REFERÊNCIAS BIBLIOGRÁFICAS

BALTAZART, D. & KISTER, L. Is it Possible to Predetermine a Referent Included in a French *N* de *N* Strucutre ? In: S. P. Botley & A.M. Mc Ennery (eds) *Discourse Anaphora and Anaphor Resolution Colloquium*. Lancaster University, 17-18th julho 1996, Lancaster UK.

CHAROLLES, M. Introdução aos problemas da coerência dos textos (abordagem teórica e estudo das práticas pedagógicas). In: GALVES, C; ORLANDI, E.; OTONI, P. (orgs). *O texto: escrita e leitur*a. Campinas, Pontes, 1988.

Dagan,I.; ITAI, A. Automatic processing of large corpora for the resolution of anaphora refer-ences. *Proceedings of the 13th International Conference on Computational Linguistics*, COLING'90, Helsinki, 1990

DU BOIS, John. Beyond definiteness: the trace of identity in discourse. In: CHAFE, Wallace (ed.). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex, 1980. p. 203-274.

FOX, Barbara. Introduction. In: FOX, Barbara (ed.). *Studies in anaphora*. Amsterdam: John Benjamins, 1996. p. vii-xi.

HOEY, Michael. *Patterns of lexis in tex*t. Oxford: University Press, 1991.

INDURSKY, Freda. Da anáfora textual à anáfora discursiva. *Anais do 1º. Encontro do Círculo de Estudos Lingüísticos do Sul – CelSul.* Florianópolis: UFSC, 1997. p. 713-

LANGACKER, Ronald W. Conceptual groupings and pronominal anaphora. in: FOX, Barbara (ed.). *Studies in anaphora*. Amsterdam: John Benjamins, 1996. p. 333-378.

McEnery,T.; Botley, S. (Eds) *Discourse Anaphora and Anaphor Resolution*. Amsterdam, John Benjamins, 1998.

Mitkov R. - Anaphora resolution: a combination of linguistic and statistical approaches. Proceedings of the Discourse Anaphora and Anaphor Resolution. Lancaster University, UK, 17-19 July 1996

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

Mitkov, R. A new approach for tracking center. In Proceedings of the International Conference *New Methods in Language Processing*, UMIST, Manchester, UK, 13-16 September 1994.

Mitkov, Rusla; & Belguith,Lamia. Pronoun resolution made simple: a robust, knowledge-poor approach in action. *Proceedings of the International Conference "Traduction Automatique et Langage Naturel" (TALN'98*). Paris, France, 1998.

Mitkov, Rusla; & Belguith,Lamia. Pronoun resolution made simple: a robust, knowledge-poor approach in action. *Proceedings of the International Conference "Traduction Automatique et Langage Naturel" (TALN'98*). Paris, France, 1998.

Nasukawa, T. Robust method of pronoun resolution using full-text information. *Proceedings of the 15th International Conference on Computational Linguistics* COLING'94, Kyoto, Japan, 5-9 August 1994.

Nasukawa, T. Robust method of pronoun resolution using full-text information. *Proceedings of the 15th International Conference on Computational Linguistics* COLING'94, Kyoto, Japan, 5-9 August 1994.

PARABONI, Ivandré. *Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa*. Dissertação de mestrado. Porto Alegre: PUCRS, 1997.

van HOEK, Karen. *Paths through conceptual structure: Constraints on pronominal anaphora.* Doctoral dissertation. San Diego: University of California, 1992.