Cad.Est.Ling., Campinas, (45):119-132, Jul./Dez. 2003

O CONHECIMENTO DE COLOCAÇÕES NO PROCESSAMENTO DE RELAÇÕES ANAFÓRICAS

MARCO ROCHA
(UFSC)

ABSTRACT This paper presents a corpus-based approach to the study of anaphora. A sample of anaphoric demonstratives collected in a dialogue corpus is analysed with the purpose of investigating collocations and patterns of text, associating anaphors with antecedents by means of a data-driven classification which attempts to capture cognitive processes.

1. INTRODUÇÃO

Abordagens que incorporam a noção de colocação a uma teoria da língua, no sentido da associação de uma forma lingüística a um sentido, representam, na maioria dos autores que as propõem, uma contestação da composicionalidade da semântica. Mais especificamente, Bolinger (1975, 1976) não apenas aponta as limitações da semântica composicional, mas vê as colocações, as quais chama de "prefabrications" ou "prefabs", como um dos dois principais traços da organização textual. Sinclair (1987) define estes dois princípios da organização textual como o "open-choice principle" e o "collocational principle".

O primeiro compreende a língua, do ponto de vista da produção, como uma série contínua de decisões livres, isto é, decisões restringidas apenas pela gramática, ou sistema formal da língua, quanto às possíveis classes das palavras a serem empregadas em cada posição na seqüência sintagmática. Na realidade, porém, estas escolhas são fortemente reduzidas pela ação de restrições de natureza não-gramatical. O falante possui, armazenadas no seu léxico mental, uma grande quantidade de elementos prontos que, embora aparentemente analisáveis em segmentos, constituem na verdade escolhas únicas. O princípio das colocações é assim compreendido, neste estudo. Deste modo, desenvolve uma das idéias centrais da lingüística de *corpus* (ver, por exemplo, McEnery e Wilson 1996), resumida no aforismo de Firth (1957), "thou shalt know a word by the company it keeps".

Segundo esta abordagem, é possível estudar fenômenos lingüísticos através de levantamentos exaustivos de ocorrências deste fenômeno em *corpora*. A capacidade de busca e recuperação de dados do computador digital permite a realização rápida destes levantamentos com uma taxa de erro desprezível. A manipulação destes dados de modo a permitir a definição dos padrões de co-ocorrência também é relativamente simples e segura, do ponto de vista operacional. A utilização sistemática de recursos tecnológicos na análise de fenômenos lingüísticos passa a ser, nesta abordagem, um aspecto indispensável da metodologia de pesquisa. Noções de base estatística, como freqüência e probabilidade, tornam-se parte integrante do arcabouço teórico que sustenta a análise, afastando-se de abordagens baseadas em regras deterministas.

Ao mesmo tempo, as abordagens a partir de *corpus* abriram perspectivas importantes para a solução de problemas difíceis em tecnologia das línguas humanas. As abordagens lógicas, baseadas

em regras previamente especificadas a partir de uma teoria lingüística, ou uma combinação de teorias, não conseguem abranger a variedade do uso da língua, além de gerar dificuldades quanto à implementação de decisões relacionadas ao uso da regra adequada em um dado momento do processamento. A lingüística de *corpus*, como alternativa de abordagem em lingüística, reflete, em grande parte, o debate entre simbolismo e conexionismo, em termos de soluções computacionais.

A pesquisa na qual se insere o presente estudo diz respeito ao processamento de relações anafóricas em interfaces de diálogo, talvez a aplicação que mais claramente caracteriza o objetivo tecnológico da lingüística computacional, ou seja, a "máquina que fala". A abordagem utilizada na pesquisa parte de um *corpus* de diálogos autênticos, o Corpus de Diálogos Clínicos do Rio de Janeiro[1] (doravante, CDC-RJ), para classificar as relações anafóricas, conforme observadas no corpus, segundo um modelo definido à medida em que se desenvolvia a análise de ocorrências de termos anafóricos e seus antecedentes. Esta análise indica que a atuação do princípio de colocações é real também no nível das relações textuais complexas. O estudo concentra-se nos demonstrativos anafóricos, associando os padrões de co-ocorrência destes termos, ou seja, as colocações em que aparecem no corpus, à identificação dos antecedentes, na tentativa de compreender melhor as estratégias usadas em situações particularmente complexas de identificação de antecedentes implícitos e textuais.

O restante deste trabalho está organizado da seguinte maneira: na próxima seção, o conceito amplo de colocação utilizado no estudo é definido; na terceira seção, é detalhada a utilização deste conceito no âmbito da investigação das relações anafóricas aqui descrita; na quarta seção, as relações entre colocação, anafóra e topicalidade são explicitadas com base na observação de ocorrências de demonstrativos anafóricos encontrados no corpus; na quinta, a utilização destes resultados para a construção de um modelo cognitivo do processamento de anáforas é discutida, explorando também suas relações com aprendizado de máquina, com vistas à construção de interpretadores de anáfora mais eficazes. Finalmente, algumas observações quanto aos aspectos cognitivos e computacionais deste estudo são brevemente apresentadas, servindo de conclusão provisória ainda dependente de resultados futuros.

2. O CONCEITO DE COLOCAÇÃO

A partir das idéias de Firth (1957), foram desenvolvidos aspectos distintos da noção de colocação (ver Partington 1998 para uma resenha mais completa)[2], sobretudo após o "renascimento" das abordagens a partir de *corpus* da segunda metade da década de oitenta, impulsionado pela disseminação do computador digital e pela elaboração de programas de manipulação de *corpus*. Sinclair (1991) apresenta a seguinte definição para a noção de colocação:

> Collocation is the occurrence of two or more words within a short space of each other in a text. (Sinclair 1991:170)

A base desta definição é de natureza *textual*, isto é, um item coloca-se com outro se aparece em algum ponto próximo deste outro no texto. Alternativamente, uma definição de "sentido colocacional", apresentado por Leech (1974) como um dos sete tipos de sentido possíveis, pode ser caracterizada como uma definição *psicológica*, uma vez que o tipo de sentido em questão é definido como "the associations a word acquires on account of the meanings of words which tend to occur in its

[1] O corpus ainda não se encontra disponível para a comunidade de pesquisa. Possui 46660 palavras e foi coletado em um hospital do Rio de Janeiro.

[2] O histórico aqui apresentado faz uso sistemático da resenha em Partington (1998).

environment" (Leech 1974: 20). Deste modo, a noção de colocação é incorporada à competência comunicativa do falante nativo, como uma das formas de construção do sentido.

Através da exposição sistemática à língua de sua comunidade, o falante acaba por conhecer quais são as colocações normais e quais as incomuns em circunstâncias dadas. Também Aitchinson (1994: 21) observa este processo de construção do sentido com base nos padrões de co-ocorrência, ao constatar que "humans learn word-meaning from what occurs alongside". Problemas de compreensão, sobretudo em situações novas ou desconhecidas, são resolvidos pelo exame do contexto imediato, ou co-texto, em busca de informações que possam esclarecer o sentido de um item desconhecido ou permitir que um item polissêmico possa ser interpretado de maneira correta.

A visão do fenômeno das colocações sob uma ótica probabilística, expressa pela noção de normalidade, dadas as circunstâncias, é resumida por Hoey (1991) em uma definição, chamada de *estatística* em Partington (1998), que se adequa bem como base prática para investigações em lingüística de *corpus* de um modo geral, uma vez que pode ser facilmente transformada em um procedimento estatístico incorporado a um programa de manipulação de *corpus*. Esta definição, apresentada abaixo, destaca o interesse dos fenômenos de co-ocorrência quando sistematicamente repetidos, isto é, quando formam padrões com um propósito específico, em termos de sentido.

> Collocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context. (Hoey 1991: 6-7)

Embora o termo *colocação* seja geralmente associado aos padrões de co-ocorrência de duas palavras isoladas, a idéia pode ser aplicada a todos os níveis de análise lingüística. Não existe nenhuma diferença qualitativa, em termos teóricos, entre colocações de palavra com palavra e colocações de palavra com sintagma, sintagma com sintagma, sintagma com oração e mesmo oração com oração. Conforme apontado por Partington (1998), todas estas associações podem ser estudadas como fenômenos de padronização textual, uma área de investigação lingüística para a qual o autor sugere a denominação de *combinatorics*.

Estes padrões podem ser definidos como troncos lexicalizados de sentenças ou como schemata (Barlow 1996). Estas unidades fixas ou semi-fixas não se encaixam facilmente no léxico, uma vez que são longas demais e contêm regularidades internas, mas também não são passíveis de análise em termos sintáticos, pois se organizam como unidades. Os troncos lexicalizados de sentença contêm um ou mais elementos que constituem uma classe, e não apenas um item. Em uma situação dada, como, por exemplo, desejar a alguém uma boa viagem como forma de expressar consideração, o falante poderia utilizar diversas formas, de acordo com as circunstâncias do momento, tais como:

> Espero que faças uma boa viagem.
> Espero que tenhas feito uma boa viagem.
> O diretor manda dizer que espera que o senhor tenha feito boa viagem.

Estas formas poderiam ser caracterizadas como variantes de um padrão de co-ocorrência, chamado anteriormente de tronco lexicalizado, tal como o mostrado abaixo, onde os elementos verbais não-flexionados terão suas realizações determinadas pelos contextos de uso (os elementos entre parênteses são opcionais e FLEX significa flexão):

> *esperar*-FLEX *que fazer*-FLEX (uma) *boa viagem*

Este padrão é o tronco de sentença, e suas diferentes realizações são resultado das exigências contextuais impostas pela situação de comunicação. Também são possíveis expansões deste padrão

básico, como, por exemplo, o acréscimo de expressões como *de volta* ou o uso de discurso indireto, como na terceira realização acima.

Os *schemata*, dada uma associação entre forma e sentido, são o componente da forma que contém algumas das características de uma expressão fixa, mas que também contém partes variáveis as quais conseguem adaptar-se de modo a transmitir informações que variam com o contexto. Deste modo, na expressão *deixar-se levar*, há duas variáveis: o verbo *deixar*, que muda conforme a concordância com o sujeito e o contexto temporal do evento; e o pronome reflexivo, que varia dentro desta classe fechada também conforme o sujeito. Não obstante, esta variabilidade está delimitada pela estrutura do esquema, que é constituída por *deixar-FLEX* + reflexivo + *levar*.

O mesmo acontece com expressões tais como *dar* + reflexivo + *por vencido* ou *fazer* + possessivo + *independência financeira*, onde elementos de significado fixo são associados às variações de natureza gramatical, compondo esquemas que são usados sempre que uma determinada situação se repete. O reconhecimento da situação baseia-se na vivência social do falante, que a associa aos esquemas lingüísticos. Esta associação, porém, nunca é determinista, pois sempre há possibilidade de variação. Deste modo, elementos de raciocínio probabilístico são associados às possíveis variações lingüísticas. Na próxima seção, este conceito estendido de colocação será utilizado para a construção de hipóteses explicativas que especifiquem possíveis estratégias de processamento para a resolução de referências anafóricas onde o termo anafórico é um demonstrativo.

3. COLOCAÇÕES E RELAÇÕES ANAFÓRICAS

As relações anafóricas são um aspecto intensivamente estudado nas pesquisas relacionadas aos estudos da linguagem, uma vez que a complexidade destas relações permanece sendo uma capacidade de difícil explicação em termos cognitivos, isto é, a identificação de antecedentes no âmbito de relações anafóricas continua a colocar questões nada triviais para uma teoria que se proponha a explicar como as relações discursivas operam no sentido de associar os termos anafóricos a estes antecedentes. Não obstante, os demonstrativos anafóricos têm recebido pouca atenção nestes estudos, talvez porque, quase que necessariamente, sua resolução, em termos de identificação do antecedente, exija uma análise de relações textuais em relação às quais existe ainda pouco consenso entre os investigadores.

Na abordagem aqui desenvolvida, foram analisadas 171 ocorrências de demonstrativos anafóricos em seis diálogos do CDC-RJ. As ocorrências foram divididas em quatro categorias segundo o tipo de demonstrativo anafórico. A distribuição dos demonstrativos anafóricos analisados na amostra é mostrada na Tabela 1 abaixo:

Tabela 1: Distribuição dos demonstrativos anafóricos incluídos na amostra segundo tipos de termo anafórico

isso e contrações	1 18
o, a, os, as	22
aquilo e contrações	10
esse, aquele, e suas flexões e contrações	21
Total	1 71

O reconhecimento destes quatro tipos de termo anafórico dentre os demonstrativos não é problemático, uma vez que está baseado estritamente na realização. O demonstrativo *esse* e suas flexões[3] precisam ser reconhecidos como casos de anáfora, excluindo da amostra as ocorrências - mais freqüentes na proporção de cinco para um nos diálogos analisados - onde a função é de determinante. A classificação não constitui problema para o analista, e, em princípio, também não deveria representar dificuldade maior para um procedimento automático, tal como um etiquetador de categorias morfossintáticas. Igualmente, classificar os casos em que *o* e suas flexões são demonstrativos anafóricos, separando-os das ocorrências em que têm função de determinante, é, de um modo geral, simples.

A análise dos antecedentes é, comparativamente, bastante mais problemática que a dos termos anafóricos. O primeiro passo é reconhecer este antecedente, uma tarefa que pode ser difícil, sobretudo no caso dos demonstrativos. Talvez pelo fato dos fenômenos anafóricos serem percebidos, em muitos estudos, como uma análise das relações textuais onde o termo anafórico é um pronome pessoal, espera-se que o antecedente de um termo anafórico dado seja um elemento do texto previamente apresentado sob a forma de um sintagma nominal[4] (doravante, SN) que descreve ou representa um referente do discurso. A estratégia de processamento fundamental para os fenômenos anafóricos seria a busca entre os candidatos possíveis, dentre os já apresentados anteriormente. No caso de catáforas, seria necessário detectar que o antecedente é apresentado mais adiante no texto. Não é raro, porém, que sejam necessárias operações bem mais complexas para que o antecedente possa ser identificado. A comparação dos dois exemplos[5] abaixo, ambos do demonstrativo anafórico *isso*, mostra que, no exemplo (2), a identificação de um antecedente não é trivial.

(1) A: o arroz, quanto que a senhora coloca no prato?
B: ah, eu coloco duas colheres
A: só porque? de sopa?
B: duas colheres de sopa, mas não é,
ããã, colher de sopa da vovó não
A: sei
porque? a senhora não agüenta comer
mais que isso**?**

(2) B: ele me passou um exame de sangue,
que quando eu fui fazer não tinha
tinha a possibilidade de fazer
A: hãã mas quando é que foi **isso**?
B: foi em...em agosto

No exemplo (1), uma busca na fala precedente aponta, primeiramente, para *colher de sopa da vovó* e, em seguida, para *duas colheres de sopa*. É difícil ter certeza quanto às possíveis dificuldades de processamento que a existência de dois candidatos semelhantes muito próximos um ao outro, e ao termo anafórico, poderiam causar para a identificação do antecedente correto. Não obstante, o

[3] Doravante, a menção do demonstrativo *esse* refere-se também a *aquele*, suas flexões e contrações.

[4] O termo sintagma nominal é usado neste estudo sem preocupação com sua vinculação anterior a teorias sintáticas específicas.

[5] Todos os exemplos foram extraídos do CDC-RJ praticamente sem alteração, exceto por pequenas diferenças relacionadas à omissão de sinais e anotações acrescentados pelo analista que não dizem respeito à ocorrência discutida no exemplo. O termo anafórico analisado está marcado em negrito em todos os exemplos.

problema é essencialmente avaliar um ou mais candidatos apresentados no texto precedente, muito freqüentemente dentro do contexto imediato, e decidir qual é o mais adequado.

Contudo, no exemplo (2), a identificação de um antecedente exige decisões menos óbvias. O antecedente não é um referente do discurso no sentido habitualmente utilizado na literatura (ver, por exemplo, Karttunen 1976), mas sim uma passagem da fala anterior ou, possivelmente, o turno de fala como um todo, editadas as repetições e outras alterações típicas da língua falada, além das modificações relacionadas à concordância. Porém, um analista poderia tratar o antecedente como um SN implícito derivado do texto precedente, por exemplo, *a ida de B para fazer o exame de sangue que ele* (termo anafórico que se espera tenha sido resolvido) *passou, mas que B não pode fazer*. Este tipo de solução gera problemas de identificação precisa, uma vez que os ajustes considerados necessários podem variar de analista para analista, ou mesmo em duas análises distintas pelo mesmo analista. Conseqüentemente, a definição de explicitação do antecedente torna-se difícil. O antecedente deve ser tratado como sendo a enunciação anterior, pura e simplesmente, e portanto explícito, ou não?

Ao mesmo tempo, parece importante especificar os mecanismos por meio dos quais um falante decide que um antecedente explícito, isto é, que pode ser identificado através de uma busca em uma lista de possíveis candidatos do texto precedente, não é a solução adequada para um determinado caso de referência anafórica. O mesmo parece ser verdade para o contraste entre antecedentes nominais, isto é, representados por um SN claramente delimitado, e antecedentes textuais, ou seja, fragmentos de fala de dimensões variáveis que podem ou não estar circunscritos a uma estrutura sintagmática delimitada. A análise dos casos encontrados no CDC-RJ foi vista como um primeiro passo nesta direção.

Os antecedentes foram classificados com base nestas duas dicotomias: a distinção entre antecedente *explícito* e antecedente *implícito*; e a distinção entre antecedente *nominal* e antecedente *textual*. A terminologia anteriormente utilizada para a análise das relações anafóricas, a saber, *termo anafórico* e *antecedente*[6], foi estendida aos casos de catáfora. Sempre que o analista se decidiu por um sintagma nominal ou fragmento de texto de qualquer tipo, mas claramente identificável no texto, como sendo o antecedente do caso de anáfora analisado, este foi classificado como um antecedente *explícito*. No exemplo (1), o antecedente é explícito e nominal, no (2), explícito e textual. Um exemplo de antecedente implícito é dado abaixo.

(3) A: e como é que você 'tá no geral da tua saúde assim ?
B: ah, geral da minha saúde eu 'tou toda ruim
A: hãã (risos)
B: acho que se for fazer um check-up,
não tem nada que escape
A: hãã
B: porque eu tenho problema de intestino,
eu tenho problema de estômago
A: uum
B: eu tenho problema de de de vista é o principal,
é **o** que mais 'tá me atrapalhando, agora

O demonstrativo *o* se refere a *problemas de saúde*, antecedente que pode ser inferido do discurso sem necessidade de operações particularmente complexas. Não obstante, é preciso definir, em termos de estratégia de processamento, como a identificação do antecedente ocorre, isto é, como

[6] Este tratamento é comum na literatura. Ver Bosch (1985).

a enumeração de problemas de um determinado tipo acaba levando à conclusão de um efeito local de topicalidade que permite a identificação de um antecedente implícito que é o tópico do segmento de discurso em questão. Esta relação entre topicalidade e anáfora, associada ao conehcimento de colocações, voltará a ser abordada no plano do modelo cognitivo.

Há um outro grupo de ocorrências, porém, cujos antecedentes parecem melhor classificados como irrelevantes ou inexistentes. Isto, é claro, dificulta a compreensão da idéia de referência, uma vez que uma relação anafórica depende, em princípio, do estabelecimento de um vínculo entre um termo anafórico e seu antecedente. Porém, os demonstrativos tipicamente anafóricos, como *isso*, aparecem, com alguma freqüência, como falsos termos anafóricos, cujos antecedentes, pelo menos à primeira vista, são muito difíceis de identificar, mesmo com esforços de reconstrução textual. Estas ocorrências foram classificadas como *não-referenciais*, e é apresentado abaixo um exemplo de uma ocorrência assim classificada.

(4) B: e 'tou, o pessoal foi,
A: justamente
B: foi saindo, foi casando, essa coisa toda,
e, agora, em casa, somos duas pessoas
A: por isso que se explica
B: **isso** aí, é, conta na balança, né ?
A: justamente, justamente
B: nem tenho dúvida

Neste fragmento, a descrição da seqüência de acontecimentos é apenas esboçada, resumida numa expressão como *essa coisa toda*, e a classificação do antecedente da primeira ocorrência de *isso* é feita em bases semelhantes às do exemplo anterior. Porém, a segunda ocorrência, na expressão *isso aí*, tem características bem diferentes, pois tem função muito semelhante às duas ocorrências de *justamente* que compõem o turno seguinte e à reiteração sob a forma de *nem tenho dúvida*. Confirmam o que já foi dito, expressando, além do posicionamento dos falantes em relação à questão, uma forma de aliança entre os interlocutores. O antecedente propriamente dito não parece relevante para o processamento. Mais uma vez, são os elementos prosódicos que permitem uma interpretação mais segura deste valor puramente de confirmação genérica, distinto do exemplo anterior. A Tabela 2 em seguida mostra o cruzamento dos tipos de termo anafórico com os tipos de explicitação do antecedente descritos acima.

Tabela 2 – Tabulação cruzada entre tipo de termo anafórico e explicitação do antecedente

Termo anafórico	Explicitação do antecedente			
	Explícito	Implícito	Não-refer.	Total
isso e contrações	78	24	16	1 18
o, a, os, as	12	4	6	2 2
aquilo	5	3	2	1 0
esse, essa, aquele e plurais	21	0	0	2 1
Total	116	20	24	1 71

Os antecedentes explícitos são, portanto, os mais comuns, somando 67,83% do total dos casos na amostra. O pronome *esse* e suas flexões só aparecem como anafóricos vinculados a antecedentes explícitos. O percentual de antecedentes explícitos para o demonstrativo *isso* é de 66,1%, enquanto os pronomes *o, a, os, as* apresentam uma proporção de 54%, e *aquilo* precisamente 50%. Porém, o total de casos é pequeno para estes tipos de termo anafórico, e as conclusões devem ser evitadas no presente estágio.

A Tabela 3 apresenta os mesmos tipos de termo anafórico agora tabulados contra a classificação de *nominal* ou *textual* acrescida de uma terceira categoria, denominada *mecanismo de foco*, criada para abrigar os casos de antecedente inexistente ou excessivamente vago para permitir identificação. Uma vez que não se trata de um termo anafórico no sentido estrito, já que não desempenha esta função, a denominação não classifica um antecedente, que, para todos os propósitos práticos, não existe, e sim o próprio demonstrativo.

Tabela 3 - Tabulação cruzada do tipo de termo anafórico com a estrutura do antecedente

Termo anafórico	Tipo de estrutura interna do antecedente			
	Nominal	Textual	Foco	Total
isso e contrações	51	40	27	118
o, a, os, as	12	4	6	22
aquilo	6	2	2	10
esse, essa, aquele e plurais	21	0	0	21
Total	90	46	35	171

Como pode ser observado, o pronome *esse* e suas flexões têm apenas antecedentes nominais. O percentual de nominais em *aquilo* chega a 60%, deixando exatos 20% para cada uma das outras duas categorias, enquanto os pronomes *o, a* e seus plurais têm este tipo de antecedente em 54% dos casos, mas possuem maior número de mecanismos de foco (27%) e apenas 18% de antecedentes textuais. Como dito anteriormente, o tamanho da amostra não favorece conclusões. No tipo de termo anafórico mais numeroso, a distribuição é bastante mais equilibrada, uma vez que apenas 43% dos casos, em números redondos, são classificados como nominais, enquanto 34% são textuais e 23% são mecanismos de foco.

O tipo de termo anafórico não é conhecimento suficiente para que seja possível especificar a explicitação ou a estrutura interna do antecedente. Por outro lado, em algum ponto do processo parece igualmente indispensável definir que características têm os antecedentes a serem identificados, de modo a restringir as opções possíveis. A investigação evoluiu, portanto, no sentido de uma análise mais detalhada das ocorrências da amostra, de modo a tentar associar possíveis padrões de co-ocorrência do termo anafórico a uma solução em termos de explicitação ou estrutura interna do antecedente.

A análise das ocorrências de cada tipo de termo anafórico foi feita de modo a tentar associar características do contexto imediato, quando repetidas sistematicamente, a uma estratégia de identificação do antecedente. Duas limitações importantes devem ser levadas em conta. A primeira, já suficientemente enfatizada, é o tamanho da amostra. À medida em que a especificação dos padrões de co-ocorrência aumenta, o número de casos que se enquadram nestes padrões torna-se cada vez menor, tornando as hipóteses relativas às estratégias de identificação dos antecedentes gradualmente menos confiáveis. A segunda é a inexistência de testes adequados, tanto psicolingüísticos quanto computacionais, para corroborar estas hipóteses iniciais com dados experimentais que pudessem

indicar a possibilidade de validação da hipótese. Tanto a expansão das dimensões da amostra quanto a realização dos testes mencionados são objetivos futuros da pesquisa.

Uma estratégia de resolução provavelmente útil é a separação entre os termos realmente anafóricos e aqueles que, na verdade, não se referem a nenhum antecedente no sentido estrito do termo, com base no princípio das colocações. Em expressões como *por isso ou por aquilo* e *é isso aí*, na amostra analisada, as ocorrências do demonstrativo tipicamente anafórico *isso* exercem função diversa da esperada, uma vez que a identificação de um antecedente específico, seja explícito ou implícito, nominal ou textual, não parece essencial para a interpretação semântica da fala.

Em outros padrões identificados, o antecedente é sempre ou predominantemente explícito ou implícito, nominal ou textual, e ainda anafórico ou catafórico. Através da classificação sistemática destes padrões, parece ser possível avançar no sentido da especificação de um modelo cognitivo abrangente para o processamento de relações anafóricas, cujo plano de construção está sendo exemplificado no presente estudo por meio dos demonstrativos anafóricos, justamente por apresentarem uma freqüência relativamente alta de casos cuja resolução não é trivial.

4. TOPICALIDADE, COLOCAÇÕES E ANÁFORA

Mesmo com um *corpus* anotado bem maior do que o atualmente disponível, não seria possível especificar uma estratégia de identificação do antecedente com base em padrões de co-ocorrência para todos os casos de demonstrativos anafóricos. Na amostra analisada, há casos em que o estabelecimento da relação entre um padrão de co-ocorrência e uma estratégia de identificação de antecedente é muito difícil, tendo em vista as características inespecíficas do contexto imediato. Muito freqüentemente, estas ocorrências também possuem antecedentes implícitos que constituem nominalizações baseadas no evento ou eventos descritos no turno de fala precedente. O exemplo (3), na seção anterior, é um destes casos. O exemplo (5) abaixo mostra um outro.

(5) A: e a senhora, hoje em dia,
utiliza algum tipo de remédio?
B: um... para ir ao banheiro?
A: não, para qualquer coisa
a senhora utiliza
B: não, só que essa essa a semana passada
eu estava com Binotal, né?
A: uhum, para que?
B: para porque eu extraí um dente
eu fiz o canal primeiro, mas
fiquei sentindo [uma dorzinha
A: e **isso**] dificultou a sua mastigação?
B: o quê, o o dente?
A: é
B: está dificultando sim
mas agora eu não posso colocar, enquanto não lixar né?
A: uhum
B: está muito recente

A função sintática de sujeito de um predicado verbal constitui já uma indicação de preferência por antecedente nominal. É essencial também para a análise a sobreposição de falas, marcada pelos

colchetes no exemplo. O demonstrativo anafórico na fala de *A*, transcrito acima na linha subseqüente, foi realizado, na verdade, simultaneamente ao sintagma nominal *uma dorzinha*, da fala de *B*, uma vez que *A* não esperou que *B* terminasse o seu turno. A sobreposição é um fenômeno comum em diálogos autênticos, e, nesta ocorrência em particular, é fundamental para a interpretação da referência anafórica. O SN *uma dorzinha* não pode ser o antecedente, fato que é corroborado posteriormente com o pedido de explicitação da referência por parte de *B*, que usa *o dente* em entoação interrogativa para verificar a correção do antecedente identificado. A identificação correta é confirmada por A em seguida.

Não obstante, a interpretação de uma fala reconstruída *o dente está dificultando a mastigação* não pode ser feita composicionalmente com base no sentido isolado das palavras. O sintagma *o dente* precisa ser compreendido como uma espécie de redução do sintagma *a extração do dente*, o qual não aparece no texto. Embora possa ser inferido a partir da fala *porque eu extraí um dente* com aparente facilidade, a possibilidade de uma inferência que identificasse *canal* (*do dente*) como antecedente parece, à primeira vista, mais ativada, uma vez que canal está explícito no texto anterior. Embora não tão freqüente, o fenômeno já foi registrado em vários trabalhos na literatura relacionada a anáforas sob o nome de "strained anaphora" (ver, por exemplo, Hirst). Trata-se, portanto, de um antecedente nominal implícito que deve ser inferido a partir de uma fala, e de cuja identificação depende o processamento semântico do texto.

Diferentemente dos padrões de co-ocorrência anteriormente discutidos, não há soluções diretamente baseadas nas palavras do contexto imediato que permitam distinguir estas ocorrências de outras onde o antecedente não está implícito. Uma vez que a referência anafórica acontece na fala subseqüente de *A*, dois ou três movimentos depois, dependendo de como se considere a interrupção, especula-se que *A* tenha detectado a mudança de tópico que a fala de *B porque eu extraí um dente* realizou, afastando a atenção da questão dos remédios. É possível que, se a interrupção de *A* não tivesse ocorrido, *B* continuasse a conduzir o diálogo como se estivesse falando de *remédios*, mas o fato é que a intervenção de *A* consolida uma mudança de tópico iniciada por *B*, mas que *B* não tinha intenção de realizar. Trata-se de um tópico de conversação negociado, onde a referência anafórica tem papel crucial na consolidação do acordo entre *A* e *B*. O curso da negociação leva *B* a procurar confirmar a mudança de tópico, e a identificação do antecedente, por meio do SN reduzido com entoação de pergunta.

Deste modo, o desenvolvimento de uma hipótese de processamento completa para o processamento de demonstrativos anafóricos terá que acabar por lidar com a questão do acompanhamento do tópico. Parece necessário incorporar a idéia de segmentação do diálogo, ou de qualquer outro texto, segundo a evolução do tópico, de modo a permitir que possíveis antecedentes implícitos possam ser hierarquicamente considerados com base no tópico atual. Isto significa especificar, em cada momento do diálogo, qual o tópico vigente, ou, pelo menos, quais os possíveis tópicos vigentes, a partir da fala dos participantes.

Poder-se-ia imaginar uma estratégia de processamento que, inicialmente, esgotasse todas as possibilidades de correspondência com padrões de co-ocorrência previamente estabelecidos. Caso não fossem encontradas correspondências, guardadas proporções aceitáveis de adaptação, elementos do discurso em destaque seriam considerados como candidatos, incluindo tópicos locais e globais em uma hierarquia. No caso dos demonstrativos anafóricos do tipo do exemplo (11), as ocorrências analisadas parecem apontar para uma prioridade do tópico local do segmento, mas seu número ainda é muito pequeno para conclusões. Foi encontrado pelo menos um caso em que o tópico global do diálogo, conforme detectado pelo analista, é o antecedente correto. Há indícios, além disso, de uma interação entre topicalidade e antecedentes textuais, a qual ainda não foi coerentemente integrada ao esboço de modelo, onde o fragmento adequado seria sinalizado pela presença de um tópico local ou global.

Por outro lado, a hipótese de processamento paralelo também deve ser considerada, não apenas como explicação para a velocidade das adaptações, mas como a única forma de integrar uma gama tão ampla de informações, que inclui todos os níveis de análise lingüística e aspectos de conhecimento enciclopédico e experiencial. Porém, a especificação de uma hipótese completa de processamento paralelo, sobretudo no que diz respeito ao reconhecimento das características textuais, inclusive padrões de co-ocorrência, pode ser muito difícil. De alguma maneira, porém, as ações que levam à identificação dos antecedentes precisam ser definidas em um modelo que possa ser testado, tanto nos aspectos cognitivos quanto computacionais. A próxima seção explora brevemente estas dificuldades.

5. ANÁFORA: PLANO PARA UM MODELO COGNITIVO

As investigações com o CDC-RJ encontram-se limitadas pelo tamanho da amostra. As idéias para um futuro modelo cognitivo são, no atual estágio das investigações, uma definição metodológica, um plano de como proceder para construir este modelo, muito mais do que uma especificação deste modelo. O princípio das colocações, conforme definido anteriormente, guia o procedimento inicial de levantamento dos padrões de co-ocorrência que possam se demonstrar reveladores em termos de especificação de estratégias de identificação de antecedentes. A análise da amostra discutida aqui identificou padrões que provavelmente se repetem em larga escala. Entretanto, a verificação disto em um número estatisticamente adequado de ocorrências ainda precisa ser realizada. Cada padrão será então associado a um procedimento de reconhecimento e a uma estratégia de identificação do antecedente.

Na análise da amostra, 33,9% dos casos foram identificados a padrões estáveis de co-ococorrência, como *pelo menos isso*. Utilizando o recurso WebCorp (2003), foram coletadas 55 outras ocorrências da colocação. A análise confirmou que os antecedentes são explícitos e textuais em todos os casos. Este é o único padrão já plenamente testado. O processo de testagem de cada um dos padrões isolados pela análise das ocorrências no CDC-RJ acabará por estabelecer aqueles que apontam para estratégias estáveis de identificação do antecedente, associadas à explicitação e estrutura interna deste último, conforme a classificação acima.

Nos casos em que o reconhecimento dos padrões não é suficiente para definir uma estratégia adequada de identificação do antecedente, as informações relacionadas à topicalidade podem ser a fonte de subsídios para o processamento da referência anafórica. A topicalidade foi inicialmente explorada para a resolução de anáforas em diálogos em Fox (1987) e, do ponto de vista computacional, em Grosz e Sidner (1986), e posteriormente em muitos outros estudos. Os elementos do discurso em destaque são os antecedentes preferenciais, segundo uma hierarquia que ordena o teste de adequação de cada um dos diferentes tipos de tópico ou elemento de destaque no discurso. As hierarquias podem variar de acordo com as características do co-texto, e estas variações são guiadas pela análise de ocorrências anteriores que seguem padrões iguais ou semelhantes. Também conforme os resultados da análise da amostra, seria possível resolver 26,3% dos casos.

Outros 24% dos casos da amostra foram resolvidos por meio da estratégia básica de busca no texto precedente, com a identificação do primeiro sintagma nominal adequado, segundo critérios tradicionais de concordância e restrições selecionais, como o antecedente correto. Em muitos destes casos, o termo anafórico é o pronome *esse* ou uma de suas flexões. Portanto, a capacidade de reconhecer o tipo de termo anafórico seria suficiente para distinguir as ocorrências em que há uma probabilidade maior de um antecedente ser identificado pela estratégia básica tipicamente associada ao antecedente nominal explícito. Isso não significa dizer que não existam casos do termo anafórico

isso com antecedentes que possam ser reconhecidos por meio desta estratégia. Estes são, no entanto, bastante mais raros.

Outra estratégia importante no processamento dos demonstrativos anafóricos são as referências dêiticas, onde o antecedente é um elemento presente ou visível na situação na qual o diálogo ocorre. Estas ocorrências são resolvidas por meio de recursos visuais, na enorme maioria dos casos, e os antecedentes são geralmente classificados como implícitos, pelo menos do ponto de vista do discurso, e nominais. O reconhecimento da referência dêitica também pode ser facilitado pelos padrões de co-ocorrência, uma vez que, em 72% dos casos de dêixis, o demonstrativo é seguido por um advérbio de lugar como *aqui* ou *aí*. Em relação ao total da amostra, 7,2% dos casos foram resolvidos com base nesta estratégia de utilização de dados visuais. Mais uma vez, as conclusões devem ser vistas com as devidas restrições de tamanho da amostra.

Uma última estratégia detectada na análise das ocorrências do *corpus* é a referência a um ou mais membros de um conjunto previamente definido no discurso, sem que este membro tenha sido apresentado em isolamento anteriormente, como no exemplo (6) abaixo:

(6) A: E a senhora usa adoçante ?
no lugar ?
B: Uso
B: Porque eu estava usando um que tem parece ciclomato
não é ciclomato, né ?
A: Que ?
B: Eu estava usando um que tem...
A: Tinha ciclamato.
B: Pois é.
Aí ela passou outro não é ?
Mas agora
A: Fin ?
Foi por acaso Fin ?
B: Parece que são dois nomes.
A: Doce menor ?
B: Ai meu Jesus **esse** eu não guardei não
eu sei que é um leitoso
A: Uhum
B: que **o** que estava usando parecia água, não é ?
A: É.
B: e **esse** agora é assim parecendo leite.

Nesta passagem, há uma série de referências a adoçantes específicos não mencionados anteriormente, após ser definido o conjunto dos adoçantes como tópico local. Estas referências utilizam numerais (*um*), pronomes indefinidos (*outro*) e, no final, três demonstrativos anafóricos em seqüência. Na primeira ocorrência, não há modificadores. Já na segunda, a oração relativa define o antecedente implícito no conjunto dos adoçantes, função que é realizada pela palavra *agora* na terceira. As ocorrências deste tipo foram reunidas sob a denominação de **uso da noção de conjunto** e constituem 6,5% do total de ocorrências da amostra. Há ainda um pequeno número de ocorrências (3%) cuja classificação ainda não foi definida. A Tabela 4 abaixo resume a distribuição das ocorrências segundo as estratégias de processamento brevemente descritas nesta seção (as percentagens foram arredondadas).

Tabela 4 - Distribuição das estratégias de processamento

Colocações	58 (33,9%)
Topicalidade	45 (26,3%)
Primeiro Candidato	40 (23,3%)
Dêixis	12 (7,2%)
Noção de conjunto	11 (6,4%)
Indefinidos	5 (2,9%)
Total	171 (100%)

O procedimento metodológico é, portanto, associar cada termo anafórico a uma estratégia de processamento. Um termo anafórico cujo co-texto não seja apropriado para tratamento como uma colocação, nem permita o reconhecimento de um padrão associado à estratégia baseada na topicalidade, poderia ser então resolvido através de uma busca de primeiro candidato, uma referência dêitica ou através da noção de conjunto. As prioridades podem ser alteradas de acordo com o termo anafórico. Uma vez reconhecida a estratégia adequada, a caracterização do antecedente a ela associada, segundo a explicitação e a estrutura interna, devem permitir a identificação correta.

Este tipo de modelagem cognitiva está nitidamente associada ao que se convencionou chamar de aprendizado de máquina. Existem várias formas de levar a cabo processamento computacional desta natureza, mas, de um modo geral, o aprendizado em questão diz respeito a um treinamento de uma máquina, baseado na observação de casos previamente classificados por analistas especializados, de modo que, ao analisar um caso novo do fenômeno a que o aprendizado diz respeito, a máquina seja capaz de utilizar esta classificação com base nos casos em que o treinamento foi realizado. Deste modo, espera-se caminhar tanto na direção de uma hipótese de modelagem cognitiva para as relações anafóricas, quanto contribuir para a solução do difícil problema de tecnologia das línguas humanas geralmente chamado, neste âmbito, de resolução de anáforas.

6. CONCLUSÃO

A abordagem apresentada neste estudo parte da noção de que o princípio das colocações é um dos aspectos básicos de uma teoria da língua. A partir daí, procura desenvolver um modelo da atuação deste princípio no âmbito dos conhecimentos necessários ao processamento de relações anafóricas. Os dados, conforme observados em um *corpus* de diálogos, são a fonte fundamental de informações para a formulação de hipóteses na modelagem cognitiva pretendida, para a qual um plano metodológico de prática analítica é a conclusão mais palpável deste estudo no momento. O conceito de modelagem cognitiva empregado no estudo busca inspiração tanto na especificação de uma hipótese de processamento para os fenômenos anafóricos, consubstanciada no modelo cognitivo pretendido, quanto em propostas de um plano de ação para a resolução de anáforas em sistemas computacionais capacitados a processar línguas humanas.

Neste aspecto, a abordagem sugerida aqui parece aproximar-se dos modelos cognitivos associativos atualmente conhecidos como modelos paralelos distribuídos (ver, por exemplo, McClelland e Rumelhart, 1986). Por outro lado, uma vez que o aspecto da investigação relacionado à verificação empírica de hipóteses relacionadas ao processamento cognitivo e computacional é fundamental, a proposta de modelo cognitivo pode evoluir no sentido de graus variados de hibridismo (ver Sloman 1999), na medida em que uma proposta híbrida possa explicar melhor os fenôemnos observados e resolver com maior eficácia as dificuldades de implementação em sistemas reais. Em

conseqüência, o objetivo tecnológico da lingüística computacional é visto como um aspecto central da investigação e das soluções propostas.

REFERÊNCIAS BIBLIOGRÁFICAS

AITCHINSON, J. (1997). *The language web: (The Reith lectures)*. Cambridge: Cambridge University Press.

BARLOW, M. (1996). Corpora for theory and practice. *International Journal of Corpus Linguistics*. 1/1: 1-37.

BOLINGER, D. (1975). *Aspects of language*. Nova York: Harcourt Brace Jovanovich.
_____ (1976). *Meaning and memory*. Forum Linguisticum 1/1: 1-14.

GROSZ, B. e SIDNER, C. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics* 12, 3.

FOX, B. (1987). *Discourse structure and anaphora*. Cambridge: Cambridge University Press.

HIRST, G. (1981). *Anaphora in natural language understanding*. Berlim: Springer-Verlaf.

FIRTH, J. (1957). *Papers in linguistics*. Londres: Oxford University Press.

HOEY, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

KARTTUNEN, L. (1976). Discourse referents. In: J. McCawley (org.). *Syntax and semantics, vol. 7*. Nova York: Academic Press.

LEECH, G. (1997). *Semantics*. Harmondsworth: Penguin.

MCCLELLAND, J. e RUMELHART, D. (1986). *Parallel distributed processing, vol.2*. Cambridge, MA: MIT Press.

MCENERY, A. e WILSON, A. (1996). *Corpus linguistics*. Edimburgo: Edinburgh University Press.

PARTINGTON, A. (1996). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam: John Benjaminsx.

SINCLAIR, J. (1996). Collocation: a progress report. In: R. Steele e T. Threadgold (orgs.). *Language topics: essays in honour of Michael Halliday*. Amsterdam: John Benjaminx.

SINCLAIR, J. (1996). *Corpus, concordance, collocation*. Oxford: Oxford University Pressx.

SLOMAN, S. (1999). Cognitive Architecture. In: R. A. Wilson e F. C. Keil (orgs.) *The MIT encyclopedia of the cognitive sciences*. Cambridge, MA: The MIT Press.

WEBCORP. www.webcorp.org.uk. Research and Development Unit for English Studies, University of Liverpool. Acesso em 15 de agosto de 2003.