D.E.L.T.A., Vol. 16, Nº. 2, 2000 (229-261)

Relações Anafóricas no Português Falado: Uma Abordagem Baseada em Corpus

(A Corpus-based Approach to Anaphora in Spoken Portuguese)

Marco Rocha
(Universidade Federal de Santa Catarina)

Abstract: This paper describes corpus-based research on anaphoric relations in spoken Portuguese, relying on data collected in dialogues recorded in real-life situations. The essential analitycal tool is a corpus annotation which classifies each case of anaphora according to four attributes described in the paper. The research project as a whole is concerned with possible applications in natural language processing, particularly regarding natural language interfaces to databases.
Key-words: Anaphora; Corpus annotation; Corpus linguistics; Natural language processing.

Resumo: O trabalho descreve pesquisa baseada em corpus sobre relações anafóricas no português falado, desenvolvida a partir de dados coletados em diálogos gravados em situações da vida real. A ferramenta de análise essencial da pesquisa é uma anotação de corpus que classifica cada caso de anáfora segundo quatro atributos descritos no trabalho. O projeto de pesquisa como um todo está relacionado ao desenvolvimento de possíveis aplicações no processamento de linguagens naturais em sistemas computacionais, particularmente no que diz respeito a interfaces em linguagem natural para acesso a bancos de dados.
Palavras-chave: Anáfora; Anotação de corpus; Lingüística de corpus; Processamento de linguagens naturais.

Introdução

A investigação das relações anafóricas exige sempre uma definição inicial daquilo que se pretende analisar, uma vez que, em meio à vasta quantidade de estudos produzidos sobre o assunto, o termo **anáfora** é muitas vezes utilizado para significar fenômenos distintos. Conforme

assinala Bosch (1983), a palavra **anáfora** foi, em certo sentido, uma solução hábil para os problemas causados pelo termo **pronominalização**, pois o sentido literal da palavra **pronome** pode levar a interpretações inadequadas. Pronomes são mais do que um substituto para um substantivo ou sintagma nominal que poderia ser utilizado em seu lugar. O enfoque baseado na substituição também encontra dificuldades para lidar com referências pronominais a entidades do discurso que não foram explicitamente introduzidas ou cujos referentes são passagens inteiras de discurso (ver Hirst 1981, o próprio Bosch 1983 e Carter 1987 para levantamentos detalhados dos enfoques em questão).

As abordagens que permaneceram dentro dos limites da gramática sentencial – notadamente a gramática gerativa – desenvolveram estudos sobre anáforas sintaticamente controladas. Com isto, um grande número de casos foram desconsiderados como anáforas "pragmaticamente controladas", que não tinham lugar na teoria lingüística. Além disto, desenvolveu-se, como parte integrante destas abordagens, um hábito de criar exemplos, ao invés de extraí-los de dados observáveis no uso cotidiano da língua. Esta prática foi justificada com base na crença de que o verdadeiro conhecimento lingüístico deveria ser procurado fora da linguagem cotidiana conforme usada em contexto para fins de comunicação.

Uma consequência positiva da mudança de terminologia foi a possibilidade de associar aos pronomes fenômenos que não constituem referência pronominal, utilizando, não obstante, uma nomenclatura adequada. Sob o nome de anáfora, os pronomes podem ser analisados como uma manifestação de um processo muito mais amplo: o uso de uma variedade de mecanismos lingüísticos para gerar coesão, conforme definida em Halliday e Hasan (1976). Embora os pronomes permaneçam sendo o objeto de análise mais freqüente das pesquisas relacionadas às relações anafóricas, diversos estudos buscam discutir outras formas de referência anafórica, tais como sintagmas nominais anafóricos não-pronominais e elipses verbais – ver , por exemplo, Webber (1979) e Hoey (1991).

Esta expansão do conceito aconteceu em grande parte através de pesquisas que focalizavam fenômenos discursivos, como Fox (1987), as quais foram realizadas não apenas por lingüistas, mas também por pesquisadores nas áreas de psicolingüística e processamento

computacional de linguagens naturais (doravante, PLN). A necessidade de apresentar alternativas de explicação para problemas ainda difíceis de tratar – dentre eles a resolução de referências anafóricas –, aliada à dificuldade de mapear os modelos abstratos da lingüística de base sentencial até as enunciações da língua cotidiana, com os quais estes campos do conhecimento têm necessariamente que lidar, motivou o esforço tanto para incorporar aspectos textuais à análise dos fenômenos estudados, quanto para fortalecer a base empírica das investigações.

Diante destas variações no arcabouço teórico e metodológico no qual as pesquisas se inserem, não é surpreendente que a literatura produzida a respeito das relações anafóricas utilize o termo para significar uma gama variável de fenômenos lingüísticos[1]. Especificamente, as diferenças mais freqüentes nas investigações em questão dizem respeito à inclusão ou não de referências anafóricas intersentenciais; à utilização ou não de amostras de uso real da língua; e à inclusão ou não de uma variedade maior de termos anafóricos, embora o pronome de terceira pessoa permaneça sendo o termo anafórico prototípico e mais freqüentemente estudado.

A metodologia da lingüística de corpus oferece uma alternativa para aqueles pesquisadores que resistem ao distanciamento da teoria lingüística em relação à língua usada no cotidiano. As gramáticas e teorias, nas pesquisas baseadas em corpus, são desenvolvidas a partir de um levantamento abrangente de um número significativo de ocorrências de um fenômeno dado, em amostras de uso da língua em situações da vida real. Exemplos criados são a exceção, e não a regra. Todos os casos do fenômeno estudado são incluídos na análise, e noções de estatística, tais como freqüência e probabilidade, desempenham um papel central na formulação da teoria.

Além disto, as abordagens baseadas em corpus podem ser associadas aos modelos conexionistas em PLN, uma vez que estes modelos pressupõem habitualmente um corpus de treinamento. As redes conexionistas são atualmente uma alternativa de abordagem relativamente estabelecida em inteligência artificial, em parte devido às sérias

[1] Isto também é verdade no que diz respeito a outras noções importantes em lingüística.

dificuldades enfrentadas pelos sistemas precedentes de PLN ao lidar com a linguagem natural irrestrita. Segundo os que advogam a adoção de modelos conexionistas ou baseados em corpus em PLN, parte do problema é a preocupação excessiva, até então, com a formulação de regras de base lógica para lidar com as linguagens naturais, com uma contrapartida de descaso pela coleta e análise de dados e exemplos de uso da língua para comunicação (ver Harris 1992).

A quantidade de pesquisa produzida com uso de abordagens baseadas em corpus tem crescido ininterruptamente nos últimos anos, embora seja ainda pequena em termos de estudos orientados para as relações discursivas, como é o caso das relações anafóricas. A metodologia da lingüística de corpus não é exatamente uma abordagem nova, como demonstrado em Francis (1992). Contudo, o advento do computador digital alterou radicalmente as possibilidades deste tipo de abordagem, uma vez que o armazenamento de enormes quantidades de dados, sob a forma de corpora de grande porte, tornou-se relativamente fácil, particularmente com o barateamento do custo das máquinas nos últimos anos. A eficiência com que os computadores realizam operações de busca e recuperação permite que uma grande quantidade de ocorrências de um dado fenômeno seja analisada com rapidez e precisão.

Deste modo, a lingüística de corpus está intimamente relacionada à lingüística computacional, um termo genérico utilizado para abranger praticamente qualquer uso de computadores para a análise e geração de línguas humanas. Compreende-se, portanto, que uma parcela substancial da pesquisa produzida segundo abordagens baseadas em corpus venha da área de inteligência artificial, muitas vezes em projetos conjuntos com lingüistas. As abordagens baseadas em corpus constituem-se em uma alternativa importante para a solução de problemas de PLN que as abordagens baseadas em regras têm dificuldade de resolver. Este trabalho busca contribuir para este esforço de pesquisa e, portanto, discute sucintamente a possibilidade de utilizar os resultados aqui descritos em aplicações tais como interfaces em linguagem natural para acesso a banco de dados, tradução de máquina e aprendizado de línguas com ajuda de computador.

O estudo descrito em seguida teve como objetivo fundamental investigar as relações anafóricas em diálogos na língua portuguesa, de maneira a estabelecer padrões de ocorrência baseados no uso cotidiano da língua para comunicação. A fonte de dados utilizada é o Corpus de Diálogos Clínicos do Rio de Janeiro (doravante, CDC-RJ), cujas características serão descritas no decorrer do trabalho. O processo de formulação das conclusões partiu de um mínimo de noções teóricas a priori, buscando evoluir no sentido de uma gramática das relações anafóricas baseada na observação (Aarts 1991).

O estudo dos fenômenos anafóricos em diálogos reais envolve uma variedade de formas de referência, realizadas por pronomes, sintagmas nominais e formas verbais, muitas vezes organizados em cadeias de referência. Além disso, a interpretação correta do discurso falado requer o controle dos diferentes referentes em tempo real. Os fenômenos anafóricos são, na verdade, tão ubíquos, e aparecem sob formas tão diversas, que a definição do objeto de estudo é de fato bastante difícil. Ainda mais importante, as exigências de processamento que a resolução destas referências requer também são diferentes e não variam de maneira simétrica em relação aos diversos termos anafóricos[2]. Sendo assim, ocorrências distintas do mesmo pronome podem requerer processamento diferente, com uso diferenciado dos meios de resolução de anáforas que o discurso e o conhecimento lingüístico dos participantes fornecem. Por outro lado, ocorrências de termos anafóricos de tipo diferente podem ser resolvidas por meio de processos semelhantes.

Uma descrição dos fenômenos anafóricos adequada a uma abordagem baseada em corpus, e, portanto, tão desprovida de pressupostos teóricos quanto possível, parte da existência de elementos do discurso que estabelecem uma relação especial com um outro elemento deste mesmo discurso. A interpretação semântica, em seu aspecto textual, depende não apenas do reconhecimento da existência desta relação, mas também da identificação do antecedente correto, uma operação complexa que ultrapassa o estabelecimento de uma corres-

[2] Ver, a esse respeito, Koch e Marcuschi (1998).

pondência trivial entre os elementos em questão. Esta operação é chamada freqüentemente de resolução da anáfora.

Neste enfoque, portanto, anáfora é o nome dado a esta relação ou processo no qual um **termo anafórico,** em uma instância de discurso dada, se vincula a um elemento identificável – chamado de **antecedente** – para que a interpretação semântica seja realizada com êxito[3]. Estes elementos têm que estar presentes no discurso ou ser inferíveis do que foi dito. O ambiente físico circundante e a situação em que o discurso ocorre são também fontes cruciais de informação para que a interpretação correta se concretize, sobretudo nos casos de dêixis[4] na língua falada. Esta será a definição adotada neste trabalho.

Os casos de anáfora foram analisados com base no que estava foneticamente realizado, sem pressupor qualquer processo de resolução da referência em questão. Deste modo, as noções de pronome zero e categoria vazia não foram a priori consideradas necessárias para a construção do modelo de classificação utilizado nesta pesquisa para a análise de fenômenos anafóricos[5]. Conseqüentemente, a noção de **verbo anafórico** foi utilizada para classificar o termo anafórico nas ocorrências de anáfora em que a estrutura argumental do verbo requer a recuperação de um elemento do discurso. Sintagmas preposicionais e adverbiais foram também analisados como anafóricos em ocorrências em que a estrutura sintagmática incompleta de uma enunciação requer a recuperação de elementos do discurso para a interpretação semântica. A classificação de termos anafóricos, juntamente com os demais atributos associados a cada caso de anáfora, será descrita em maior detalhe na seção que aborda o modelo de classificação.

A forma de discurso na qual a pesquisa se concentra é o diálogo de obtenção de informações ou orientado para a realização de uma tarefa de qualquer tipo. No caso da língua portuguesa, dois aspectos das rela-

---

[3] Termo anafórico e antecedente são comumente usados também na análise de catáforas e dêixis, uma prática igualmente adotada neste estudo.

[4] Ver Bosch (1983) para uma discussão da distinção entre anafóra e dêixis.

[5] A observação dos dados do corpus acabou por demonstrar, posteriormente, que tais noções não seriam úteis ou mesmo plausíveis para os propósitos em questão.

ções anafóricas em diálogos chamam a atenção do analista. O primeiro aspecto é a omissão do sujeito ou do objeto, ou ainda de ambos, uma característica comum do português falado. A referência a uma entidade do discurso a ser identificada é detectada através da estrutura argumental, a qual estabelece o conjunto de argumentos essenciais aos diferentes verbos da língua. Uma vez que o sujeito não esteja realizado, é preciso identificá-lo no discurso ou inferi-lo com base nas informações transmitidas através deste contexto discursivo, sejam elas de natureza estritamente lingüística ou relacionadas ao conhecimento que decorre da experiência de mundo e da situação em que a conversação se passa.

O segundo aspecto digno de nota é o fenômeno das cadeias de referência, isto é, de termos anafóricos vinculados a outros termos anafóricos que os precedem em cadeia. Em última análise, a resolução ocorre através de um termo anafórico no início da cadeia. Estas cadeias são de extrema importância na língua falada, sobretudo em diálogos, onde são muito mais freqüentes. Se comparados à língua escrita ou à língua falada expositiva formal, os diálogos tipicamente lidam com um número bem menor de referentes aos quais se faz referência repetidamente (ver Biber 1992 para um análise comparativa dos sistemas de referência no discurso escrito e falado na língua inglesa). No caso do português, porém, estas cadeias são em parte construídas também com base na estrutura argumental.

Se comparado com línguas como o inglês e outras do ramo germânico, é possível obervar que o português não possui um pronome neutro que possa ser utilizado como termo anafórico nos casos em que o referente é um ser inanimado ou abstração. Ainda que a função de pronome neutro sobreviva em português, em certa medida, nos pronomes demonstrativos *isto, isso* e *aquilo*, a repetição sistemática destes pronomes soaria no mínimo estranha e intuitivamente inadequada em muitos contextos. Pode-se observar, deste modo, que o controle constante da estrutura argumental é um aspecto fundamental da interpretação semântica relacionada às referências anafóricas, permitindo a identificação de sujeitos e objetos que não estejam foneticamente realizados.

Ainda explorando a análise comparativa com a língua inglesa, poder-se-ia dizer que esta última se baseia em pronomes (como *he, she, it*

and *they*) e operadores (os verbos auxiliares de modo geral) para sinalizar a necessidade de recuperar elementos no discurso anterior, possibilitando a interpretação semântica. Em conseqüência, a omissão de elementos apresentados na pergunta torna as respostas uma forma de referência anafórica, uma vez que sua interpretação depende da recuperação destes elementos[6]. Em português, esta mesma função é realizada por formas verbais com argumentos omitidos.

Os dois aspectos mencionados acima e sua realização no sistema de referências do português falado foram sistematicamente explorados no estudo. O restante do artigo está organizado da seguinte maneira: na próxima seção, é descrita a metodologia empregada na coleta de dados e na análise propriamente dita; a seção subseqüente apresenta o modelo de classificação utilizado na análise dos casos de anáfora encontrados no corpus; a quarta seção discute os resultados do estudo e aponta possíveis desdobramentos significativos a serem desenvolvidos a partir destes resultados. A última seção resume a investigação realizada e sugere aplicações possíveis.

1. Metodologia

A descrição da metodologia utilizada no estudo está dividida em duas subseções. Na primeira, descreve-se o corpus coletado para os propósitos da pesquisa e o processo de coleta. Na segunda subseção, os fundamentos das abordagens de base em corpus são definidos em maior detalhe.

1.1. O corpus

Uma vez decidido que a pesquisa utilizaria uma abordagem baseada em corpus, concentrando-se no português falado, o próximo passo foi selecionar um corpus adequado aos propósitos da pesquisa. A idéia

[6] Esta interpretação da combinação pronome-operador típica das respostas na língua inglesa é discutida em detalhe em Quirk et al. (1985), seções 6.12-16.

de um corpus como fonte de material para pesquisa sobre linguagem não é nova para os pesquisadores brasileiros. Sob a influência de abordagens orientadas para a investigação sociolingüística, pelo menos uma iniciativa nacional de coleta de material da língua falada – o Norma Urbana Culta (NURC) – foi implementada e levada a cabo em várias capitais brasileiras. Diversos outros projetos de caráter local foram também realizados.

A maior parte das pesquisas, tanto no Brasil como em outras partes do mundo, se concentra em fenômenos fonéticos, morfológicos e sintáticos. O nível do discurso é menos freqüentemente enfocado. Várias razões contribuem para esta tendência, entre elas o grau muito menor de consenso em relação às teorias explicativas relacionadas aos fenômenos do discurso. Um segundo fator, diretamente relacionado aos corpora de diálogos, é que as exigências habituais de autenticidade não são tão facilmente atendidas quanto na língua escrita. A autenticidade de dados extraídos de um jornal, enquanto amostras de língua escrita, por exemplo, é praticamente garantida, uma vez que o texto não é produzido em conseqüência de uma iniciativa de pesquisa lingüística. Isto também é verdade para a quantidade enorme de textos escritos diariamente em muitas línguas, criando assim um montante substancial de dados disponíveis para os lingüistas.

A fim de atingir o mesmo nível de autenticidade em um corpus de língua falada, é necessário registrar diálogos que ocorram naturalmente em interações entre pessoas enquanto se dedicam a suas atividades diárias. As dificuldades que isto implica não são de pequena monta. A primeira decisão crucial é escolher hora e local apropriados para realizar as gravações. Isto geralmente exige negociações, já que as pessoas tendem a não aceitar muito facilmente a idéia de serem gravadas em situações que digam respeito a seu trabalho. As condições de gravação podem ser desfavoráveis ou mesmo imprevisíveis. Dependendo do ambiente onde as gravações ocorram, pode ser inteiramente impossível controlar interferências potencialmente desastrosas na rotina das gravações.

Deste modo, não é surpreendente que os pesquisadores que tentam coletar dados da língua falada prefiram métodos menos arriscados, o que geralmente significa gravar em ambientes protegidos, tais como

estúdios ou dependências das universidades. Os informantes recebem algum tipo de tarefa, a qual gera uma interação mediada pela fala, ou são simplesmente entrevistados por um pesquisador sobre algum tópico considerado adequado. Este é, sem dúvida, um método válido de obter dados da língua falada, mas as limitações, para propósitos de pesquisa que incluem a investigação de fenômenos do discurso, são inegáveis. Os dados coletados desta maneira não são autênticos *stricto sensu*, uma vez que as conversações não teriam ocorrido se uma iniciativa de pesquisa dada não estivesse em curso.

Neste sentido, o CDC-RJ é um corpus autêntico. As gravações foram feitas nas dependências da UnATI (Universidade Aberta da Terceira Idade), um projeto de tratamento holístico, pesquisa e formação de pessoal qualificado para a terceira idade, ligado à Universidade do Estado do Rio de Janeiro.A UnATI opera em um dos andares do edifício principal da Universidade do Estado do Rio de Janeiro. Suas atividades institucionais incluem vários cursos – tais como yoga, dança de salão, oficina de poesia e línguas estrangeiras – psicoterapia, recreação e aconselhamento alimentar e legal. No aspecto clínico, há consultas com médicos e enfermeiros, fisioterapia e entrevistas com os assistentes sociais. Estas últimas são geralmente voltadas para a seleção de novos alunos-pacientes para admissão na UnATI, segundo uma variedade de critérios em sua maioria relacionados à impossibilidade de obter tratamento de outro modo.

As gravações foram feitas em ambos os locais e contêm diálogos entre pacientes ou parentes de pacientes e a equipe, incluindo profissionais de assistência de saúde de todos os níveis, assim como alguns diálogos entre membros da equipe. Os gravadores foram operados pelos próprios membros da equipe, de modo que, uma vez que os procedimentos básicos haviam se tornado claros, o pesquisador nada fez, além de trazer os gravadores e fitas pela manhã e recolhê-los ao final do dia. Em conseqüência do espírito altamente cooperativo da equipe da UnATI, muitas horas de diálogos foram gravadas durante aproximadamente duas semanas.

Devido às limitações habituais de tempo e financiamento, a maior parte deste material não foi sequer transcrito, uma vez que excede em muito as exigências da pesquisa para a qual foi coletado, em termos de

dados, assim como do estudo apresentado aqui. Cerca de dez diálogos já foram de fato digitalizados. Seis deles foram suficientes para suprir os 3045 casos de anáfora analisados no presente estudo.

1.2. A abordagem

A lingüística de corpus não se constitui em um ramo da lingüística, no sentido que o são disciplinas como a sociolingüística ou a psicolingüística. Trata-se, na verdade, de uma metodologia de análise lingüística, e não de uma área de pesquisa. É possível, portanto, estudar fonética, sintaxe ou semântica, além dos próprios ramos acima citados, por meio de um corpus, uma vez que este seja adequado para a iniciativa de pesquisa em questão (ver Leech 1992, McEnery e Wilson 1996[7]). Uma vez que a ferramenta fundamental para a investigação de um corpus qualquer é o computador (ver Leech 1992), fica pressuposto que o corpus seja legível por máquina, e a área comum entre a lingüística de corpus e a lingüística computacional torna-se naturalmente ampla e em constante expansão.

Em relação à dicotomia chomskyana entre competência e desempenho ou seus desenvolvimentos mais recentes, a lingüística de corpus se concentra no desempenho lingüístico e não na competência. Em termos sucintos (ver Sampson 1987 e Leech 1992 para um tratamento mais completo), a metodologia de base em corpus não considera a competência como o assunto por excelência da lingüística, e, na verdade, vê a separação entre a competência mental de um falante da língua e sua manifestação no uso cotidiano como superdimensionada nas abordagens gerativistas. Por isso mesmo, a ênfase das investigações baseadas em corpus recai sobre a descrição lingüística, ao invés de sobre a busca de universais lingüísticos.

A tendência a encarar a descrição e análise da língua, conforme usada na vida real, como uma atividade menor ou de cunho "não-teóri-

[7] "Corpus linguistics is a methodology that may be used in almost any area of linguistics, but it does not truly delimit an area of linguistics itself."(McEnery e Wilson 1996)

co" – da mesma maneira como se poderia separar lepidopterologistas de colecionadores de borboletas – é inteiramente rejeitada na lingüística de corpus. A análise do corpus envolve o processamento mental da linguagem investigada, trazendo consigo, portanto, a necessidade de desenvolver modelos psicológicos do processamento. Estes modelos, porém, são desenvolvidos a partir da observação da linguagem em uso, e não à revelia desta linguagem.

Vale destacar que a maioria das aplicações do conhecimento lingüístico – seja em educação, tradução ou PLN – dizem respeito a línguas específicas e não a universais. Desta forma, as abordagens de base em corpus associam a lingüística, como ciência, à tecnologia e à verificação independente de resultados, como já é verdadeiro há séculos nas ciências naturais. Encarar resultados como um aspecto menor da investigação científica compromete gravemente qualquer iniciativa no sentido da avaliação da qualidade dos modelos e teorias construídos.

Ainda dentro da mesma vertente de análise, termos como "quantitativa" ou "empirista", quando associados a uma metodologia, parecem trazer, desde a ascensão e subseqüente predomínio da gramática gerativista, algum tipo de conotação pejorativa cuja validade é no mínimo discutível. O uso de noções como freqüência e probabilidade não exclui a análise qualitativa, nem muito menos o uso de regras e modelos, apenas fundamenta estes construtos com números, o que, em si, dificilmente pode ser encarado como metodologicamente inadequado.

Na realidade, a questão da freqüência faz invariavelmente parte da seleção de material a ser incluído em atividades didáticas relacionadas a línguas, como é fácil verificar em qualquer método de ensino de língua estrangeira. Não há, portanto, nenhuma razão para desprezar os métodos estatísticos bastante úteis que já foram desenvolvidos em outras áreas para fazer previsões quanto ao comportamento lingüístico de, por exemplo, usuários de um sistema computacional capacitado a processar linguagem natural. Na verdade, isto já é sistematicamente praticado, e seria positivo que os lingüistas participassem com maior intensidade neste florescente ramo da pesquisa científica.

Em suma, a lingüística de corpus baseia-se no desenvolvimento de gramáticas a partir da observação da linguagem em uso. Nas páginas

que se seguem, espera-se poder exemplificar como esta abordagem lida com um fenômeno reconhecidamente difícil de tratar como as relações anafóricas, e que alternativas de solução tem a oferecer para as aplicações mais comuns do conhecimento lingüístico, onde as relações anafóricas continuam a colocar dificuldades consideráveis para professores, tradutores e projetistas de sistemas.

2. O modelo de classificação

Os fenômenos anafóricos foram classificados segundo quatro atributos, a saber: o tipo de termo anafórico; o tipo de antecedente; o papel topical do antecedente; e a estratégia de processamento. Cada caso de anáfora encontrado na amostra foi classificado segundo estes atributos, de acordo com as categorias possíveis para cada um deles. O processo de desenvolvimento deste modelo de classificação será apresentado aqui como algo acabado, mas foi, na verdade, desenvolvido a partir do processo de análise dos casos de anáfora encontrados no corpus, e, assim, reiteradamente corrigido e aperfeiçoado até que tivesse sido alcançado um padrão considerado satisfatório para a análise coerente dos dados do corpus (ver Rocha 1998 para uma descrição completa).

2.1. O tipo de termo anafórico

Os termos anafóricos foram classificados em três grandes grupos, a saber:

- pronomes;
- verbos e adjuntos adverbiais;
- nomes.

No primeiro grupo, foram incluídos todos os pronomes pessoais de terceira pessoa, invariavelmente considerados como termos anafóricos, assim como: todos os pronomes possessivos substantivos; os pronomes possessivos adjetivos de terceira pessoa; todos os pronomes demonstrativos substantivos; todos os pronomes reflexivos de terceira pessoa; os

pronomes indefinidos *algum*, *nenhum*, *todo*, *muito*, *pouco*, *vários*, *tanto* e *quanto*, em todas as suas flexões, quando usados como pronomes substantivos; e os numerais empregados com função de núcleo de sintagma nominal.

Na categoria dos verbos e adjuntos adverbiais foram agrupados os verbos anafóricos, abrangendo as formas verbais de terceira pessoa sem sujeito explícito, inclusive os verbos de ligação; todas as formas verbais de verbos transitivos sem objeto explícito; todas as formas verbais de verbos de ligação sem predicativo do sujeito explícito; todos os advérbios utilizados em enunciações onde o sintagma verbal a que se relacionam não está explícito, inclusive os sinais de resposta *sim* e *não*; todos os sintagmas preposicionais utilizados em enunciações onde o sintagma verbal a que se relacionam não está explícito; e alguns outros casos raros envolvendo orações subordinadas que exigem a recuperação da principal a que estão vinculadas. Abaixo é dado um exemplo de sintagma preposicional anafórico[8].

(1)

A: mas a senhora continua com a mesma com o mesmo sintoma?

B: com o mesmo problema

No fragmento de conversação acima, o sintagma preposicional *com o mesmo problema* só pode ser interpretado se for vinculado à enunciação precedente adequadamente. Deste modo, tanto os sintagmas preposicionais quanto os advérbios anafóricos são muitas vezes respostas a perguntas ou reações a declarações feitas pelo interlocutor. O terceiro grupo dos nomes inclui sintagmas nominais anafóricos, inclusive as repetições literais, e adjetivos que qualificam núcleos omitidos de sintagmas nominais, os quais têm que ser recuperados no contexto do discurso.

Em relação aos pronomes, algumas opções foram feitas no que diz respeito ao caso oblíquo dos pronomes pessoais de terceira pessoa. Embora os pronomes átonos do caso oblíquo de primeira e segunda

[8] Todos os exemplos foram extraídos do CDC-RJ.

pessoa sejam utilizados regularmente no português falado, o mesmo não é verdade em relação às formas de terceira pessoa. Os pronomes retos são empregados com freqüência, mesmo quando se trata de um objeto direto. Estas ocorrências foram incluídas como pronomes objetos na amostra de casos de anáfora, sem qualquer distinção em relação às demais formas. Um exemplo é mostrado abaixo.

(2)
B: eu consegui matricular ele no INPS perto de casa

Nos usos de pronomes de terceira pessoa em contração com a preposição *de*, as ocorrências foram invariavelmente classificadas como casos do tipo de termo anafórico **pronome objeto**, mesmo quando a função semântica da contração é, claramente, de possessivo, como no exemplo abaixo:

(3)
A: quais são as queixas dele?

A posse, no português falado, é quase que invariavelmente denotada através destas contrações nos casos de terceira pessoa. Isto resultou em freqüências muito baixas de termos anafóricos classificados como qualquer dos dois tipos de pronomes possessivos mencionados acima, uma vez que os pronomes de primeira e segunda pessoa, onde a incidência é maior, não foram, de modo geral, incluídos na amostra, já que, na maioria dos casos, não sinalizam referência anafórica, exceto nos casos de discurso relatado.

Em relação aos verbos anafóricos, também foi necessário rever a definição estabelecida inicialmente para que ocorrências cuja classificação se mostrou problemática pudessem ser incluídas na amostra. Os verbos anafóricos foram descritos acima como uma forma verbal que exige a recuperação de elementos da sua estrutura argumental no contexto do discurso para sua interpretação semântica. Conforme assinalado anteriormente, isto requer, para fins de processamento, que a presença de argumentos essenciais dos verbos seja constantemente verificada

em relação a um padrão de estrutura argumental incorporado a um léxico pré-existente em uma máquina ou na mente de um usuário da língua.

Ao realizar a coleta de casos de anáfora no corpus, o analista utilizou um procedimento básico: verificar a presença de um sujeito em todos os sintagmas verbais, e dos objetos necessários em todos os verbos transitivos, além dos predicativos nos verbos de ligação. Sempre que um dos argumentos essenciais não era encontrado, a ocorrência era classificada como um verbo anafórico. O contexto discursivo era então analisado para identificar o antecedente e a estratégia de processamento, conforme classificação apresentada mais adiante. Um exemplo de verbo anafórico é mostrado abaixo.

(4)

A: a senhora sabe se tem algum exame de sangue da senhora? de colesterol, de glicídio?

B: 'tava ... foi a foi a a doutora pediu, né?

A: pediu?

A primeira ocorrência da forma verbal *pediu* tem um sujeito explícito, mas não há objeto direto na enunciação. Entretanto, *pedir* é um verbo transitivo que requer um objeto direto e, muitas vezes, um objeto indireto também, embora, neste caso, pareça ser desnecessário incluir este último na análise. A ocorrência é então analisada como um verbo anafórico, o que significa que o discurso foneticamente realizado não fornece os argumentos essenciais do verbo conforme esperado. O objeto direto tem que ser recuperado no turno precedente, embora o verbo da enunciação não seja *pedir*, mas sim *ter* em seu sentido existencial, o qual não requer um sujeito, mas necessita de um objeto.

A ocorrência subseqüente de *pediu* não explicita nenhum dos dois argumentos, e por isso é analisada como um caso de referência anafórica dupla por meio de um único termo anafórico, já que é preciso recuperar dois antecedentes no discurso anterior. Casos como este não são incomuns. Contudo, há sentenças do português que são analisadas, segundo Cunha (1985), como não tendo sujeito, identificadas por certos verbos e usos típicos de formas verbais, listados abaixo:

- sintagmas verbais que expressam fenômenos naturais
- o verbo *haver* quando denota existência
- os verbos *haver*, *fazer* e *ir* quando se referem a tempo transcorrido
- o verbo *ser* quando se refere a tempo

A esta lista devem ser acrescentadas as ocorrências do verbo *ter* que também denotem existência, já que são bastante comuns no português falado. A inexistência de sujeito nestas formas verbais não foi considerada um caso de anáfora. As ocorrências destes verbos em que o objeto ou o predicativo do sujeito estavam omitidos foram, porém, incluídas na amostra como casos de anáfora, já que estes argumentos são necessários à interpretação semântica. Um problema de solução mais difícil se relaciona às formas verbais que têm função de marcadores do discurso, seja em perguntas de confirmação ou em respostas a perguntas. Estas ocorrências não podem ser tratadas de maneira uniforme, uma vez que elas, em muitos casos, desempenham funções discursivas simultaneamente ao papel esperado determinado pelo sentido lexical do verbo, como no exemplo abaixo.

(5)

A: e ele, como é que ele ‘tá de saúde, Joana?
quais são as queixas dele?

B: olha saúde ele não ... não ‘tá bem, né?
ele é aposentado, mas continua trabalhando, entendeu?

O sentido literal de *entendeu* não se adequa muito claramente ao contexto, já que a enunciação precedente é declarativa e transmite informações simples, que não exigem nenhuma forma especial de entendimento. A forma verbal em questão cumpre, principalmente, a função pragmática de certificar-se da atenção do ouvinte e mantê-la focalizada no que está sendo dito. Não obstante, o verbo *entender*, nesta acepção, é um transitivo direto que exige um objeto. Nada nos verbetes de dicionários da língua portuguesa (ver Hollanda 1986) sugere a possibilidade deste tipo de ocorrência como intransitivo, e também não há qualquer comentário em Cunha (1985). Porém, estas ocorrências são muito freqüentes, e é extremamente difícil, e, em certos casos, impossível especificar o objeto direto destas formas verbais.

Uma maneira de lidar com estas ocorrências seria deixá-las também de fora da amostra, como ocorrências do verbo *entender* com sentido alterado, as quais não incluiriam um objeto direto em sua estrutura argumental. O primeiro problema com este tipo de solução é, evidentemente, que não existe nenhum precedente deste tipo de análise na literatura de referência, à diferença dos verbos sem sujeito discutidos anteriormente. O segundo problema decorre do fato de que é possível detectar pelo menos parte do sentido lexical do verbo ainda preservado. O terceiro problema deriva da existência de ocorrências muito semelhantes no corpus em que a interpretação semântica pode ser muito mais "referencial". Na realidade, parece existir um continuum de referencialidade (ver Schiffrin 1987) neste tipo de ocorrência, variando desde o sentido lexical estrito do verbo até o uso para funções exclusivamente pragmáticas, em que o sentido lexical da forma verbal é praticamente irrelevante.

No decorrer da análise dos dados do corpus, foi possível observar a rica complexidade da interação entre relações anafóricas, estrutura argumental, marcadores do discurso e perda de sentido lexical, com a contrapartida de um aumento do peso da função pragmática das formas verbais. Em muitos casos envolvendo perguntas de confirmação ou respostas com os argumentos omitidos, formas verbais dos verbos *entender* e *saber* são utilizadas com variados graus de preservação do sentido e de contrapartida em termos de reforço do papel pragmático.

Estas ocorrências apresentam complexidade ainda maior se forem consideradas também as ocorrências dos verbos de ligação *ser* e *estar*, cuja estrutura argumental exige a identificação de um sujeito e de um predicativo do sujeito. Em termos de processamento, torna-se necessário especificar quais as situações em que um determinado verbo será considerado com sua estrutura argumental padrão e quais exigem soluções em que esta estrutura é descartada em favor de uma interpretação como marcador discursivo. O problema é semelhante ao do tratamento de termos como pronomes demonstrativos que, embora tipicamente anafóricos, ocorrem como não-referenciais em colocações específicas, ainda que, no caso destes últimos, o levantamento dos ambientes que propiciam as alterações não tenha que lidar com uma complexidade tão grande de graus de referencialidade.

A solução encontrada para o problema será apresentada mais adiante, uma vez que extrapola o âmbito da classificação dos tipos de termo anafórico, envolvendo também o tipo de antecedente e a estratégia de processamento.

3.2. O tipo de antecedente

A classificação do tipo de antecedente diz respeito basicamente à dicotomia implícito/explícito, sendo que o segundo tipo predomina fortemente, pelo menos em diálogos. Abaixo há dois exemplos de anáfora. No exemplo (6), o antecedente *não sinto sede durante o dia* está explícito na enunciação anterior. A variação de pessoa não foi considerada no estudo como suficiente para que a classificação do antecedente oracional fosse classificada de maneira diferente.

(6)

A: e sede, a senhora sente muita sede durante o dia?

B: nenhuma

No exemplo (7), o antecedente *açúcar* está implícito devido à forte ligação semântica com *glicose*. Parece razoável afirmar que, em termos de processamento, a ativação de elementos próximos do campo semântico tem participação importante na identificação de referências deste tipo.

(7)

A: mas a senhora alguma vez já fez algum exame de glicose
para ver se tem algum problema

B: bom, quando eu fiz estava passando uns dez pontinhos
mas o médico falou que era também da idade
e não ia passar remédio só suspender o açúcar

Há ocorrências, porém, em que a classificação do antecedente é difícil, tais como a expressão *por isso ou por aquilo* no exemplo (8) abaixo.

(8)

B: sobre a urina?

A: é

B: não, eu nunca prestei atenção se era por isso ou por aquilo que eu comesse, entendeu

A expressão cristalizada não se refere especificamente a nenhum alimento em particular, embora seja composta por dois pronomes demonstrativos, ambos tipicamente empregados na função de termo anafórico. Tendo em vista o processamento automático de relações anafóricas, para o qual o estudo pretende contribuir, é importante incluir todas as ocorrências de palavras tipicamente anafóricas, de modo a estabelecer os padrões de ocorrência. No caso em questão, não há um antecedente *stricto sensu* para os termos anafóricos. Casos como este receberam a classificação de **não-referencial**.

É certamente relevante assinalar que estes casos não constituem relação anafórica propriamente dita, uma vez que não há antecedente a ser identificado. Por outro lado, sua inclusão permite avaliar com que freqüência palavras tipicamente utilizadas como termos anafóricos, como é o caso dos demonstrativos em questão, são empregadas em situações em que a relação anafórica não se concretiza conforme esperado, e, sobretudo, permite estudar estas situações de modo a estabelecer em que contextos ocorrem. Isto pode ser da maior importância para o processamento automático de anáforas, uma vez que um interpretador de anafóras em um sistema qualquer não persistiria em tentativas inúteis de identificar um antecedente específico.

Também é verdade que a análise destes padrões de ocorrência contribui para o esclarecimento inclusive do aspecto psicolingüístico do processamento de relações anafóricas, ao levantar a questão das expressões cristalizadas e outras formas mais complexas de processamento com base em esquemas, geralmente discutidas na literatura da lingüística de corpus sob a denominação de colocações. A noção de colocação pode ser definida como a co-ocorrência sistemática de itens lexicais, com a possibilidade de um tratamento estendido para abranger as estruturas léxico-gramáticas propostas na lingüística sistêmica. A questão voltará a ser discutida na subseção relativa às estratégias de processamento.

A análise dos dados do corpus revelou, gradativamente, que seria necessário criar uma quarta possibilidade de classificação de antecedentes para lidar com casos como o do exemplo (9) abaixo. Segundo o critério estabelecido para a determinação do caráter anafórico de um verbo qualquer, verifica-se a presença dos argumentos essenciais em forma explícita na enunciação. Caso haja omissão, fica caracterizada a anáfora. No caso da forma de terceira pessoa do singular do verbo *ser* abaixo, o sujeito está omitido e precisa ser recuperado no contexto do discurso para que a interpretação semântica se complete com êxito.

(9)

A: a senhora prefere fazer o exame num hospital
aqui no Hospital Carlos Pinto

B: não, não é preferência, eu fui fazer aí, mas teve esses problemas,

A: teve esses problemas

A análise do contexto aponta para um antecedente implícito de difícil especificação. A melhor solução está provavelmente em uma expressão genérica, tal como *o problema* ou *a questão*, que se referem àquilo que está sendo discutido ou falado no momento de maneira relativamente vaga. Contudo, é preciso que haja alguma maneira pré-definida de lidar com ocorrências como estas, já que elas não são particularmente incomuns. Foi criada a categoria do antecedente **implícito no discurso** para classificar este tipo de ocorrência. Tais situações demonstram a importância das abordagens baseadas em corpus, uma vez que a análise dos dados do corpus coloca questões autênticas que provavelmente não seriam lembradas. A confiança na própria intuição, sem o confronto com dados da língua conforme utilizada para comunicação, tem conseqüências sobre o processamento automático de linguagens naturais, tornando os sistemas construídos excessivamente frágeis diante da enorme variedade de situações possíveis da vida cotidiana.

2.3. O papel topical do antecedente

Este atributo constitui uma tentativa de incorporar a relação freqüentemente mencionada entre topicalidade e anáfora ao modelo de

classificação. Utilizando informações estatísticas simples, tais como freqüência e distribuição, juntamente com dados tais como a posição da primeira ocorrência no texto e a classificação do constituinte como sintagma nominal, foi especificado um **tópico do discurso** para cada diálogo, além de um **tópico de segmento** para cada passagem de diálogo em que se verificasse continuidade de tópico. Foram definidos também **elementos temáticos do discurso** e **elementos temáticos de segmento**, a fim de classificar entidades do discurso (no sentido utilizado em Weber 1979) importantes que não fossem tópicos.

A análise do corpus demonstrou que, em muitos casos, uma divisão em subsegmentos se faz necessária, e cada um dos subsegmentos recebeu um tópico, chamado de **tópico de subsegmento**. Algumas outras categorias foram utilizadas para casos raros de difícil classificação que não é necessário analisar aqui. Estas categorias foram utilizadas para definir um papel topical para cada antecedente detectado em todos os casos de anáfora incluídos na amostra. Esta definição de papéis topicais resulta, em última análise, numa especificação da estrutura da topicalidade de um diálogo dado.

Uma discussão completa desta estrutura da topicalidade é impossível dentro dos limites deste trabalho. Não obstante, vale destacar que estas informações desempenham papel fundamental na resolução de casos de anáfora particularmente complexos e difíceis de resolver, onde o termo anafórico está distante o suficiente de seu antecedente para que existam várias alternativas de antecedentes sintaticamente viáveis entre o termo anafórico e o antecedente correto.

2.4. A estratégia de processamento

A estratégia de processamento foi considerada como uma variável necessária para a análise das relações anafóricas, já que o tipo de termo anafórico e o tipo de antecedente não são suficientes para definir com exatidão o caminho a ser percorrido para a resolução de uma referência anafórica. Uma vez que o estudo pretende contribuir para o processamento automático de relações anafóricas em sistemas

computacionais capacitados a lidar com linguagens naturais, foi considerada essencial uma maior preocupação com definições relativas ao processamento, investigando as associações entre o termo anafórico, o antecedente e o processamento, de modo a estabelecer padrões de ocorrência que pudessem orientar um interpretador de anáforas em um sistema atuante no mundo real.

Foram estabelecidas quatro categorias abrangentes para definir as possíveis estratégias de processamento empregadas na resolução de anáforas. A primeira delas foi chamada de **processos sintáticos**, e diz respeito a resoluções de referências anafóricas baseadas em concordância e proximidade, isto é, o antecedente é o primeiro candidato adequado encontrado no discurso precedente, levando em conta gênero e número do termo anafórico empregado. Tais soluções podem ser implementadas com relativa facilidade em um sistema de computadores através de um algoritmo "ingênuo", isto é, uma seqüência de procedimentos que ignora toda a informação de natureza semântica, como o descrito em Hobbs (1986). Um exemplo deste tipo de resolução para anáfora é dado abaixo.

(10)

B: fiz a ... aquele negócio que anda na esteira

A: uhum foi o teste ergométrico, né?

B: é, fiz aquilo

As cadeias de referência foram também consideradas como um processo sintático, uma vez que o algoritmo "ingênuo" localizaria um outro termo anafórico, o qual, em princípio, já teria sido previamente analisado e resolvido. Embora as cadeias de referência apresentem problemas de solução não tão simples quanto a simples escolha do primeiro candidato adequado para o processamento, não parece excessivo pressupor que as informações de natureza sintática seriam suficientes para garantir o êxito da interpretação. Porém, no exemplo abaixo, há três ocorrências de pronome demonstrativo anafórico, sendo que a terceira faz parte de uma contração com a preposição *de*. Esta ocorrência não faz parte de uma cadeia de primeiros candidatos e seria resolvida incorretamente com uso do algoritmo "ingênuo".

(11)

B: eu tinha assim um pelágio

A: sei, aquilo que cai assim embaixo do olho

B: não, filha, é um ... é do colesterol
dá assim feito umas gordurinhas

A: aham

B: então, eu tinha demais

A: uhum

B: aí eles tiraram

A: isso é em função do colesterol?

B: o médico diz que é

A: nossa

B: do colesterol

A: e aí você fez uma uma pequena cirurgia <2syl>

B: é, ele aproveitou, tirou pele também e

A: aham

B: e tirou a <4syl>

A: mas você ficou legal disso, depois controlou o colesterol e ficou, ficou, não volta não né? não reincide não?

O pronome demonstrativo na última enunciação do fragmento acima refere-se a *pelágio*, mas o primeiro candidato adequado seria *pele*, uma vez que o demonstrativo *isso* não discrimina o gênero do referente. Mesmo que elementos de semântica lexical fossem utilizados, a expressão *ficar legal* não eliminaria a possibilidade do referente ser *pele* com base em restrições selecionais. Somente fatores de natureza discursiva podem garantir o processamento com êxito deste tipo de referência, onde é necessário ignorar um ou mais candidatos adequados mais próximos e localizar um antecedente mais distante. Este tipo de estratégia de processamento, e algumas variantes dela, foi denominada como **conhecimento discursivo**, uma vez que é preciso incluir fatores do contexto do discurso para viabilizar o processamento. As referências dêiticas também estão incluídas nesta categoria.

O terceiro tipo de estratégia de processamento diz respeito a um fenômeno já mencionado, as colocações, e recebeu o nome de **conhe-**

cimento de colocações. Imagine-se que o léxico mental possua uma lista de expressões cristalizadas cujo processamento está pré-determinado em conseqüência da experiência acumulada no uso da língua. Esta lista pode incluir combinações entre itens lexicais, traços semânticos comuns a um grupo de itens lexicais e estruturas sintáticas (ver Rocha 1998 para uma lista detalhada). O levantamento realizado no corpus, através da classificação da estratégia de processamento, permitiria, portanto, que esta mesma lista hipotética fosse construída e, potencialmente, incorporada ao interpretador de anáforas como conhecimento essencial para o êxito do processamento.

Sendo assim, a colocação *por isso ou por aquilo* estaria associada a uma resolução em que a ocorrência é **não-referencial**. Vale frisar que todos os exemplos discutidos até agora são de pronomes demonstrativos, o que demonstra a importância de incluir uma variável como a estratégia de processamento, evitando, assim, que fenômenos anafóricos tão diversos fossem agrupados sob a mesma classificação sem maiores especificações.

O quarto tipo de estratégia de processamento diz respeito ao uso de informações de natureza lexical e recebeu, portanto, o nome de **conhecimento lexical**. O exemplo (7) é um caso típico deste tipo de estratégia, onde o antecedente *açúcar* pode ser identificado devido à ativação anterior causada pela menção da *glicose*. As referências anafóricas por meio de repetição lexical foram incluídas nesta categoria. Embora esta estratégia de processamento esteja fortemente associada aos sintagmas nominais anafóricos não-pronominais, a análise dos dados do corpus revelou que o conhecimento lexical também é importante para a resolução de referências por meio de sintagmas preposicionais e advérbios anafóricos.

O processo de análise foi feito através do exame dos diálogos do CDC-RJ. Cada caso de anáfora incluído na amostra foi anotado manualmente com a classificação definida por estas quatro variáveis. Foram analisados 3045 casos de anáfora em seis diálogos do CDC-RJ. Alguns resultados da análise serão discutidos na próxima seção.

3. Os resultados do estudo

A Tabela 1 abaixo resume as freqüências dos tipos de termo anafórico na amostra coletada.

Tabela 1 – Tipos de termo anafórico

	Freqüência	Percentagem
Pronomes	530	17.4
Verbos e Adj. Adv.	1507	49.5
Nomes	1008	33.1
Total	3045	100.0

Pode-se observar, na Tabela 1, que o tipo de termo anafórico mais freqüente na amostra de português falado analisada é o verbo[9]. Uma vez que este estudo presume ter conseguido coletar uma amostra autêntica de diálogos em português, parece seguro afirmar que o verbo anafórico é a forma predominante de referência anafórica da língua. Vale frisar que uma análise de diálogos em inglês segundo a mesma classificação registrou uma pequena variação na percentagem de nomes anafóricos, e percentagens diametralmente opostas de pronomes e verbos (ver Rocha 1998). Isto reforça a observação, feita anteriormente, de que o sistema de referência, em inglês, baseia-se fundamentalmente em sinais explícitos da necessidade de recuperar um elemento do discurso para a interpretação semântica, enquanto a língua portuguesa utiliza a estrutura argumental dos verbos para detectar argumentos essenciais omitidos que sinalizam a referência anafórica. A Tabela 2 abaixo mostra os números relativos ao tipo de antecedente.

[9] Embora os adjuntos adverbiais anafóricos estejam agregados ao total, os verbos constituem 81,55% (1229) dos casos.

Tabela 2 – Tipo de antecedente

	Freqüência	Percentagem
Explícito	2265	74.4
Implícito	382	12.5
Implícito no discurso	49	1.6
Não-referencial	349	11.5
Total	3045	100.0

A predominância dos antecedentes explícitos é indiscutível. Isto demonstra que a resolução de referências anafóricas depende fundamentalmente de um processamento adequado dos elementos diretamente introduzidos no discurso, e não de inferências a partir de informações no discurso, em busca de antecedentes implícitos. A percentagem relativamente alta de ocorrências não-referenciais é provavelmente conseqüência do grande número de marcadores do discurso com funções estritamente pragmáticas. A Tabela 3 resume os resultados da amostragem em termos de estratégia de processamento.

Tabela 3 – Estratégias de processamento

	Freqüência	Percentagem
Proc. Sintáticos	1005	33.0
Conh. Colocações	656	21.5
Conh. Discursivo	446	14.6
Conh. Lexical	938	30.8
Total	3045	100.0

A importância do conhecimento de colocações fica bastante evidenciada nos números relativos à estratégia de processamento, ainda que os processos sintáticos predominem como forma de resolver as re-

ferências anafóricas que integram a amostra. O conhecimento lexical, onde a repetição lexical é a forma predominante de sinalizar o caminho para a identificação do antecedente, também atinge um percentual bastante alto de estratégias de processamento. O conhecimento discursivo, onde se concentram os casos difíceis, cuja solução exige um processamento conjunto de diversos elementos do contexto discursivo, concentra o menor número de casos, mas, ainda assim, suficientes para inviabilizar a interpretação semântica global de uma instância de discurso, caso não se obtenha êxito em sua resolução.

A Tabela 4 apresenta o cruzamento dos números relativos ao tipo de termo anafórico com o tipo de antecedente. Os antecedentes implícitos no discurso foram agrupados aos implícitos em geral, de modo a facilitar os testes de chi-quadrado e associação descritos em seguida (ver Walsh 1990 para uma discussão dos problemas gerados por células com valores muito baixos para os testes de chi-quadrado).

Tabela 4 – Tipo de termo anafórico por tipo de antecedente

	Explícito	Implícito	Não-refer.	Total
Pronomes	449	57	24	530
Verbos Adj. Adv.	1057	128	322	1507
Nomes	759	246	3	1008
Total	2265	431	349	3045

As células com os números mais interessantes se concentram na coluna dos não-referenciais. Em forte contraste com resultados do estudo realizado com diálogos em língua inglesa em Rocha (1998), utilizando a mesma classificação, o número de pronomes não-referenciais é muito baixo, enquanto o de verbos e adjuntos adverbiais é muito alto. Isto parece revelar que os verbos anafóricos, particularmente os verbos de ligação anafóricos, são mais freqüentes quando suas formas equivalentes em língua inglesa são pronomes neutros ao invés das formas usadas para referentes humanos, já que estes nunca são não-referenciais.

Fica assim delineado, mais uma vez, o contraste entre dois sistemas de referência que utilizam, respectivamente, pronomes e verbos como suas formas de termo anafórico por excelência.

Os testes de chi-quadrado para estas duas variáveis demonstraram alta significância de sua relação, do ponto de vista estatístico, uma vez que a possibilidade de sua relação se dever ao acaso é menor do que $p < 0.00005$. No entanto, a medida de associação, usando o *tau* de Goodman e Kruskal, revelou um nível de associação baixo, com uma redução proporcional do erro de 0.4. Isto significa que a probabilidade de prever com acerto o tipo de antecedente, uma vez que se saiba o tipo de termo anafórico, aumenta apenas em 4%, se comparada ao acaso. A Tabela 5 apresenta o cruzamento dos números do tipo de termo anafórico com as estratégias de processamento.

Tabela 5 – Tipo de termo anafórico por estratégia de processamento

	Proc. Sintáticos	Conh. Lexical	Conh. Discurs.	Conh. Coloc.	Total
Pronomes	307	23	145	55	530
Vbs. e Adj.Adv.	666	32	246	563	1507
Nomes	32	883	55	38	1008
Total	1005	938	446	656	3045

Apesar da predominância dos processos sintáticos por pequena margem, pode ser observado que as resoluções baseadas em conhecimento de colocações atingem um nível muito alto. Isto se deve à influência de um grande número de verbos de ligação anafóricos não-referenciais, cuja estratégia de processamento é fundamentalmente o conhecimento de colocações, conforme discutido anteriormente. As ocorrências de resoluções com base em conhecimento lexical se concentram nos nomes, o que seria de se esperar. As ocorrências de resoluções com base em conhecimento discursivo também predominam entre

os verbos, mas, considerando a quantidade muito maior de ocorrências de verbos anafóricos do que de pronomes, o nível das ocorrências de pronomes com este tipo de resolução é alto, uma vez que atinge cerca de trinta por cento dos casos, enquanto os pronomes chegam apenas a 17.4% dos casos no cômputo geral.

Os testes de chi-quadrado com estas duas variáveis tiveram resultados semelhantes aos feitos com as duas variáveis da Tabela 4. Contudo, o nível de associação medido pelo *tau* de Goodman e Kruskal chega a 0.36, o que significa que a possibilidade de prever a estratégia de processamento com acerto aumenta em 36% quando se conhece o tipo de termo anafórico. Isto significa que, dado um diálogo onde são conhecidos os termos anafóricos, as estratégias de processamento a serem utilizadas têm boa possibilidade de serem previstas com precisão, sobretudo se informações contextuais coletadas através da observação dos tipos de termo anafórico específicos puderem ser utilizadas.

Sabe-se que, atualmente, já existem programas capazes de atribuir classes gramaticais às palavras de um texto legível por máquina automaticamente. Estes programas são geralmente chamados de etiquetadores de estruturas morfossintáticas. Embora o nível de exatidão que obtêm em transcrições de diálogos tenha que ser verificado, a perspectiva de obter informações efetivas sobre estratégias de processamento de termos anafóricos a partir de sua classe gramatical parece real, ainda que este estudo não seja em absoluto suficiente para conclusões mais definitivas. Finalmente, a Tabela 6 mostra os resultados do cruzamento dos dados entre as estratégias de processamento e os tipos de antecedente.

Tabela 6 – Estratégias de processamento por tipos de antecedente

	Explícito	Implícito	Não-refer.	Total
Proc. Sintáticos	963	42	0	1005
Conh. Lexical	714	224	0	938
Conh. Discurs	336	110	0	446
Conh. Coloc.	252	55	349	656
Total	2265	431	349	3045

Fica claro aqui que a associação entre colocações e não-referenciais é absoluta. Deste modo, os termos aparentemente anafóricos que na realidade não se referem a um antecedente identificável podem ser detectados através de uma lista de colocações, ainda que as formas verbais discutidas anteriormente possam aparecer tanto como colocações, ou seja, com seu sentido alterado, quanto como ocorrências com seu sentido lexical esperado.

Os testes de chi-quadrado realizados com estas duas variáveis obtiveram significância em nível idêntico aos registrados nas duas outras tabulações cruzadas. O nível de associação registrado para estas duas variáveis foi razoavelmente alto, chegando a 0.22, o que sinaliza um aumento de 22% em relação ao acaso na possibilidade de prever o tipo de antecedente, uma vez que a estratégia de processamento seja conhecida. Parece razoável concluir destas medições de associação que a estratégia de processamento age como um elo de ligação entre as duas outras variáveis, uma vez que apresenta nível de associação alto com o tipo de termo anafórico, como a variável dependente, e também nível de associação alto com o tipo de antecedente, desta vez com este último como a variável dependente.

Estes números não apenas justificam a inclusão da variável na classificação de termos anafóricos, mas parecem sinalizar que é possível aumentar a eficiência de interpretadores de anáforas em sistemas de PLN através desta abordagem, ainda que o estudo tenha limitações óbvias de dimensão e abrangência. Quando a interação entre as quatro variáveis foi medida através da análise loglinear, somente foram consideradas estatisticamente significativas as interações entre três variáveis nas quais uma delas era a estratégia de processamento, o que torna ainda mais clara a tendência detectada através das medidas de associação.

5. Conclusão

A classificação criada para a análise das relações anafóricas no português falado parece ter possibilidades de tornar-se uma contribuição real para a compreensão deste complexo fenômeno do discurso.

Espera-se, igualmente, que o estudo possa representar um primeiro passo para uma maior eficiência na resolução de referências anafóricas em sistemas de processamento de linguagens naturais. Pressupondo um sistema com a capacidade de realizar a rotulação de estruturas morfossintáticas em tempo real, à medida em que um usuário fala, por exemplo, poderia ser possível aprimorar a resolução de anáforas em interfaces em linguagens naturais para acesso a banco de dados, uma das aplicações mais desejáveis do processamento de linguagens naturais em computadores. A utilização destas mesmas capacidades em sistemas de tradução de máquina e aprendizado de línguas com ajuda de computadores também apresenta perspectivas atraentes.

No processo geral de investigação científica relacionada à linguagem, as abordagens baseadas em corpus desempenham um papel fundamental na renovação da pesquisa lingüística e em suas diversas aplicações, introduzindo um elemento de realidade da língua que vinha sendo deixado de lado e mesmo condenado. Não se pretende, com isso, menosprezar abordagens de natureza mais formal ou abstrata, mas lembrar que a análise do uso da língua no contexto da vida real é, pelo menos, tão importante quanto estas últimas.

Referências Bibliográficas

Aarts, J. (1991) Intuition-based and observation-based grammars. In: K. Aijmer e B. Altenberg (org.) *English corpus linguistics: Studies in honour of Jan Svartvik*. Harlow: Longman.

Biber, D. (1992) Using computer-based corpora to analyse the referential strategies of spoken and written texts. In: Jan Svartvik (org.) *Directions in corpus linguistics*. Berlim: Mouton de Gruyter:215-252.

Carter, D. (1987) *Interpreting anaphora in natural language texts*. Bognor Regis: Ellis Horwood.

Bosch, P. (1983) *Agreement and anaphora*. Nova York:Academic Press.

Cunha, C. & Cintra, L. (1985) *Nova gramática do português contemporâneo*. Rio de Janeiro: Nova Fronteira.

Fox, B. (1987) *Discourse structure and anaphora*. Cambridge: CUP.

FRANCIS, N. (1992) Language corpora B.C. In: Jan Svartvik (org.) *Directions in corpus linguistics*. Berlim:Mouton de Gruyter: 215-252.

HALLIDAY, M.A.K. e HASAN, R. (1976) *Cohesion in English*. Londres:Longman.

HARRIS, C. (1992) Connectionism and cognitive linguistics. In: Noel Sharkey (org.) *Connectionist natural language processing*. Oxford: Intellect.

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

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

HOBBS, J. (1986) Resolving pronoun references. In: B.L. Webber; B. Grosz e K. Sparck-Jones (org.) *Readings in natural language processing*. Palo Alto: Morgan Kaufmann.

HOLLANDA, A. (1986) *Novo dicionário da língua portuguesa*. Rio de Janeiro:Nova Fronteira.

KOCH, I.V. & MARCUSCHI, L.A. (1998) Processos de referenciação na produção discursiva. *D.E.L.T.A*, **14** especial. São Paulo: EDUC:169-190.

LEECH, G. (1992) Corpora and theories of linguistic performance. In: Jan Svartvik (org.) *Directions in corpus linguistics*. Berlim: Mouton de Gruyter:105-22.

MC ENERY, T. & WILSON, A. (1996) *Corpus linguistics*. Edinburgo: Edinburgh University Press.

QUIRK, R.; GREENBAUM, S.; SVARTVIK, J. e LEECH, G. (1985) *A comprehensive grammar of the English language*. Londres: Longman.

ROCHA, M. (1998) *A corpus-based study of anaphora in dialogues in English and Portuguese*. Tese de doutorado. Falmer: University of Sussex.

SAMPSON, G. (1987) Probabilistic models of analysis. In: R. Garside, G. Leech and G.Sampson (orgs.) *The computational analysis of English*. Harlow: Longman.

SCHIFFRIN, D. (1987) *Discourse markers*. Londres: Cambridge University Press.

WALSH, A. (1990) *Statistics for the social sciences*. Nova York: Harper e Row.

WEBBER, B.L. (1979) *A formal approach to discourse anaphora*. Nova York: Garland.