3
Pressupostos Teóricos e Metodologia

Nesse capítulo, apresentamos os pressupostos teóricos e metodológicos que orientam a nossa reflexão em relação aos pontos principais desta pesquisa: o potencial de significação e utilização apresentado pelo diminutivo sufixal no português do Brasil e a efetuação de uma pesquisa baseada em um corpus de língua oral.

Como já foi sinalizado, pretendemos questionar o tratamento tradicional do diminutivo como um mecanismo morfológico essencialmente ligado à noção semântica de tamanho reduzido do referente. Para a corroboração (ou não) da nossa hipótese de maior relevância das funções pragmáticas, buscamos evidência em dados reais do uso que os falantes nativos fazem do diminutivo. Para tanto, precisamos estabelecer critérios segundo os quais uma dada ocorrência diminutiva pode ser analisada como pertencendo a um ou outro dos dois seguintes pólos: no primeiro pólo, temos os casos veiculando a noção semântica de tamanho reduzido das dimensões do referente, e no segundo pólo, temos os casos do diminutivo veiculando funções de ordem pragmática. Na primeira parte desse capítulo, procura-se delinear como o significado do diminutivo pode ser descrito em termos da problemática dicotomia semântica >< pragmática. Para tanto, pretendemos nos basear em propostas trazidas pela Lingüística Cognitiva. Na segunda parte do capítulo, estabeleceremos os pressupostos teórico-metodológicos que orientam a parte empírica do nosso estudo.

3.1
Sobre a descrição do significado

Como vimos nos capítulos anteriores, a questão de descrição do potencial de significação do diminutivo é bastante complexa: por um lado, encontram-se descrições com visão bastante restrita em relação à capacidade de significação, por outro lado, descrições que mesclam conceitos da semântica e da pragmática nas suas tentativas de dar conta da variedade dos sentidos possíveis.

Como foi observado no capítulo 2, muitas vezes os autores usam tanto os conceitos semânticos quanto pragmáticos para se referirem aos valores do

diminutivo, mas não especificam o que entendem por esses conceitos, usando expressões como “os valores semântico-pragmáticos do diminutivo”. Em outros casos, procura-se estabelecer uma distinção mais explícita entre os variados valores e empregos do diminutivo, porém, raramente explicitando os critérios usados para a distinção desses valores. Nesse trabalho, dado o nosso objetivo de explicitar a importância da dimensão pragmática do diminutivo, procuramos definir melhor em que termos os diferentes aspectos do significado podem ser descritos. Quando se trabalha com a descrição de significado, é necessário tomar uma posição sobre a difícil questão de delimitação da fronteira entre as diferentes dimensões do significado. As questões como em que consiste o significado, qual é a relação entre as palavras e o mundo, entre as palavras e o enunciador, ou entre o enunciador e o mundo, são tópicos de debates filosóficos e alvos de polêmica há já pelo menos dois milênios. Obviamente, essa pesquisa não objetiva trazer nenhum tipo de solução para essas questões, que podem ser abordadas de tantas perspectivas quanto haja observadores. No entanto, acreditamos que qualquer lingüista que pretende trabalhar com as questões ligadas ao significado, precisa refletir sobre a questão de construção de significado e definir uma posição sobre essa questão, para que a elaboração de uma proposta de descrição não seja prejudicada pelo uso de conceitos não compatíveis. Na nossa proposta de descrição do potencial de significação do diminutivo, fundamentaremos o nosso posicionamento teórico e metodológico nas contribuições da Lingüística Cognitiva para o tratamento dos fenômenos ligados ao ato de significar.

3.1.1
Significado na Lingüística Cognitiva

A Lingüística Cognitiva se desenvolveu nos anos 1980 sob a liderança dos autores norte-americanos George Lakoff (Lakoff & Johnson, 1980; Lakoff, 1987), Ronald Langacker (1987, 1990, 1991) e Leonard Talmy (1983, 1988), entre outros, e tem como pressuposto básico a idéia de que não é possível tratar da língua separadamente da cognição humana. Em vez de um módulo separado, a linguagem é concebida como parte integrante da cognição. Como observa Salomão (1990, p.2-3), o termo “Lingüística Cognitiva” é usado, na verdade, para designar uma empreitada coletiva, compreendendo vários trabalhos

independentes. No entanto, todos eles têm como ponto de partida a afirmação que "a língua é sistematicamente baseada na cognição humana" (Sweetser, 1990, p.1), e deve ser descrita e explicada como conseqüência dessa premissa. Desta maneira, a Lingüística Cognitiva se apresenta como alternativa para o paradigma clássico, chamado "objetivista" pelos cognitivistas, que pressupõe a existência de um mundo que tem uma ordem e estrutura independentemente de quem o percebe. Como foi observado por Novais (2002), numa perspectiva cognitiva, a linguagem fundamenta-se em processos cognitivos e interacionais e deve ser estudada "no seu uso e no contexto de conceptualização, da categorização, do processamento mental e da experiência individual, social e cultural" (ibid., p.10). O significado lingüístico não pode ser dissociado do conhecimento do mundo, dado que a linguagem é um meio de interpretar e construir esse mesmo mundo. Desta forma, trata-se de uma corrente de pensamento ao mesmo tempo racionalista e empiricista. Como observa Sommer (2004, p.13), a Lingüística Cognitiva concebe a linguagem como "uma entre as outras faculdades cognitivas do ser humano, mas que é também altamente motivada e não-autônoma, pois depende do contexto, dos participantes e da socialização para construir sentido".

Por associar-se à tradição funcionalista, a Lingüística Cognitiva concebe a língua como sendo moldurada pelas funções que ela serve e por uma variedade de fatores relacionados. Numa oposição funcionalismo – formalismo, o primeiro concebe o entendimento desses fatores como um pré-requisito para a caracterização das estruturas lingüísticas enquanto o segundo concebe estes como subseqüentes de tal caracterização. Para Langacker (1999, p.14) a linguagem serve, por um lado, à **função semiológica** que permite que as conceptualizações sejam simbolizadas por meio de sons e gestos, e, por outro, à multifacetada **função interativa** que envolve comunicação, manipulação, expressividade e comunhão social. Essas duas funções não são distintas nem independentes, mas devem ser entendidas como os dois lados da mesma moeda. Como observam Evans & Green (2006, p.6), a língua nos oferece um sistema elaborado de **codificação** e **transmissão** de idéias. A codificação e a transmissão são possibilitadas pela capacidade humana de simbolizar. A língua transmite pensamentos através de símbolos, que, por sua vez, são constituídos da forma, falada ou escrita, e do significado convencionalmente associado com ela. Vale observar, no entanto, que a estipulação de um par "forma-significado" não

corresponde, nas abordagens cognitivas, a conceber essa relação como uma relação canônica e linear, uma concepção bastante freqüente na teorização lingüística e que pode ser ilustrada pela celebre metáfora de conceitos concebidos como um trem de mercadorias, com os vagões, ou seja, significantes, transportando os significados como cargas fixas e estáveis. Pelo contrário, a perspectiva cognitiva rejeita uma visão da linguagem que pressupõe um núcleo semântico estável para o significado de unidades lingüísticas, e defende que o significado de uma palavra não pode ser entendido independentemente do vasto repositório do conhecimento enciclopédico associado. Como ressaltam Evans & Green (ibid., p.214), em uma perspectiva cognitiva, em vez de recipientes (*containers*) contendo significado, as palavras são entendidas como pistas (*prompts*) para a construção do significado. O signo lingüístico não carrega o significado, mas o ativa. Logo, a decisão de incluir ou de excluir determinados tipos de informação do "núcleo" semântico da palavra é necessariamente arbitrária. A natureza do significado das palavras é proteiforme (*protean*): o significado associado com uma dada palavra tem a tendência de variar dependendo do exato contexto de uso.

Um dos aspectos centrais na teorização no âmbito da Lingüística Cognitiva é o entendimento de mecanismos metafóricos e metonímicos como fundamentalmente infiltrados na linguagem assim como em nosso pensamento e ação de forma geral. Esta posição está em forte oposição com a visão tradicional sobre a metáfora como um recurso retórico e de caráter ornamental. Autores como Lakoff & Johnson (1980) foram especialmente influentes no deslocamento da metáfora da posição periférica tradicionalmente atribuída a ela para uma posição decisivamente central na descrição lingüística.

Diferentes termos e conceitos correlatos são usados em Linguística Cognitiva para realçar determinados aspectos das estruturas de conhecimento que estão na base das expressões linguísticas. Entre esses conceitos, Silva (1997) enumera as noções de 'domínio' (Langacker, 1987), 'modelo cognitivo' (Lakoff, 1987) e 'frame' (Fillmore, 1975, 1982, 1985). Croft & Cruse (2004, p.16) adicionam a essa lista ainda a noção de 'base' de Langacker (1987) e observam que todos esses termos remetem ao mesmo arcabouço teórico. Neste trabalho, faremos uso principalmente da noção de 'enquadre' (frame) de Fillmore e da noção vinculada de 'perfilamento' de Langacker (1987). Ou seja, o significado de

uma determinada palavra ou expressão é entendido em termos de estruturas ou esquemas de conhecimento, emergindo das nossas experiências quotidianas. A idéia básica nessa teoria é que não é possível entender o significado de uma determinada palavra ou expressão sem acessar todo o conhecimento relacionado a ela. Determinados conceitos relacionam-se em nossas mentes porque eles estão relacionados na nossa experiência com determinadas situações, cenários ou instituições sociais. Um exemplo clássico é o modelo cognitivo do restaurante: para que possamos atribuir significado a palavras como *garçom*, *conta*, *cliente*, etc., é necessário conhecer toda a teia de relações correspondente a esta instituição. Seria impossível entender o significado de uma dessas palavras sem saber nada sobre as situações de interação que acontecem em restaurantes. O conhecimento semântico das teias de relações liga diretamente o significado lexical ao conhecimento enciclopédico: uma determinada palavra ativa o conjunto inteiro de noções de conhecimento do mundo relacionadas àquele conceito que a palavra em questão ativa.

Essas teias de relações não são estáticas, mas se ajustam a novas situações. Um dos méritos do conceito de enquadres ou modelos cognitivos é o fato de possibilitar a atribuição de significado a palavras independentemente da questão dos referentes. Além disso, o conceito possibilita a identificação de uma palavra ou expressão tanto dentro como fora de contexto: quando usadas fora de contexto, as palavras ativam as informações mais freqüentemente relacionadas a elas, e, quando contextualizadas, elas ativam as teias de relações mais diretamente envolvidas com a situação em questão, fazendo com que o sentido adequado à palavra em questão seja mais diretamente identificado (Coscarelli, 2003, p.4).

Um conceito intimamente relacionado a esta noção é o de 'perfil' (Langacker, 1987). *Grosso modo*, o mecanismo de perfilamento[1] corresponde a uma habilidade cognitiva humana bem básica: a capacidade de focar a atenção de um aspecto para outro dentro de um esquema. Como foi observado por Silva (1997), o construto descritivo de *perfil/base* (ou *figura/base*, na terminologia de Talmy 1978), que corresponde, de certa forma, à oposição da psicologia gestaltista *figura/fundo* (embora estes conceitos se refiram mais propriamente a

[1] O conceito de perfilamento é comparável à "metáfora do holofote" de Geeraerts (1993): "o sentido é o espaço semântico iluminado pelo holofote que, a medida que se desloca, foca diferentes espaços" (*apud* Novais 2002, p.13).

relações perceptivas, e aqueles a relações conceptuais), é amplamente usado tanto na descrição da semântica das palavras como das construções gramaticais. As palavras e expressões não só ativam conceitos individuais, mas também especificam a *perspectiva* a partir da qual algo é observado. Por exemplo, numa transação comercial, o verbo "vender" focaliza um ponto de vista da transação, enquanto o verbo "comprar" perfila o ponto de vista oposto. O modelo do restaurante, já mencionado, serve como exemplo clássico para ilustrar como o mecanismo de metonímia pode interagir com o mecanismo de perfilamento na atribuição de significado a uma expressão. Diferentemente da metáfora, que opera entre dois domínios conceituais, o mecanismo de metonímia opera dentro de um só domínio. Para Croft (1993, p.354), a metonímia funciona "salientando um aspecto no perfil de um conceito" num modelo cognitivo ou domínio maior. No famoso exemplo de Lakoff & Johnson (1980, p.35), citado por inúmeros estudiosos, "*O sanduíche de presunto está esperando a sua conta*", temos um caso de metonímia "pedido por cliente". Ou seja, o componente "pedido", em vez do componente "cliente", é perfilado, este primeiro representando o último.

Enquanto originariamente aplicados a lexemas, a semântica de enquadres pode ser aplicada também a construções gramaticais. E, como veremos mais adiante, o mecanismo se manifesta útil também na nossa tarefa de descrição de aspectos semânticos e pragmáticos de formações diminutivas.

Outro aspecto da abordagem cognitivista que evita os problemas das teorias semânticas objetivistas, é a noção de *protótipo* (Rosch, 1978). Numa abordagem cognitivista à categorização (Lakoff, 1987), em vez de "membros" ou "não-membros", as categorias consistem de membros "mais centrais" (prototípicos) e "mais periféricos" (não-prototípicos). Seguindo Wittgenstein (Investigações Filosóficas), Lakoff observou que alguns membros de uma determinada categoria podem até não ter nenhum traço em comum: nesses casos, a conexão entre os membros se dá em termos de semelhanças de família. A noção de prototipicidade também será útil em nosso trabalho, justificando sentidos diferentes e até distantes num único elemento morfológico.

Podemos ver que uma perspectiva cognitiva nos oferece um caminho alternativo entre as diversas possibilidades de tratar da questão "o que significa significar". Por um lado, temos as abordagens de semântica formal, que defendem a circunscrição do significado ao domínio literal, não enciclopédico e não

pragmático e definem o significado em termos de referência e verdade. Por outro lado, temos as abordagens radicalmente pragmáticas que não aceitam nenhuma possibilidade de uma palavra ou expressão significar independentemente do contexto: nessa visão, a semântica "não existe", porque o significado não pode ser definido de maneira *apriorística*, mas apenas na realização da fala. Entre esses dois extremos, temos inumeráveis posições intermediárias. Entre a variedade de propostas teóricas, consideramos particularmente bem justificada a alternativa trazida pelas abordagens cognitivas. Uma perspectiva cognitiva torna possível a inclusão dos aspectos semânticos *e* pragmáticos na análise do significado, dado que reconhece a possibilidade de existência do conteúdo semântico associado a elementos da língua, mas, ao mesmo tempo, salienta que os seus significados não podem ser inteiramente determinados isoladamente do contexto. Em contraste às abordagens formais e objetivistas, o método cognitivo inclui na sua concepção do significado lingüístico o não literal, o enciclopédico e o pragmático. E, em contraste às abordagens radicalmente pragmáticas, o método cognitivo reconhece a existência de propriedades semânticas associadas a entidades de linguagem.

Em relação à dicotomia "semântica versus pragmática" que permeia a teorização e a análise do significado lingüístico, as abordagens cognitivas trazem, então, alternativas interessantes. Como vimos já no capítulo 2., esta dicotomia se manifesta praticamente em todo estudo que envolve fenômenos ligados ao significado, porém raramente vem tratada de forma explícita. Para Evans & Green (2006, p.211), a separação entre esses dois campos de análise deriva de razões por um lado históricas, e, por outro lado, teóricas. O estudo de semântica surgiu já com os filósofos Gregos, mas foi reconhecido como uma sub-disciplina da lingüística apenas no século XIX. Ainda no século XX, Bloomfield descreveu o estudo semântico como "o ponto fraco dos estudos lingüísticos" ([1933], 1976, p.141). Algumas décadas depois, a abordagem mentalista, liderada por Chomsky, levantou um novo interesse para o significado lingüístico como parte da competência do falante nativo, mas, como observam Evans & Green (2006, p.213), devido à influência da tradição filosófica, este interesse resultou em modelos formais enfatizando apenas aqueles aspectos do significado que podem ser facilmente modelados dentro do paradigma de condições de verdade. No entanto, nos anos 1950 e 1960, filósofos como Austin (1962) e Grice (1975), argumentando que um modelo que se baseia em condições de verdade traz

limitações artificiais ao estudo de linguagem, começaram a focar a sua atenção nos princípios que governam o *uso* da linguagem em contextos interacionais. Assim, a disciplina de Pragmática surgiu desde o início como uma abordagem bastante independente, e, até hoje, continua freqüentemente sendo vista como periférica em relação à lingüística "propriamente dita", que se concentra em modelar o *conhecimento* em vez do *uso*, e a *competência* em vez do *desempenho*. No entanto, dentro do panorama dos estudos contemporâneos da linguagem, que levam, cada vez mais, em consideração os aspectos relativos ao uso que os falantes de fato fazem da língua, a posição periférica da pragmática, assim como o seu tratamento separadamente da semântica, vem sendo questionada. Muitos dos lingüistas de hoje concordam sobre a dificuldade de impor uma separação nítida entre semântica e pragmática: o resultado acaba sempre sendo bastante artificial e dificilmente resiste a uma observação crítica. Como sabemos, o contexto de uso é muitas vezes fundamental para a determinação do significado associado a um enunciado, e há fenômenos lingüísticos que não podem ser explicados em termos puramente semânticos isoladamente, como é o caso, por exemplo, das expressões dêiticas (p.ex. *aqui*, *hoje*, *ele*, *levar*, *trazer*). Enquanto essas expressões possuem um certo teor de conteúdo semântico, os seus sentidos são impossíveis de serem determinados fora do seu contexto de uso.

Dentro da visão cognitivista da linguagem, a posição periférica da pragmática é rejeitada desde o início: no paradigma cognitivista, a dimensão pragmática é considerada como um aspecto fundamental para a descrição lingüística, uma vez que a estrutura lingüística não pode ser separada do uso. A solução da Lingüística Cognitiva para o problema de tratamento dos fenômenos lingüísticos em termos semânticos versus pragmáticos consiste em conceber esses dois tipos de conhecimento como constituindo um *continuum*. Para uma visão cognitivista, a dicotomia entre a semântica e a pragmática seria uma distinção arbitrária, dado que não há distinções baseadas em princípios entre essas dimensões: as aparentes distinções entre elas seriam, na verdade, uma questão de escala. Não há princípios ou critérios suficientemente sólidos para distinguir o conhecimento lingüístico do conhecimento do mundo, assim como o conhecimento semântico não pode ser separado do conhecimento pragmático. Enquanto esta abordagem reconhece a possibilidade de um significado convencional associado a uma dada palavra ou expressão, esses, no fundo, são

abstrações feitas a partir de um leque de contextos de uso associados com estas entidades lingüísticas. Como observam Evans & Green (2006, p.356), o significado de palavras envolve uma interação complexa entre polissemia, contexto e conhecimento enciclopédico.

Kuiri (2004, p.4) assinala que o fato de o significado de uma determinada unidade lingüística (palavra ou frase) permanecer "em aberto" até a situação atual de uso, não equivale à impossibilidade de observar ou estudar o significado fora do contexto de uso: nesses casos, os signos são analisados com base nos significados acumulados a partir dos seus usos até o momento de observação. No entanto, esse tipo de análise não revela como o signo é de fato usado, dado que os usuários podem aproveitar do signo de maneira que lhes sirva melhor – é claro, tendo em mente que as maneiras de uso não são independentes do significado convencional: este, de alguma forma, estabelece os limites para o uso do signo. No entanto, o significado convencional é tão indeterminado que a partir dele não é possível prever com precisão os possíveis usos que o signo pode vir a ter. Nesse contexto é importante lembrar, como veremos mais adiante na nossa análise de dados, que a investigação de um significado que se baseia em informação oferecida pelo contexto também é sempre uma interpretação. E essa interpretação nunca pode ser tomada como "a única verdadeira", mas, no melhor das hipóteses, apenas a melhor possível a partir da informação disponibilizada pela situação do uso. Em resumo: o significado é dinâmico, em constante movimento.

A partir do que foi apresentado acima, definimos o nosso entendimento sobre o significado lingüístico da seguinte maneira: As palavras existem como unidades básicas significativas que usamos para construir os nossos enunciados. No entanto, as palavras não carregam significados estáveis, mas funcionam como pistas para a construção do significado. Os significados não são imanentes, mas construídos (e negociados) pelos falantes (e ouvintes). Dada a dificuldade de tratar do significado *in vitro*, ou seja, fora do seu contexto de uso, a dimensão pragmática deve ser considerada como parte do significado lingüístico, e não como algo periférico e extra-lingüístico. A relação da dimensão pragmática com a dimensão semântica do significado se dá em termos de um continuum: é impossível estabelecer uma fronteira fixa entre essas duas dimensões. No entanto, aceitando a existência dessas duas dimensões, é necessário, para os fins analíticos, definir que tipos de fenômenos podem ser considerados como predominantemente

alocáveis à dimensão semântica e que tipos de fenômenos seriam atribuídos à dimensão pragmática no continuum entre os dois pólos. No caso da descrição do diminutivo, nós propomos essa divisão nos seguintes termos: O pólo semântico englobaria os casos do diminutivo veiculando a noção de redução das dimensões concretas da entidade referida. O pólo pragmático englobaria os casos do diminutivo sinalizando subjetividade e expressividade do falante, assim como os seus usos estratégicos interacionais.

Na prática da análise do sentido de uma dada ocorrência, é muitas vezes difícil distinguir entre os pólos semântico e pragmático, até porque muitas vezes estes são sobrepostos. Nessa tarefa, entre os instrumentos disponibilizados pela abordagem cognitivista, o mecanismo de perfilamento (Langacker, 1987), vem a ser especialmente útil. Normalmente, se diz que palavras e construções são *polissêmicas* e formam redes complexas de significados: como observa, entre outros, Company Company (2002, p.42), em uma determinada situação comunicativa, certos traços semânticos são enfatizados, criando a *figura*, enquanto os outros ficam fora do enfoque, constituindo o *fundo*. A nosso ver, essa análise pode ser estendida para incluir traços pragmáticos: certas situações comunicativas perfilam certos traços pragmáticos. Na medida em que avançarmos na apresentação dos nossos dados, veremos que esse é o caso do diminutivo, e sobretudo saliente no *diminutivum puerile* (introduzido no capítulo 4.3.5.). Veremos que uma dada formação diminutiva pode ter o significado de “tamanho reduzido”, como, por exemplo, em *sandalinha*: dado que se trata de um contexto infantil, de fato podemos presumir que a sandália em questão seja de tamanho pequeno. No entanto, o contexto discursivo do *diminutivum puerile* faz com que os traços pragmáticos de carinho, afeto etc. entrem para o primeiro plano. Através do mecanismo de perfilamento, são esses valores que se tornam mais relevantes em determinados modelos de situações em contextos infantis.

É a partir dessas considerações teóricas que vamos conduzir a nossa análise dos significados e funções apresentados pelo diminutivo *-inho* no nosso corpus. Na próxima parte desse capítulo, apresentaremos as linhas metodológicas segundo as quais a nossa análise de dados será efetuada.

3.2
Sobre a análise de corpus

Nessa parte do capítulo, apresentamos o arcabouço metodológico que orienta a parte empírica da nossa pesquisa sobre o diminutivo, baseada na observação de dados reais em um corpus oral informatizado.

3.2.1
A Lingüística de Corpus

A Lingüística de Corpus estuda fenômenos da linguagem por meio da observação de grandes quantidades de dados lingüísticos reais. Tais dados podem consistir em textos falados ou escritos, e são necessariamente oriundos de situações comunicativas no mundo real, correspondendo, assim, à língua em uso. A pesquisa de grandes quantidades de dados lingüísticos é possibilitada pelo auxílio de ferramentas computacionais. Através do uso de exemplos derivados de um corpus de dados reais é possível, como observa Philip (2003, p.86), evitar os problemas ligados ao uso de exemplos e experimentos especificamente construídos para verificar ou rejeitar hipóteses. Em pesquisa de corpus, cada item de língua pode ser investigado no seu contexto de ocorrência, possibilitando a avaliação do impacto real do contexto sobre o significado de um item usado.

Até os anos 1950, a pesquisa pré-computacional de corpora foi um método bastante usado em lingüística. No entanto, como foi observado por Gries (2007), com o paradigma gerativo-transformacional, a lingüística do século XX viu-se tomando um rumo racionalista e formalista, com o foco no estudo da competência lingüística de um falante idealizado. Essa mudança de perspectiva teórica se juntou com uma mudança metodológica: uma pesquisa baseada em dados foi substituída por uma pesquisa baseada em julgamentos de aceitabilidade, baseados por grande parte em introspecção. No entanto, o crescente interesse em dados baseados no desempenho dos falantes como reflexão do sistema lingüístico, junto com as tecnologias informáticas cada vez mais poderosas e disponíveis, resultou em uma revitalização das pesquisas empíricas, baseadas no uso que os falantes fazem dos recursos lingüísticos. Os corpora são usados para obter conhecimento empírico sobre a linguagem e evitam os problemas inerentes a metodologias que se apóiam na introspecção e na intuição do falante como fontes lingüísticas. Por

exemplo, listas de concordância computarizadas podem tornar evidentes fatos de linguagem que não são acessíveis à introspecção. Philip (2003, p.119) observa que os sentidos salientes das palavras e expressões são normalmente identificados com mais facilidade nas intuições do lingüista do que os sentidos menos salientes. Por isso, há fatos de linguagem que são imperceptíveis na introspecção. A evidência a partir de análise de corpora computarizadas, baseadas em amostras reais da língua, pode vir a questionar as intuições do falante às vezes até de maneira drástica.

Assim como a observação de Salomão sobre a Lingüística Cognitiva no capítulo 3.1.1, o termo "Lingüística de Corpus" também pode ser entendido como um termo designando antes uma empreitada coletiva do que uma linha teórica ou metodológica com fronteiras estreitamente delimitadas. Na primeira fase das pesquisas lingüísticas baseadas em corpus, o termo "corpus" foi usado para significar qualquer conjunto de dados arrolados ou de textos escritos, geralmente produzidos por um único autor. Com o surgimento e crescimento da Lingüística de Corpus como uma abordagem metodológica consolidada nas últimas décadas, simultaneamente com o rápido desenvolvimento das tecnologias de informática, a definição do termo se alterou[2], ressaltando-se três aspectos, citados por Baker (1995, p.225):

(i) O termo "corpus" é hoje usado para significar um conjunto de textos em formato processável por computador, passível de uma análise automática ou semi-automática de diversas formas.
(ii) Um corpus não é mais restrito à língua escrita, mas pode incluir tanto textos escritos quanto orais.
(iii) Um corpus pode conter um grande número de textos de fontes diferentes, produzidos por vários autores e versando sobre vários tópicos.

Os desenvolvimentos da Lingüística de Corpus e das tecnologias da informática estão estreitamente ligados, dado que o desenvolvimento dos computadores tem possibilitado a manipulação e análise de quantidades muito maiores de dados do que tem sido possível até recentemente. O processamento por computador é uma característica essencial nas pesquisas de corpora hoje em

[2] Observa-se que o uso do termo continua vigente também na pesquisa lexical em geral, independentemente do seu uso na Lingüística de Corpus.

dia, e, como observam McEnery & Wilson (1996, p.14), o termo "corpus" é hoje praticamente um sinônimo do termo "corpus manipulável por computador" (*machine readable corpus*). Devida à capacidade do computador de buscar, classificar e calcular dados, grandes quantidades de amostras da linguagem podem ser manipuladas e analisadas de forma rápida, eficiente e exata. Mais ainda, esses dados prestam-se para verificações ou reproduções posteriores, assim como para críticas dos outros pesquisadores, contribuindo, assim, à transparência da ciência.

Berber Sardinha (2000, p.336) chama atenção para a proliferação das definições para o termo "corpus" na literatura. É importante lembrar que entre as coletâneas de dados lingüísticos naturais, legíveis por computador, nem todo conjunto de dados pode ser considerado um corpus. Sinclair (1991) salienta que os textos escolhidos para constituir um corpus devem ser "naturais". Para o autor (1991, p.171, *apud* Berber Sardinha 2000, p.336), um corpus pode ser definido como "uma coletânea de textos naturais, escolhidos para caracterizar um estado ou variedade de linguagem". É claro que deve-se notar o caráter relativo do termo "natural" – até que ponto a linguagem acadêmica, por exemplo, é "natural"? Também, vale lembrar que ao deslocar um determinado texto do seu ambiente de ocorrência (por exemplo, ao trazer um jornal para a sala de aula), este obrigatoriamente perde a sua "naturalidade". Sobre a problemática do caráter natural dos textos, Berber Sardinha (2000, p.336) observa que "se por um lado os textos devam ser naturais (autênticos e independentes do corpus), o corpus em si é artificial, um objeto criado com fins específicos de pesquisa". Desta maneira, podemos concordar com o autor na sua observação que por textos "naturais" deve-se entender "autênticos" no sentido daqueles que "existem na linguagem e que não foram criados com o propósito de figurarem no corpus" (ibid., p.336). A idéia de "natural" pode também ser ampliada para incluir apenas textos produzidos por humanos, excluindo, desta forma, os textos gerados por programas criados para este fim. No entanto, é importante ter em mente o caráter artificial do corpus, assim como a importância da explicitação do propósito de criação de um corpus.

Entre as várias tentativas de definição do termo "corpus", Berber Sardinha (ibid., p.338), considera a definição de Sanchez (1995) aquela que melhor representa as características principais de um corpus computadorizado:

> "Um conjunto de dados lingüísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso lingüístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise."
>
> (Sanchez, 1995, p.8-9, *apud* Berber Sardinha, 2000, p.338)

Para Berber Sardinha (2000, p.338), essa definição tem o mérito de reunir todos os mais importantes aspectos a serem contemplados na compilação de um corpus, reproduzidos a seguir:

(a) A origem: os dados devem ser autênticos.

(b) O propósito: o corpus deve ter a finalidade de ser um objeto de estudo lingüístico.

(c) A composição: o conteúdo do corpus deve ser criteriosamente escolhido.

(d) A formação: os dados do corpus devem ser legíveis por computador.

(e) A representatividade: o corpus deve ser representativo de uma língua ou variedade.

(f) A extensão: o corpus deve ser vasto para ser representativo.

No capítulo 4 veremos como o nosso corpus corresponde a esses critérios.

Não resta dúvida sobre os benefícios que a Lingüística de Corpus tem trazido para uma pesquisa lingüística empírica. No entanto, é claro que nenhum corpus pode conter todas as informações lingüísticas. Apesar disso, o valor de uma abordagem que envolve análise de corpus fica evidente na seguinte observação de Fillmore (1992, p.35, *apud* Berber Sardinha, 2000, p.362): "não há nenhum corpus que contenha toda a informação que eu quero explorar", mas mesmo assim "todo corpus me ensinou coisas sobre a linguagem que eu não teria descoberto de nenhum outro modo".

3.2.1.1
Abordagem baseada em corpus (*corpus-based*) e abordagem dirigida por corpus (*corpus-driven*)

Um corpus pode ser usado de várias maneiras para desenvolver, exemplificar ou verificar uma teoria da linguagem. Apesar do termo "análise baseada em corpus" ser freqüentemente usado como um termo genérico para toda pesquisa lingüística que usa dados empíricos, dentro da Lingüística de Corpus, temos, pelo menos, duas abordagens distintas. Segundo Tognini-Bonelli (2001, p.84), o termo "baseada em corpus" (*corpus-based*) é usado para pesquisas nas quais um corpus é usado principalmente para exemplificar ou corroborar teorias ou análises existentes, enquanto o termo "dirigido por corpus[3]" (*corpus-driven*) é usado para pesquisas nas quais as afirmações teóricas são subordinadas à evidência fornecida pelo corpus.

Oliveira (2006, p.17) fornece uma apresentação ilustrativa das principais diferenças entre essas duas abordagens na forma de tabela, reproduzida embaixo:

[3] Em português, a terminologia parece não ser consolidada sobre esse termo: adotamos o termo "dirigido por corpus", usado por Oliveira (2006, p.16). Observa-se, no entanto, que Berber Sardinha (2002, p.34) usa o termo "movido a corpus", Gonçalves (2006, p.61) o termo "conduzido por corpus" e Carvalho (2007, p.39) o termo "orientado pelo corpus".

Lingüística baseada em corpus	Lingüística dirigida por corpus
o corpus é utilizado para validar, verificar e melhorar observações lingüísticas que já tenham sido realizadas	um corpus é de importância essencial no surgimento de novas idéias de como examinar os dados
o lingüista não questiona posições teóricas pre-estabelecidas ou categorias descritivas aceitas; sua posição com respeito à estrutura da língua já se estabilizou	o lingüista acredita que pode conciliar o tipo de evidências que emerge do corpus com as posições estabelecidas; ele deixa abertas as possibilidades de mudanças radicais na teoria para lidar com as evidências
o corpus é utilizado para ajudar a estender e melhorar a descrição lingüística	a evidência do corpus é soberana portanto o lingüista minimiza os pressupostos sobre a natureza das categorias teóricas e descritivas
um exemplo de questão relevante: WHOM ainda é utilizado em inglês? como?	um exemplo de questão relevante: a distinção entre gramática e léxico é necessária?

Tabela 1: Lingüística baseada em corpus vs. lingüística dirigida por corpus (Oliveira, 2006)

Uma **abordagem baseada em corpus** possibilita que teorias já existentes sejam corroboradas e refinadas através da análise de amostras autênticas da língua. Philip (2003, p.114-115) observa que nesse tipo de abordagem, no entanto, novas descobertas não acontecem com freqüência, já que os dados analisados são selecionados de antemão, com critérios baseados nas próprias características que o pesquisador se propõe a investigar. A autora assinala que, como conseqüência, esse tipo de abordagem pode criar um círculo vicioso de exclusão de dados já que exemplos que não se encaixam nas categorias pré-determinadas nem sempre recebem a devida atenção ou podem ser até deliberadamente excluídos, com o objetivo de manter a ilusão de ordem em um sistema que é inegavelmente complexo e multifacetado, e, há quem diga, até caótico. Philip cita autores como Aarts (1991), que justifica a exclusão das frases agramaticais da pesquisa baseando-se no argumento de que em qualquer corpus podem aparecer frases que foram criadas pelos falantes que, por motivos variados, deliberadamente

infringem as regras gramaticais, e que se essas violações são levadas em consideração, a generalizabilidade das postulações teóricas sobre a língua "como um todo" fica comprometida. No entanto, como observa Philip (ibid., p.115-116), esse tipo de abordagem traz o seguinte problema: se queremos impor uma definição de gramaticalidade muito estreita, definindo o que pode ser considerado como linguagem aceitável e compreensível em termos rígidos, ignorando os fatos de linguagem que não estão em harmonia com essa definição, a própria Gramática acaba se distanciando cada vez mais da linguagem que se propõe a classificar e teorizar. É claro, no entanto, que a crítica é pertinente apenas a análises com viés exclusivista e não se aplica à adequação da metodologia enquanto tal.

Com a **abordagem dirigida por corpus**, em oposição à abordagem baseada em corpus, entende-se uma metodologia na qual o corpus é visto como algo mais do que um simples depósito de exemplos para apoiar teorias pré-existentes. Para Tognini-Bonelli (2001, p.84), o comprometimento do pesquisador é com a integridade dos dados na sua totalidade, e as afirmações teóricas devem refletir diretamente a evidência extraída do corpus. Kerbrat-Orecchioni (1990, p.47) observa que na lógica direcionada pelos dados (*data driven*), é a partir da observação meticulosa dos fatos que se deve avançar para a elaboração de uma teoria: as construções teóricas devem ser inteiramente colocadas a serviço dos dados empíricos, e não o contrário. Segundo Berber Sardinha (2002, p.34), a metodologia envolvida nesse tipo de pesquisa visa à descrição abrangente dos dados, sem a intenção de selecionar exemplos para ilustrar elementos oriundos de uma teoria específica. A evidência fornecida pelo corpus vem em primeiro lugar. Desta maneira, mais do que verificar ou corroborar teorias existentes ou buscar exemplos para provar um argumento lingüístico, os proponentes dessa abordagem estão preocupados em descrever uma determinada amostra da língua de forma abrangente, evitando impor categorias pré-concebidas sobre os dados.

Apesar da crítica sobre a abordagem baseada em corpus apresentada por Philip (2003), assim como pelos defensores da abordagem dirigida por corpus em geral, acreditamos que esse tipo de abordagem pode na verdade revelar bastante sobre a linguagem. Mesmo em pesquisas nas quais os dados para a análise são selecionados de antemão, seguindo os critérios baseados nas características do objeto do estudo, várias coisas novas podem surgir nos olhos de um pesquisador atento. Embora respeitando a capacidade de geração de resultados surpreendentes

através de análise quantitativa de amostras gigantes de linguagem numa abordagem dirigida por corpus, nessa tese estamos adotando a outra posição, especificamente com o intuito de mostrar que uma análise em termos de abordagem baseada em corpus é de muita relevância tanto como instrumento de testagem de hipóteses como reveladora de aspectos novos da linguagem.

Philip (2003, p.119), ao descrever as características de uma pesquisa dirigida por corpus, observa que o pesquisador se preocupa tanto com aqueles exemplos que não correspondem à hipótese, ou que formam sub-grupos dentro da tendência hipotetizada, como com aqueles que corroboram a questão inicial da pesquisa. Cita mais uma vez Sinclair: "nenhuma instância deve ser desdenhada em nenhuma amostra, não importa o quanto inconveniente ou estranha seja" (Sinclair, 1991, p.94, *apud* Philip, 2003, p.119). No entanto, a nosso ver, esse tipo de atitude não é de propriedade exclusiva dos adeptos da pesquisa dirigida por corpus, mas é de suma importância também em toda análise baseada em corpus.

De fato, apesar da distinção entre essas duas formas de conduzir uma pesquisa lingüística na qual dados de corpus são usados, na prática da pesquisa, contudo, uma mistura das duas abordagens é muitas vezes necessária. Esse é o caso também da nossa pesquisa. Por um lado, procurou-se observar ocorrências do uso real que os falantes do português do Brasil fazem do diminutivo, com o objetivo de derivar regularidades a partir dessa observação de dados. Por outro lado, é obvio que essa observação foi influenciada pela teorização pré-existente sobre o tema. Na verdade, como foi observado por Passot (2004, p.154), na prática de uma pesquisa, a aparente dualidade entre a observação e a teorização consiste nas constantes idas e voltas entre uma e outra.

3.2.1.2 Representatividade

Apesar das inegáveis vantagens que uma análise de corpus traz para a teorização lingüística, deve-se lembrar que um corpus não é capaz, nem tem por objetivo, esgotar a riqueza de uma língua ou de fornecer uma imagem absolutamente fiel dela. Como salienta Passot (2004, p.151), uma compilação de dados, não importa o quão grande, não passa de um reservatório de exemplos atestados, frutos da criatividade potencialmente infinita dos falantes dentro dos

limites de uma língua. No entanto, a questão de representatividade se coloca inevitavelmente: em que medida os dados derivados de um conjunto limitado de textos pode refletir a língua como um todo?

Berber Sardinha (2000, p.343) observa que não existem critérios objetivos para a determinação da representatividade. Trata-se de algo que, apesar de ser almejado por todos, é difícil definir. Leech (1997, p.27 *apud* Berber Sardinha, 2000, p.345) chega a afirmar que a representatividade deve ser considerada "um ato de fé". A característica mais facilmente associada à representatividade é a extensão do corpus, o que, em termos simples, significa que para ter representatividade o corpus deve ser o maior possível (Sinclair, 1991, *apud* Berber Sardinha, 2000, p.342). Berber Sardinha observa, ainda, que dado que a linguagem é um sistema probabilístico, certos traços são mais freqüentes que outros. Por isso, para garantir a possibilidade de ocorrência de palavras ou significados de menor freqüência, é necessária a incorporação de uma quantidade grande de palavras em um corpus. Segundo o autor (ibid., p.344), a extensão do corpus comporta as seguintes três dimensões:

1. número de palavras
2. número de textos
3. número de gêneros, registros ou tipos textuais.

Segundo o autor, quanto maior o número de palavras, maior será a chance de o corpus conter palavras de baixa freqüência, as quais formam a maioria das palavras de uma língua. O número de textos maior garante que um determinado tipo textual, gênero, ou registro esteja mais adequadamente representado. Já o número de gêneros, registros ou tipos textuais permite uma maior abrangência do espectro genérico da língua.

Em relação ao tamanho do corpus, Berber Sardinha (ibid., p.346), baseando-se na observação dos corpora utilizados durante quatro anos de conferências de Lingüística de Corpus, sugere a seguinte classificação:

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Tabela 2: Classificação do tamanho do corpus (Berber Sardinha, 2000)

No próximo capítulo veremos como o nosso corpus corresponde a critérios de representatividade, assim como a outros critérios apresentados nesse capítulo.