ASPECTOS COGNITIVOS E COMPUTACIONAIS DO TRATAMENTO DA POLISSEMIA ATRAVÉS DE MÉTODOS ESTATÍSTICOS

Marco Rocha
Universidade Federal de Santa Catarina

Resumo

Este artigo discute o uso de métodos estatísticos para o tratamento de palavras ou expressões polissêmicas. A discussão aborda sobretudo a construção de modelos probabilísticos para a desambiguação de palavras polissêmicas em sistemas computacionais que realizam análise semântica e textual automaticamente. Procura-se caracterizar a metodologia utilizada para a análise de um corpus de treinamento e construção destes modelos probabilísticos a partir deste corpus de treinamento. Um exemplo de palavra polissêmica é utilizado para tornar clara a metodologia em questão. Uma breve análise do uso de métodos estatísticos em lingüística fundamenta a discussão subseqüente relativa aos modelos probabilísticos propriamente ditos.

1. Introdução

Desde, aproximadamente, meados da década de oitenta, tem sido produzida uma quantidade cada vez maior de trabalhos científicos relacionados ao processamento de linguagens naturais[1] - ou, alternativamente, à lingüística computacional - com base em abordagens que poderiam ser, grosso modo, caracterizadas como **métodos estatísticos**, embora possam ser encontradas denominações tais como **métodos estocásticos**, **métodos probabilísticos** ou **abordagens a partir de dados** (*data-driven approaches*) na literatura da área. Estas variações terminológicas possivelmente ocultam diferenças que não são triviais, mas o aspecto essencial que é comum a todas estas con-

1 Parece existir uma forte tendência recente para substituir esta expressão por tecnologia da linguagem humana. As razões para esta mudança não serão discutidas aqui. O termo processamento de linguagens naturais será a denominação utilizada neste artigo.

cepções é o uso intensivo de um corpus de grande porte, composto por textos coletados segundo critérios bem determinados, a fim de obter as informações lingüísticas necessárias ao processamento automático de linguagens naturais em sistemas de computador, geralmente através da construção de um modelo probabilístico do fenômeno focalizado no processamento.

A construção destes modelos probabilísticos com base nas informações contidas em um corpus consiste, em linhas gerais, na análise prévia de uma quantidade significativa de ocorrências do fenômeno que se pretende processar. Esta análise é feita conforme uma classificação segundo categorias apropriadas para a descrição do fenômeno em questão. A definição destas categorias é trabalho da equipe de analistas. Na maioria dos casos, é criada uma anotação de corpus que permita especificar a categoria de classificação adequada para cada ocorrência encontrada no corpus de maneira padronizada. Tanto as características de uma anotação de corpus quanto o processo de anotação - e as dificuldades derivadas, por exemplo, da discrepância entre anotadores – são objeto de pesquisas intensas (ver, por exemplo, Garside et al. 1997) e várias iniciativas de padronização (ver, por exemplo, a página da *Corpus Encoding Standard* ou a página *Linguistic Annotation*).

No que diz respeito à desambiguação automática de sentidos, o conjunto de sentidos possíveis de uma palavra polissêmica deve ser previamente determinado, por exemplo, com base em trabalhos de referência, tais como os dicionários. Isto não quer dizer que todas as ocorrências do corpus serão classificadas sem problemas com base nestes trabalhos de referência. A classificação, muito provavelmente, terá que ser modificada de modo a dar conta dos usos da palavra em situações reais, como tão freqüentemente acontece no processo de anotação de corpus. Por princípio de análise da lingüística de corpus, espera-se que o analista esteja pronto a ceder diante das evidências que o corpus fornece. Por isso mesmo, as abordagens baseadas em corpus são freqüentemente chamadas de abordagens a partir de dados.

Uma vez que uma quantidade suficiente de ocorrências da palavra polissêmica tenha sido anotada, está constituído o chamado corpus de treinamento, isto é, o conjunto de ocorrências de uma determinada palavra que será utilizado para que uma máquina "aprenda" os sentidos possíveis da palavra ou expressão polissêmica e detecte, no contexto imediato de ocorrência, características distintas associadas a cada um dos sentidos possíveis. Deste modo, um computador pode vir a distinguir qual dos sentidos possíveis de uma expressão polissêmica é o adequado em uma ocorrência nova. Esta distinção é feita com base em um modelo probabilístico, isto é, na probabilidade, dadas as características do contexto imediato desta ocorrência a ser processada, de seu sentido pertencer a esta ou aquela categoria dentre as que integram o conjunto de sentidos possíveis especificado no corpus de treinamento.

O processo de elaboração e aperfeiçoamento deste tipo de modelo probabilístico deverá, espera-se, ficar claro para o leitor no decorrer deste artigo. É importante notar, contudo, que a importância destes métodos estatísticos para a lingüística computacional tornou-se um fato consumado. Uma vez que não é possível transmitir a uma máquina o conhecimento lingüístico que permite a um falante de uma língua realizar esta desambiguação, seja este conhecimento oriundo de experiência no uso da língua ou de capacidades inatas, os métodos estatísticos constituem-se em uma alternativa de grande importância para a lingüística computacional.

A confrontação entre estes métodos estatísticos e os métodos anteriormente predominantes, chamados de **abordagens de base lógica** ou **de base em regras,** vem gerando controvérsias acaloradas. Pesquisas e resultados em sistemas construídos segundo cada uma destas abordagens têm sido comparados e discutidos. Tais sistemas incluem tanto os chamados módulos essenciais de um sistema que processa linguagem natural - como os etiquetadores de categorias morfossintáticas, os analisadores sintagmáticos e os sistemas de desambiguação de

sentido - quanto as aplicações típicas da lingüística computacional[2], ou seja, tradução de máquina, extração de informações, sumarização de texto, interfaces de diálogo e ensino de línguas com ajuda de computador, entre outras. No decorrer da refrega, os métodos estatísticos passaram a fazer parte das opções disponíveis para os projetistas de sistemas e, em maior ou menor grau, dos cursos de processamento de linguagens naturais (doravante, PLN) nas várias instituições de ensino e pesquisa que lidam com inteligência artificial.

Porém, o impacto destas tendências sobre a lingüística propriamente dita ainda é bastante limitado. Conforme observado em Abney (1995), os recursos de tratamento estatístico das línguas, que se tornaram lugar comum entre pesquisadores vinculados à área do PLN, não conseguiram atrair os lingüistas de maneira suficiente para que passassem a integrar, de fato, o arsenal de instrumentos de investigação da ciência que estuda as línguas humanas. Não resta dúvida que a lingüística de corpus[3] já tem uma penetração considerável na área da lexicografia, sobretudo de língua inglesa. Contudo, também é verdade que os lexicógrafos pertencem a um grupo de estudiosos da língua que não se integra facilmente ao corpo principal de pesquisa realizada em lingüística.

Por outro lado, ao omitir-se neste debate, a lingüística corre o risco de ter sua participação cada vez mais reduzida em uma área de atividade cientificamente rica no momento atual, a ciência cognitiva. A minimização do papel da lingüística neste ramo do pensamento científi-

2 O termo lingüística computacional, embora freqüentemente usado em contextos onde processamento de linguagem natural também seria terminologia adequada, é mais abrangente. De um modo geral, inclui áreas de aplicação, como o ensino de línguas com ajuda de computador, que pressupõem um grau de interdisciplinaridade maior do que as tipicamente presentes no âmbito do ramo da inteligência artificial chamado de processamento de linguagem natural.

3 Existem algumas variações terminológicas também em relação a esta denominação, tais como lingüística do corpus e lingüística com corpus. As diferenças conceituais e metodológicas em relação ao uso de um corpus que estas distinções de terminologia buscam representar não serão assunto deste artigo.

co seria particularmente indesejável tendo em vista o forte impacto que as idéias de Chomsky tiveram na sua gênese. Contudo, o crescente distanciamento das teorias lingüísticas em relação a várias questões atuais da ciência cognitiva — como o conexionismo, os algoritmos genéticos e o aprendizado de máquina, com a conseqüente renovação do debate relacionado ao contraste entre capacidades inatas e adquiridas — torna suas formulações, cada vez mais abstratas e impérvias à verificação empírica, desinteressantes para a definição dos processos cognitivos que de fato permitem o uso da linguagem conforme observada na realidade. A questão será discutida em maior detalhe adiante.

Neste artigo, será apresentada, na próxima seção, uma visão geral dos motivos que parecem ter levado a esta rejeição dos métodos estatísticos em lingüística. A argumentação contrária a esta rejeição inclui uma análise das contribuições que os métodos estatísticos podem trazer a níveis fundamentais da descrição lingüística, como a sintaxe e o léxico, incluindo-se aí o aspecto cognitivo, com alguns acréscimos que dizem respeito a outras áreas onde parece visível que a contribuição destes métodos já é uma realidade. Na terceira seção, as abordagens estatísticas do léxico, sobretudo as relacionadas a sistemas de desambiguação automática de sentidos de palavras, serão especificamente discutidas. A quarta seção resume as perspectivas destas abordagens tanto em termos de ciência cognitiva quanto em relação aos sistemas de desambiguação de sentidos de palavras polissêmicas.

2. Métodos estatísticos e lingüística

A rejeição dos métodos estatísticos por parte da lingüística está relacionada a uma opção por abordagens racionalistas da linguagem, em detrimento das concepções empiristas. O debate que opõe estas duas correntes já faz parte das discussões filosóficas desde a era clássica e, portanto, se estende além da lingüística. Na verdade, o confronto entre estas duas abordagens focaliza aspectos essenciais dos métodos de

análise da realidade, sobretudo na relação entre o mundo real e a mente humana, particularmente no que diz respeito à maneira pela qual ocorre a apreensão deste mundo real por parte da mente.

Não é propósito deste artigo examinar em detalhe questões mais claramente relevantes no âmbito da filosofia da linguagem, ou, talvez mais adequadamente, de uma filosofia da ciência no âmbito dos estudos das linguagens humanas. Na subseção seguinte, entretanto, será apresentado um resumo sucinto do impacto que estas questões tiveram na metodologia dos estudos em lingüística desde o final da década de cinqüenta, sem qualquer pretensão de exaustividade. A segunda subseção discute as áreas de estudo da lingüística onde o uso de métodos estatísticos já é aceito normalmente, procurando definir quais aspectos destes estudos podem ser generalizados para a pesquisa lingüística de um modo geral. A terceira subseção examina alguns dos argumentos mais recentes contra o uso de métodos estatísticos nos estudos lingüísticos, na tentativa de demonstrar que estes argumentos derivam de idéias equivocadas baseadas em preconceitos quanto ao objeto de estudo da lingüística.

2.1. Racionalismo e empirismo, de novo

O predomínio das concepções racionalistas em lingüística, a partir do final da década de cinqüenta até meados da de oitenta, acabou por causar uma predisposição para a rejeição dos métodos estatísticos em estudos da linguagem e da cognição. Esta rejeição passou a ser vista como um fato intrínseco às abordagens verdadeiramente científicas no estudo das línguas humanas, e, portanto, um aspecto positivo. A predisposição também se fez presente tanto em psicologia quanto em PLN, assim como em inteligência artificial de um modo geral. Em poucas palavras, as abordagens racionalistas, nas concepções mais recentes, podem ser caracterizadas pela descrição do conhecimento contido na mente humana com base na crença de que tal conhecimento não é resultado da percepção sensorial, mas já se encontra previamente definido no nascimento, possivelmente por meio de herança genética.

No caso da lingüística, a aceitação generalizada das concepções de Noam Chomsky, que postulam a existência de uma faculdade inata da linguagem, forneceram o fundamento para esta rejeição radical do tratamento estatístico de problemas lingüísticos. Não é difícil encontrar lingüistas que nem sequer consideram a discussão digna de atenção, uma vez que Chomsky já libertou a área destas concepções "behavioristas" há mais de quarenta anos. O adjetivo na sentença precedente é, na maioria dos casos, suficiente para condenar qualquer argumentação em contrário ao opróbrio universal, e nada mais é necessário acrescentar.

Não obstante, a distinção entre estas duas abordagens é descrita de maneira bem mais útil como uma questão de grau, ao invés de em termos de uma separação absoluta entre conhecimento inato e aprendizado a partir da interação com o meio. Exceto por algumas correntes extremas, a idéia da *tabula rasa*, isto é, da inexistência total de capacidades cognitivas pré-existentes no cérebro humano, com a conseqüente defesa da experiência como fonte exclusiva de todo conhecimento, não é levada a sério, mesmo pelos defensores das concepções empiristas, uma vez que torna muito difícil explicar a presença ubíqua de determinadas formas de conhecimento na espécie humana, como é o caso da faculdade da linguagem. Tais formas de conhecimento só são observadas em outras espécies em níveis de desenvolvimento muito rudimentares e qualitativamente diferentes. Por outro lado, as concepções que explicam a cognição com base apenas em conhecimento inato têm dificuldade em demonstrar como a atuação destas capacidades inatas, sob formas altamente especializadas, resulta em comportamentos observáveis que apresentam características claramente adquiridas, pelo menos em certa medida, através do convívio social.

Deste modo, a discussão torna-se bastante mais interessante, do ponto de vista científico, se colocada em termos mais "fracos". As concepções racionalistas propõem um grau muito maior de especialização das capacidades cognitivas inatas do que o proposto pelas concepções empiristas. O ponto de partida do que se compreende como

conhecimento humano incluiria um conjunto de princípios especializados para cada um dos domínios cognitivos. No caso da linguagem, aspectos diretamente relacionados à estrutura morfológica e sintática fariam parte da herança genética já codificada no cérebro no momento do nascimento. Segundo Chomsky, esta concepção permitiria resolver o problema da pobreza do estímulo, isto é, como as crianças conseguem aprender algo tão complexo quanto uma língua a partir dos dados assistemáticos de que dispõem em seus primeiros anos, dados estes que exigem interpretação nada trivial para que o conhecimento básico da língua seja extraído destas informações.

As concepções empiristas são desenvolvidas com base na existência de princípios gerais inatos de organização cognitiva, tais como associação, reconhecimento de padrões e generalização. A aplicação destes princípios aos estímulos do meio físico e social percebidos sensorialmente é suficiente para que a criança aprenda o sistema complexo da linguagem, assim como para o desenvolvimento de outras capacidades cognitivas, tais como a manipulação precisa de instrumentos para a realização de tarefas específicas. A aquisição de conhecimento como a linguagem não depende de uma capacidade cognitiva específica para a linguagem, expressa em princípios detalhados. Esta diferença de concepção faz com que os objetos de análise das duas abordagens sejam também distintos. Enquanto as abordagens empiristas concentram-se na análise dos processos cognitivos conforme observados em situações reais, o aspecto mais importante, no âmbito das concepções racionalistas, é a especificação destes princípios inatos que não podem ser depreendidos a partir da observação de comportamentos.

A tarefa da lingüística, nas concepções racionalistas, notadamente no paradigma gerativista predominante, é, portanto, desvendar, por meio da introspecção e da própria intuição dos analistas, os princípios específicos da competência lingüística ou linguagem internalizada, termos usados para se referir a este conhecimento

lingüístico inato. A análise da língua conforme usada em situações de comunicação no mundo real, deste modo, não apenas é insuficiente para revelar o funcionamento deste conjunto de princípios, mas fornece uma imagem distorcida deste conhecimento nuclear, uma vez que fatores extra-lingüísticos alteram as características da linguagem produzida. Estes fatores incluem, entre outros, o cansaço, as limitações de memória, hierarquias sociais refletidas na linguagem, hábitos da comunidade a que pertencem os falantes e a possibilidade de terem ingerido muita cerveja.

No âmbito deste esforço analítico, as investigações da lingüística de orientação racionalista concentraram-se na definição de **regras** que especificassem este conhecimento nuclear, base do conhecimento que permite aos falantes de uma língua reconhecerem as sentenças que a ela pertencem. Esta forma de lidar com os fenômenos lingüísticos faz parte da tradição dos estudos da linguagem há pelo menos dois mil anos, mas, na concepção gerativista, evoluiu no sentido de um nível cada vez maior de formalização segundo critérios rigorosos para a construção de gramáticas altamente detalhadas que definiam as sentenças bem formadas de uma língua, ao mesmo tempo em que rejeitavam as mal formadas ou agramaticais. No que diz respeito à análise semântica, passou a predominar a mesma ênfase em regras formais, muitas vezes fortemente dependentes da análise da estrutura sintagmática, para a produção das chamadas formas lógicas, as quais freqüentemente incluíam representações baseadas em gramática de casos (Fillmore 1968; 1977) ou escopo de quantificadores (Cooper 1983).

A abordagem foi adotada com entusiasmo em PLN (ver Bruce (1975) para aplicações de gramática de casos e Woods (1978) para aplicações de quantificação em interpretação semântica). As exigências de formalização coadunavam-se com as características e possibilidades das linguagens de programação. Além disso, reforçavam e expandiam para o importante campo da linguagem humana a concepção racionalista dos processos cog-

nitivos, igualmente predominante em inteligência artificial. A despeito da atmosfera de otimismo que transparece na maioria dos trabalhos da década de setenta e parte da de oitenta, os resultados obtidos em sistemas projetados para lidar com linguagens humanas com base em regras pré-definidas são modestos. O êxito destes sistemas dependia de um tal número de restrições quanto aos textos a serem processados, que ficaram pejorativamente conhecidos por seus críticos como "sistemas de brinquedo" (ver Button et al. 1997 para uma discussão crítica das promessas e resultados reais da lingüística computacional).

Esta breve retrospectiva não tem a intenção de diminuir as conquistas, não obstante reais, do PLN como ramo ativo da área de pesquisa denominada inteligência artificial. Seu propósito fundamental é tentar demonstrar que os problemas encontrados pelos projetistas de sistemas em PLN são fruto da concepção racionalista da linguagem humana e sua expressão típica em lingüística, isto é, a definição da gramática de uma língua por meio de regras de boa formação de sentenças que distinguem, de maneira categórica, os enunciados gramaticais dos agramaticais, e suas extensões no campo da interpretação semântica. Alguns aspectos desta questão são analisados na subseção que se segue, inclusive uma discussão de alguns problemas de descrição lingüística que também derivam de uma concepção racionalista, a despeito de sua maior ou menor eficiência em termos tecnológicos.

2.2. Estatística e análise sintagmática

Não seria difícil, dentre as várias tendências da lingüística, encontrar questionamentos à distinção categórica entre sentenças gramaticais e agramaticais, sobretudo se feita com base em uma gramática previamente determinada. O problema foi resumido na frase freqüentemente citada de Sapir (1921): "Todas as gramáticas vazam" (p.38). Não parece possível definir de forma exata quais são os enunciados gramaticais de uma língua, de modo a separá-los inequivocamente de todas as outras seqüências de palavras, consideradas então como agra-

maticais. Isto porque a análise morfossintática e sintagmática[4] de qualquer sentença sempre inclui mais de uma possibilidade. Abaixo há três exemplos que exploram o problema, procurando demonstrar que os métodos estatísticos constituem, no mínimo, uma alternativa viável para superar esta dificuldade.

(1) O teu sete em lá volta na baixo.
(2) Peça um de ferro.
(3) Maria limpa o quarto.

A decisão quanto à gramaticalidade das sentenças acima parece simples. Os exemplos (2) e (3) são enunciados comuns do português, enquanto (1) é um amontoado de palavras, e, portanto, agramatical. Entretanto, não apenas (1) é uma sentença perfeitamente inteligível, mas foi coletada "de ouvido" em uma situação real de discurso falado. Os leitores versados no vocabulário dos músicos talvez tenham compreendido (1) logo na primeira leitura, ou tenham sido posteriormente despertados para esta possibilidade de análise pela "dica" do método de coleta na sentença precedente. Para os demais, a análise de (1) tem que ser esclarecida.

4 A expressão análise sintagmática é usada neste artigo como uma tradução da palavra inglesa parsing, no sentido em que é utilizada na literatura de PLN, isto é, a definição da estrutura sintagmática desde o componente mais amplo – a sentença ou enunciado – até

(4) SN ← O teu sete em lá
SN ← O teu sete
DET ← O
DET ← teu
N ← sete
SP ← em lá
P ← em
N ← lá
SV ← volta na baixo
V ← volta
SP ← na baixo
CONTR ← na
P ← em
DET ← a
N ← baixo

O sintagma nominal inicial pode parecer ainda um tanto misterioso, mas basta interpretar *lá* como um dos tons musicais possíveis e *sete* como a definição do compasso, com o possessivo indicando que o ouvinte compôs a passagem musical em questão. Portanto, o SN significa *aquela passagem no compasso de sete no tom de lá que tu compuseste*. O sintagma verbal completa a sentença ao definir que a passagem *volta*, isto é, é executada de novo, *na* (flauta) *baixo*. Embora a concordância em questão possa parecer estranha à primeira vista, a classificação das vozes humanas é usada em português conforme criada em italiano. Ninguém se espanta com concordâncias como "a soprano está rouca", uma vez que a palavra *soprano* só é usada em português na classificação de vozes, e nem sequer se considera que o *o* final denote masculinidade no mundo real. Há mesmo quem prefira usar o artigo masculino.

A classificação de instrumentos com base nas vozes humanas exige um pouco mais de conhecimento especializado, mas isto não altera o fato de que *baixo* no exemplo (1) faz referência ao registro da voz

masculina grave na escala musical. A interpretação é clara para os falantes familiarizados com o vocabulário musical. O fato de existir, na língua portuguesa, o adjetivo *baixo*, o qual, como os demais adjetivos, exige concordância com o núcleo do sintagma nominal quando usado em seu sentido mais comum, não muda a concordância para a classificação dos instrumentos segundo as vozes humanas. Deste modo, só faria sentido realizar a concordância se o adjetivo se referisse ao volume sonoro da flauta. Do contrário, como não se tem notícia, na história da música, de nenhuma mulher que cantasse na região da voz de baixo, a concordância com o substantivo omitido, neste uso, pareceria um tanto ridícula para os que tiveram educação musical.

Os exemplos (2) e (3) colocam a questão bastante discutida da multiplicidade de análises possíveis. Em (2), à primeira vista, trata-se de uma sentença com o verbo no imperativo, na qual alguém comunica ao ouvinte que deve pedir apenas uma unidade de algum objeto não especificado que seja de ferro. A definição do material pressupõe uma provável existência de alternativas. Porém, é possível interpretar o mesmo enunciado como um sintagma nominal isolado, em que uma *peça* qualquer do tipo *um* feita de *ferro* fosse especificada como resposta a uma pergunta ou reação perfeitamente compreensível em determinado contexto. A polissemia do substantivo *peça* ainda permite outras interpretações menos prováveis.

No caso (3), tanto a polissemia do verbo *limpar – eliminar a sujeira* ou *roubar*? – quanto a de *quarto – cômodo* ou *o objeto que se segue ao terceiro e precede o quinto*? – possibilitam pelo menos quatro interpretações. Pessoas de boa imaginação também poderiam encontrar mais interpretações menos prováveis. As decisões relativas à análise morfossintática, que define a classe das palavras, a análise sintagmática e a desambiguação dos sentidos das palavras polissêmicas interagem em situações reais de uso da língua. Não se quer com isto dizer que a sintaxe não é autônoma como nível de análise lingüística, mas que autonomia não significa isolamento. O tratamento da gramática de uma língua

como um conjunto de regras que separa categoricamente as sentenças gramaticais das agramaticais não tem como integrar estas interações à análise sintagmática sem romper com o determinismo embutido no formalismo.

A realidade de uso da língua, portanto, coloca a ambigüidade de natureza morfossintática e a polissemia como aspectos essenciais da investigação lingüística. Não é preciso apontar aqui as dificuldades que os julgamentos de gramaticalidade geram. Todos, seja como leitores ou ouvintes em palestras, já se depararam com exemplos de sentenças tratadas como gramaticais, em determinados contextos analíticos, pelo autor ou palestrante, quando parecem claramente agramaticais para o leitor ou ouvinte, e vice-versa. Freqüentemente, a discussão termina com uma afirmação como "no meu dialeto, é gramatical", o que não parece muito científico. Qualquer análise sintagmática inclui a possibilidade de várias interpretações, e, portanto, a incerteza.

Uma vez que a incerteza seja vista como parte integrante da análise lingüística, os métodos estatísticos passam a ser uma opção natural de tratamento dos fenômenos analisados. A estatística é, por excelência, a ciência da incerteza, onde o tratamento determinístico é substituído por modelos baseados em probabilidades altas e baixas, mas nunca nulas. Se a afirmação do parágrafo anterior relativa ao dialeto fosse feita com base em um número suficientemente grande de ocorrências coletadas em um corpus de dimensões conhecidas da variedade lingüística em questão, o problema da gramaticalidade "teórica" da sentença tornar-se-ia uma discussão irrelevante, sobretudo se a probabilidade de ocorrência da forma em questão fosse especificada com base em variáveis estatisticamente mensuráveis.

2.3. Aquisição da linguagem, mudança e variação

Algumas áreas da investigação lingüística já incorporaram métodos estatísticos há algum tempo. Isto porque as questões contidas nos seus respectivos objetos de pesquisa apontam para os métodos estatísticos

como a solução para o tratamento científico dos problemas a serem analisados. Deste modo, em estudos de aquisição da linguagem, as mudanças nas gramáticas das crianças, as quais podem ser observadas nas freqüências relativas com que determinadas estruturas são usadas, são tratadas por meio de especificações destas freqüências relativas, e, portanto, de estatísticas. Modelos probabilísticos podem definir quais os elementos - tais como contextos lingüísticos de qualquer natureza, fatores sociais ou faixa etária – que influenciam na probabilidade de ocorrência de uma ou outra estrutura, especificando, inclusive, o momento em que uma determinada estrutura é descartada, e uma outra se torna estável. Um tratamento categórico do processo de aquisição prevê mudanças abruptas que na realidade não são observadas.

É possível argumentar dentro desta mesma linha em relação tanto à mudança lingüística quanto à variação lingüística. A dificuldade em distinguir com precisão entre estes dois fenômenos só reforça a adequação dos métodos estatísticos para tratá-los. O redator da Folha Ilustrada, na edição de 7 de dezembro de 1994, registrou uma forma de mudança lingüística no nome próprio *Edmundo* da seguinte maneira:

(5) EDMUNDO

> Para ser usada **tipo** "**rolou** o maior Edmundo".
> Inspirada no esquentado jogador do Palmeiras, serve para designar briga, **barraco**.

O registro do jornalista, na verdade, não tinha a intenção de constatar mudança lingüística, mas de ironizar a irascibilidade do indivíduo. Não obstante, o redator inadvertidamente caracteriza três usos recentemente integrados à língua portuguesa de formas lexicais existentes anteriormente, marcados em negrito no exemplo. O uso do substantivo *tipo* como uma locução adverbial, com o sentido aproximadamente de *mais ou menos assim*, é um desenvolvimento de seu uso no sentido

de *classe* ou *forma*. A análise de um corpus histórico poderia detectar um período em que este último sentido, sob forma estrita, era o único encontrado. O exemplo (6) abaixo, também extraído da Folha de São Paulo, edição de 13 de abril de 1994, demonstra este uso.

(6) Lobotomia é um **tipo** de cirurgia no cérebro, que era usado para "apagar" comportamentos mentais considerados deficientes, como a esquizofrenia.

Um aspecto facilmente detectado em abordagens a partir de corpus é a companhia em que a palavra analisada aparece, em relação aos sentidos possíveis. Parece seguro afirmar que a combinação *tipo de* é particularmente freqüente neste sentido de *classe*. Esta abordagem permite a inclusão de elementos sintáticos, tornando a interação entre os diferentes níveis de análise lingüística um padrão de investigação. Com base neste mesmo corpus histórico, infelizmente hipotético, poder-se-ia testar a hipótese de uma evolução gradual na direção da locução adverbial, passando pela forma intermediária exemplificada em (7) (Folha de São Paulo, 4 de novembro de 1994), em que o sentido de *classe* é preservado em maior medida, mas já sem o rigor das taxonomias científicas.

(7) Se você está numa grande universidade, **tipo** USP ou Unicamp, localize o "Sysop" (o operador de sistema).

É provável que uma análise detalhada das 9210 ocorrências da palavra *tipo* no corpus do NILC[5] detecte outras variações de sentido e possa associá-las a colocações e a estruturas sintagmáticas. O propósito da discussão aqui apresentada é demonstrar que a análise de fenômenos de mudança lingüística já é naturalmente feita com base em informações de freqüência ao longo do tempo. Os métodos estatísticos

5 Núcleo Interisntitucional de Lingüística Computacional

são, conseqüentemente, ideais para o tratamento de fenômenos de mudança lingüística. O verbo *rolar* e o substantivo *barraco*, também presentes no exemplo (5), poderiam ser alvo de investigações semelhantes. No caso do primeiro vocábulo, o estudo seria particularmente interessante, uma vez que o verbo *rolar* está cada vez mais freqüentemente sendo usado no sentido de *existir*, absorvendo inclusive as características sintáticas do verbo *haver*, quando usado neste mesmo sentido, já que é possível argumentar que *o maior Edmundo* é objeto de *rolou* em (5), e o sujeito, portanto, pode ser tratado como inexistente.

A variação lingüística é tratada estatisticamente no âmbito da sociolingüística desde a definição de regra variável (Labov 1972, Suppes 1971), inclusive com a elaboração de pacotes estatísticos especificamente planejados para fins de análise variacionista, baseados em procedimentos como qui-quadrado e análise loglinear, de amplo uso em muitas aplicações em estatística. O estudo dos vários dialetos de uma língua, tanto em termos geográficos quanto sociais, exige a análise de um continuum de variação, onde a freqüência de construções e usos de itens lexicais muda diferenciadamente. Os métodos estatísticos são uma alternativa metodológica ideal para definir a relevância de fatores geográficos, lingüísticos e sociais para cada dialeto ou registro.

A variação de natureza tipológica também vem sendo tratada estatisticamente com bastante êxito. Na realidade, a despeito da relação entre as abordagens racionalistas e a chamada gramática universal, a identificação de muitos dos universais lingüísticos especificados até os dias de hoje se faz com base em análises estatísticas. Os universais absolutos, conforme definido em Crystal (1995), não existem[6], exceto por afirmações tão genéricas – como, por exemplo, "todas as línguas têm vogais" -, que sua utilidade é praticamente nenhuma. Os univer-

6 Most of the time, in fact, it is clear that 'absolute' (or exceptionless) universals do not exist. (p.85)

sais relativos – os quais podem ser expressos em termos estatísticos – são os que de fato oferecem possibilidades científicas relevantes.

Deste modo, os universais lingüísticos aparecem geralmente sob a forma de conclusões extraídas de uma amostra suficientemente grande das línguas humanas. As conclusões definem, por exemplo, que, em 99% das línguas cuja ordem das palavras foi estudada, o sujeito gramatical precede o objeto (Crystal 1995). Em outro estudo de foco fonológico, abrangendo 317 línguas (Maddieson 1984), foi observado que menos de 3% não possuem consoante nasal. O conceito de forma "marcada" e "não-marcada" fica assim vinculado à predominância estatística.

Os universais implicacionais aparecem sob a forma de afirmações do tipo "se X, então Y", descrevendo uma relação constante entre duas propriedades das línguas. Greenberg (*apud* Crystal 1995) propõe uma lista de 45 universais. O universal 43, por exemplo, estabelece: "se uma língua tem categorias de gênero para os substantivos, também tem categorias de gênero para os pronomes". Estes universais implicacionais também são elaborados com base em estatísticas extraídas de amostras grandes da população, no sentido estatístico, de línguas humanas. Da mesma maneira, o estudo da variação e da mudança lingüísticas pode ser conduzido com êxito com base em um tratamento estatístico da população de gramáticas - ou léxico-gramáticas, de modo a dar conta da interação discutida acima – existentes em uma comunidade da falantes de uma determinada língua, representadas em uma amostra, ou seja, o corpus.

2.4. Isso não é lingüística !

Como conclusão desta argumentação já excessivamente longa em favor do uso de métodos estatísticos em lingüística, parece adequado tratar especificamente do aspecto cognitivo[7]. Foi mencionado anterior-

7 Esta subseção contém argumentação que utiliza intensivamente material apresentado em Manning e Schütze (1999) e Abney (1995).

mente que a separação entre a abordagem racionalista e a empirista causa uma distinção quanto ao objeto de pesquisa. Esta distinção, freqüentemente definida em termos de competência e desempenho, põe em dúvida a validade de tudo que foi discutido anteriormente, uma vez que caracterizaria as questões relacionadas à percepção de gramaticalidade e ambigüidade como dados do desempenho, e, conseqüentemente, fora do âmbito da lingüística.

É bastante difícil compreender com exatidão a diferença entre os julgamentos de gramaticalidade com base em uma teoria da gramática, por melhor que seja, e as percepções de gramaticalidade com base na inserção da sentença ou sintagma em questão em algum contexto possível. Na realidade, os julgamentos de gramaticalidade de uma sentença isolada parecem sempre incluir um tipo de teste psicolingüístico informal que procura inserir a estrutura analisada, abstraída do exemplo de sentença, em algum contexto possível, além de verificar a aceitabilidade dos elementos que compõem a estrutura.

Deste modo, a distinção entre um conhecimento lingüístico nuclear e o uso de estruturas sintagmáticas ou itens lexicais em situações reais parece fortemente superestimada. De alguma maneira, a teoria lingüística predominante desenvolveu o hábito de separar os dados lingüísticos em aspectos que refletem restrições gramaticais decorrentes deste conhecimento nuclear e aspectos que podem ser atribuídos a mecanismos não-essenciais relacionados ao processamento. Em última análise, a produção e compreensão de línguas humanas, conforme observadas em situações reais, devem ser depuradas destes aspectos não-lingüísticos, de modo a permitir que os aspectos que refletem o conhecimento lingüístico fundamental sejam analisados.

Porém, esta distinção não está baseada em nenhum indício experimentalmente constatado. Não há nenhuma divisão fisiológica conhecida que atribua uma função de armazenar regras gramaticais a um determinado conjunto de neurônios, e, em contrapartida, a função de realizar o acesso a este repositório de regras de gramática e o processamento lin-

güístico de um modo geral, em situações reais de uso da língua, a um segundo conjunto de neurônios. A distinção entre gramática e processamento pode, sem dúvida, ser defendida em termos de utilidade metodológica, mas não tem qualquer fundamento empírico e, tendo em vista a sofisticação dos desenvolvimentos recentes em neurofisiologia e ciência cognitiva, parece um tanto simplória e ingênua.

Por esta razão, a opção por analisar exclusivamente fenômenos que refletem a competência lingüística de um falante ideal foi seguramente útil e permanece como um caminho frutífero em lingüística, mas representa uma definição de interesse por determinado tipo de problema lingüístico, e não à lingüística em si. Por outro lado, a participação da lingüística nas investigações mais recentes da ciência cognitiva corre o risco de se tornar cada vez menos significativa, justamente por ignorar questões como o conexionismo e o processamento paralelo distribuído, que redefiniram a discussão da relação entre capacidades inatas e aprendizado (ver Abney 1995).

Limitações de espaço tornam impossível realizar uma análise mais profunda dos aspectos cognitivos, mas é importante frisar que o debate relativo à aquisição de conhecimento, inclusive o conhecimento lingüístico, continua bastante ativo, e é prematuro tratá-lo como se estivesse encerrado. O tratamento da linguagem e cognição humanas como fenômenos probabilísticos não apenas é possível, mas vem produzindo resultados importantes. O conhecido argumento usado por Chomsky (1957), relativo à produtividade da língua humana, só parece suficiente para os que analisam o problema a partir da conclusão, isto é, os que, já inicialmente, não acreditam na relevância dos métodos estatísticos para estudos da linguagem humana e da cognição de um modo geral.

A conclusão em questão, conforme estabelecida por Chomsky na época, afirma que a teoria da probabilidade é *inadequada* para a formalização da noção de gramaticalidade, porque o cálculo da probabilidade de sentenças a partir de um corpus atribuiria a mesma probabilidade bai-

xa para todas as sentenças que não fizessem parte do corpus, fossem estas sentenças gramaticais ou não (Chomsky 1957). Conseqüentemente, a produtividade das línguas humanas não poderia ser explicada por meio de probabilidades. Este argumento só permite chegar à conclusão pretendida se houver prevenção do analista em relação à formação de conceitos de um modo geral com base em raciocínio probabilístico.

Um conceito como, por exemplo, *gordo*, tem uma característica de relatividade, isto é, há uma área cinzenta intermediária em que diferentes observadores podem divergir quanto à descrição de uma pessoa como sendo gorda ou magra. Esta área cinzenta coincide, de um modo geral, com pesos medianos, isto é, típicos de seres humanos adultos na sociedade atual. A descrição da pessoa, alternativamente, pode ser verbalizada como *nem gorda, nem magra, normal*. A definição de *normal*, indubitavelmente, tem base probabilística, isto é, está baseada na noção de média, um dos pilares da teoria da probabilidade. O conceito de peso médio é, inclusive, utilizado com êxito em várias situações práticas importantes, como avaliações de controle de saúde e especificações obrigatórias da capacidade de elevadores em edifícios comerciais e residenciais.

Outro aspecto da relatividade da noção de *gordo* é o conhecimento que o observador tem da pessoa que está sendo descrita. Um observador, ao deparar-se com uma pessoa que atualmente pesa sessenta e cinco quilos, e com a qual conviveu durante muito tempo, poderia dizer "*como estás gordo !*", caso esta pessoa pesasse cinqüenta quilos no longo período de convivência anterior, pressupondo naturalmente um interregno suficientemente prolongado em que ambos não tivessem contato. Embora o peso de sessenta e quilos possa estar abaixo da média, por exemplo, das pessoas do sexo masculino da sociedade urbana no Brasil, a convivência anterior causou a retenção, na memória do observador, de uma pessoa de cinqüenta quilos. Isto faz com que os sessenta e cinco quilos atuais sejam motivo para a classificação do indivíduo em questão como uma pessoa que está gorda.

Vamos supor agora que este mesmo observador se depare com uma pessoa de altura mediana que pese duzentos quilos, e que jamais tenha visto uma pessoa com este peso antes. Parece razoável afirmar que o observador não teria dificuldade em classificar esta pessoa como *gorda*, e não como alguém que não é capaz de classificar, por jamais ter visto pessoa semelhante. Parece igualmente razoável esperar que uma pessoa de altura mediana e quarenta quilos seja classificada como *magra*, ou, pelo menos, não *gorda*. Os modelos probabilísticos são caracterizados exatamente por esta capacidade de apreender um tipo de regularidade e utilizá-la para avaliar situações novas.

Na realidade, é justamente a capacidade de interagir com situações imprevistas que constitui o argumento mais forte em favor das abordagens probabilísticas em ciência cognitiva. Embora a argumentação relativa à criatividade da língua tenha sido usada para fundamentar a rejeição dos métodos estatísticos em lingüística, os modelos de gramática construídos com base em um corpus de grande porte são considerados mais "robustos" do que as gramáticas em que todas as sentenças gramaticais de uma língua são previamente especificadas. A capacidade "gerativa" destas gramáticas deterministas baseadas em regras permanece uma possibilidade obscura, pelo menos em termos de implementação computacional capaz de lidar com situações reais de uso da língua.

Não resta dúvida que os modelos probabilísticos criticados por Chomsky no final da década de cinqüenta eram excessivamente simples. Neste aspecto, o computador, sobretudo após sua popularização na década de oitenta, desempenha um papel essencial. Hoje em dia, um pesquisador equipado com um computador pessoal comum pode construir um modelo probabilístico com base em um corpus de grande porte, já que as capacidades de armazenamento e processamento de máquinas de custo módico, disponíveis em muitas instituições de ensino e pesquisa, são perfeitamente suficientes. Podem ser coletados e armazenados corpora de dimensões adequadas, de modo a consti-

tuírem uma amostra representativa da língua, dialeto ou subconjunto da língua que se pretende estudar. Deste modo, as noções de **população** e **amostra**, conforme empregadas em estatística e teoria da probabilidade, podem ser utilizadas para a análise lingüística.

Uma vez considerado como uma amostra da população de textos existentes, o corpus pode servir de base para a extração de probabilidades de ocorrência de estruturas sintagmáticas e também de probabilidades de uso de palavras e enunciados em sentidos específicos, tendo como ponto de partida a distribuição de contextos em que estas acepções aparecem em um corpus. A literatura relativa à especificação de estimativas de probabilidades de ocorrência de eventos que não fazem parte de uma amostra é vasta. Os métodos já foram amplamente testados em numerosos campos do conhecimento. O argumento de que todas as sentenças (ou usos, no caso da polissemia) que não fazem parte da amostra serão tratados da mesma maneira em um modelo probabilístico não é verdadeiro.

3. Abordagens estatísticas do léxico

O tratamento do léxico com base em métodos estatísticos pode ser realizado de uma variedade de maneiras. Este artigo não tem pretensão de apresentar uma visão mesmo parcial das muitas possibilidades metodológicas, muito menos uma descrição exaustiva destas possibilidades (ver, por exemplo, Manning e Schülze (1999), para uma análise mais detalhada). Não obstante, algumas características básicas destes tratamentos serão discutidas nesta seção. Isto não significa que não existam outras formas de levar adiante com êxito análises estatísticas de informações lexicais. O panorama breve aqui apresentado visa apenas a dar uma idéia de como estas abordagens são utilizadas em lexicografia e em tecnologia da linguagem humana de um modo geral, sobretudo na desambiguação de sentidos de palavras polissêmicas.

3.1. Um corpus e um computador

O ponto de partida da equipe de projetistas de sistemas que utiliza métodos estatísticos é o corpus computadorizado e, naturalmente, o computador, usado, nestas abordagens, como ferramenta de análise lingüística. Ao invés de criar rotinas de processamento para consulta a um dicionário eletrônico, a fim de sistematizar as informações lexicais relacionadas a um item, permitindo assim a interpretação semântica de um enunciado qualquer onde este item lexical aparece, o projetista começa seu trabalho com a análise das informações contidas no corpus. Este trabalho de coleta dos dados de uso de uma palavra tem, geralmente, como ponto de partida, uma concordância desta palavra.

Uma concordância é uma lista de exemplos de uma determinada palavra, expressão ou morfema, apresentados no contexto em que ocorreram em um corpus. Pode-se obter uma concordância usando um programa concordanceador, tal como o WordSmith (Scott, 1998), ou, no caso de usuários do Linux, aproveitar os recursos do sistema operacional, sobretudo o comando *grep*, para gerar concordâncias. Existem vários outros recursos computacionais disponíveis no mercado para a manipulação de corpora. Alguns destes software são gratuitos ou de custo relativamente baixo. Abaixo temos um pequeno fragmento, composto por onze ocorrências, de uma concordância da palavra *peça* extraída do corpus do NILC, composta por 3685 ocorrências.

1. Tony Meola é **peça** decisiva nos planos de Bora Milutinovic.
2. A polícia italiana recuperou esta semana uma das **peças** arqueológicas mais
3. Nelas, aparece uma **peça**, provavelmente do aerofólio, que se desprende do
4. dúvida a respeito da autenticidade da **peça** acusatória, não acarreta nulidade
5. fiscalistas a ponto de ficar contando **peça** por **peça** bonecos, tênis ou camisetas

6. Se preciso, **peça** a um amigo que o amarre ao tronco de uma árvore no Ibirapuera.
7. A obesidade é a última **peça** do dominó que cai", diz o médico.
8. propagou-se por toda a Europa e tornou-se **peça** de resistência em todas as mesas.
9. a própria miséria parece hoje **peça** de uma engrenagem voltada à destruição do
10. Michelle Matalon, 34, atriz e produtora da **peça** "Pentesiléias" (atualmente
11. uma câmara fria para testar os componentes destes caminhões. As **peças** são expostas

As ocorrências foram selecionadas, de modo a cobrir um espectro razoavelmente amplo da polissemia da palavra, dentre os primeiros cem casos apresentados na lista gerada pelo WordSmith. Uma vez que não foi feita uma análise completa das ocorrências de **peça** encontradas no corpus, não é possível tirar conclusões abrangentes. Porém, o propósito da discussão neste artigo é sobretudo de natureza metodológica. A análise da sexta ocorrência acima permite a detecção de um caso de homonímia, uma vez que trata-se de uma forma verbal. Pressupondo um corpus etiquetado segundo categorias morfossintáticas, esta ocorrência seria retirada automaticamente do corpus de treinamento, assim como todas as outras que fossem classificadas como formas verbais. A pressuposição é razoável, já que existem etiquetadores bastante eficientes para a língua portuguesa (ver, por exemplo, Bick 1996).

A partir daí, o processo de desambiguação automática torna-se bastante mais complexo, uma vez que a distinção em termos de classe de palavra não pode ser mais utilizada como critério. Todas as ocorrências de **peça** são substantivos. A construção de um corpus de treinamento necessita agora do trabalho paciente de anotação de cada ocorrência segundo um conjunto de sentidos possíveis. A especificação deste conjunto de sentidos, conforme mencionado anteriormente, pode ser feita

com base em trabalhos de referência pré-existentes, como, por exemplo, um ou mais dicionários. Entretanto, é muito provável que a palavra em questão apresente sentidos, em contextos de uso registrados no corpus, que não estejam registrados nos trabalhos de referência.

Se fosse usado o Dicionário Caldas Aulete (1980) como referência, os analistas participantes da construção do corpus de treinamento teriam um problema inicial de considerável complexidade para resolver. Neste dicionário, a metodologia utilizada na organização dos verbetes procura fornecer ao consulente um sentido genérico da palavra definida, seguida por acepções mais específicas. No caso de **peça**, o verbete começa com a definição *parte de um todo*, seguida do exemplo *Uma peça de carne*. O leitor não consegue ter certeza quanto a que todo se refere o dicionário, já que *carne* não pode ser o *todo* em questão. Há duas maneiras de interpretar a definição apresentada pelo dicionário.

A primeira delas baseia-se na definição, em detrimento do exemplo. O sentido *parte de um todo* pode ser considerado como o sentido básico. Ao analisar as onze ocorrências acima, este sentido básico parece claro em apenas três delas, de números 3, 9 e 11, já que *aerofólio*, *engrenagem* e *caminhões* são o todo do qual **peça** é uma parte. É possível argumentar que a ocorrência 7 poderia ser incluída como também adequadamente definida como *parte de um todo*, que seria o conjunto das peças de um jogo de dominó, mas sem dúvida as peças de um jogo de dominó não se organizam em um *todo* da mesma maneira que as peças de um caminhão. Claramente, as peças têm que ser organizadas de uma única forma, com grau de variação mínimo ou, praticamente, nenhum, para que o todo resultante possa ser chamado de *caminhão* e desempenhe as funções de um caminhão, enquanto as peças de um jogo de dominó podem ser organizadas de muitas maneiras diferentes sem desrespeitar as regras do jogo de dominó.

Este tipo de distinção sinaliza que a definição com base em um sentido genérico pode ser útil como uma espécie de recurso último para a interpretação de acepções raras ou, pelo menos, inusitadas,

mas que acepções mais freqüentes serão mais facilmente processadas, tanto do ponto de vista computacional quanto cognitivo, se padronizadas em termos mais específicos, como, por exemplo, *autopeças* e *peças de um jogo*. Deste modo, a interpretação semântica acontece diretamente a partir dos elementos lexicais e sintáticos presentes no contexto imediato da ocorrência, ao invés de a partir de um sentido genérico que, na realidade, não é suficiente para lidar com as situações de uso mais freqüentes. Embora o verbete do Caldas Aulete caminhe do sentido mais genérico para o mais específico, parece mais adequado caminhar dos sentidos mais específicos - associados a outros itens lexicais do contexto imediato, sob a forma de colocações, ou a estruturas sintáticas típicas - até chegar ao mais genérico, que ficaria restrito às interpretações de ocorrências em que todas as demais acepções específicas falhassem, isto é, que não pudessem ser resolvidas com base no corpus de treinamento.

Antes de prosseguir no detalhamento da metodologia de construção de um corpus de treinamento, vale a pena discutir a segunda interpretação possível do sentido genérico apresentado no início da definição do Caldas Aulete, a qual se baseia no exemplo, *uma peça de carne*, associado, estranhamente, à definição textual *parte de um todo*. O exemplo adequa-se melhor a uma definição como *unidade de um determinado tipo*. Este tipo de matéria, material ou objeto encontra-se normalmente especificado no contexto imediato, seja explicitamente por meio de um adjetivo ou sintagma preposicional, seja implicitamente por meio de inferência com base em outras informações do texto. Estas informações têm que estar explicitadas em formas literais presentes no texto, ainda que a noção de contexto imediato precise ser expandida de modo a incluir itens lexicais que não estejam em posição adjacente à palavra analisada.

Esta segunda interpretação sugere que, na verdade, os lexicógrafos do Caldas Aulete procuraram fundir dois sentidos genéricos que não são idênticos, embora se sobreponham em alguns usos, em uma única defi-

nição inicial. Desta forma, parece razoável concluir que tanto uma *auto-peça* quanto uma *peça de jogo* são, concomitantemente, *parte de um todo* e uma *unidade de um determinado tipo*, se as acepções forem consideradas isoladamente. Porém, em contextos específicos, um dos sentidos tende a preponderar. Por exemplo, na ocorrência 11 acima, o sentido de *componente*, e, portanto, *parte de um todo*, prepondera, uma vez que o tipo, *peça de caminhão*, já se encontra definido anteriormente no texto. Em outras ocorrências, como no exemplo da *peça de carne* do Caldas Aulete, parece difícil imaginar o sentido de *parte de um todo*. A **peça** poderia ser, por exemplo, *parte de uma rês*, mas não é muito útil, do ponto de vista de um usuário de dicionário, assim como para projetistas de sistemas tecnológicos que processem linguagem humana, compreender o exemplo como *parte de uma carne*, em que esta *carne* é uma rês qualquer.

A continuação do verbete procura detalhar as acepções, utilizando a convenção característica do Caldas Aulete, a qual separa as acepções por meio de duas barras verticais, sem uso de numeração. Em seguida à definição genérica, que parece misturar dois sentidos distintos, o verbete concentra-se na acepção *parte de um todo*, definindo **peça** como *cada uma das partes ou elementos de uma coleção ou conjunto, considerada como um todo*. Seguem-se exemplos como *peças de um trajo*, *peças de um serviço de chá* e *peças de um relógio*. A próxima acepção diz respeito às *peças de um jogo*, e a subseqüente refere-se ao sentido de **peça** como *subdivisão de uma casa*. Todas estas acepções exploram a noção de *parte de um todo*, mas, conforme observado anteriormente, é possível imaginar contextos em que o sentido *unidade de um determinado tipo* preponderasse para os mesmos exemplos específicos.

As acepções que se seguem parecem explorar sentidos mais claramente relacionados à noção de *unidade de um tipo*, como **peça** no sentido de *antiga moeda de ouro*, *jóia*, *peça de linho* ou *de pano*, de um modo geral, *peça de ourivesaria*, *peça de arquitetura*, até chegar à *qualquer parte componente de um mecanismo*, e, portanto, retornar ao sentido genérico de *parte de um todo*. A seqüência se fecha com *qualquer composição literá-*

ria especialmente dramática. Daí em diante, o verbete procura listar expressões cristalizadas associadas a colocações específicas, como *engano* ou *logro*, quase sempre objeto de verbos como *pregar* ou *armar*, e mais uma variedade de acepções, algumas regionais, outras nitidamente dependentes das colocações de que fazem parte para sua interpretação semântica, como *peça por peça* e *peça de resistência.* Não parece haver nenhuma sistematização na ordem aparente, ou, pelo menos, não há maiores esclarecimentos quanto às razões que levaram à escolha da ordem de apresentação das acepções no verbete.

Deste modo, é muito difícil, seja para um usuário do dicionário, seja para um sistema projetado para realizar interpretação semântica automaticamente, utilizar as informações contidas em um verbete de dicionário para resolver problemas complexos de polissemia, isto é, decidir, dentre as muitas possibilidades de sentido para uma palavra como **peça**, qual é a acepção adequada para a ocorrência a ser interpretada em um contexto real de uso. O uso de um modelo probabilístico, construído a partir de um corpus representativo de uma língua ou sublíngua usado como corpus de treinamento, constitui-se, conseqüentemente, em uma alternativa viável para a solução de um problema real.

O ponto de partida para a anotação das ocorrências encontradas no corpus de treinamento é a lista de sentidos possíveis encontrados em trabalhos de referência, mas é quase certo, conforme é possível depreender dos exemplos acima, que será necessário acrescentar novas acepções à lista inicial, uma vez que os dicionários de língua portuguesa não são construídos com base em informações de um corpus. Além disso, ainda conservam alguns dos vícios da gramática tradicional em relação à subordinação da língua falada à escrita e à referência aos clássicos como base da pesquisa realizada para a definição dos usos possíveis de uma palavra, sem nenhuma preocupação com freqüência ou contexto de uso.

A manipulação deste corpus de treinamento, de modo a permitir a construção deste modelo probabilístico dentro de limitações razoáveis de tempo e recursos humanos e financeiros, só é possível com a

utilização sistemática de um computador. A ferramenta tecnológica, nesta abordagem, passa a ser parte integrante da metodologia de análise lingüística, e não apenas um instrumento circunstancialmente utilizado para testar concepções pré-estabelecidas. Pressupondo um corpus de treinamento manualmente adotado, a construção do modelo probabilístico será detalhada na próxima subseção.

3.2. Um modelo probabilístico

Prosseguindo com o exemplo apresentado na subseção anterior, procurar-se-á apresentar abaixo o processo analítico de construção de um modelo probabilístico para um determinado item lexical, no caso a palavra **peça**. A despeito do nome de conotação matemática, o processo analítico em questão é eminentemente lingüístico. O aspecto probabilístico em questão é relativamente simples. A partir de uma amostra suficientemente grande de ocorrências do item lexical em questão, compara-se a freqüência de uma determinada acepção com elementos do contexto imediato, incluindo aspectos lexicais e sintáticos, dentro de uma concepção léxico-gramática. Com base nestas freqüências, são especificadas probabilidades para as diferentes acepções possíveis, uma vez que estejam presentes determinadas características do contexto imediato. Esta especificação baseia-se na análise do corpus de treinamento.

O caso típico e mais evidente de uma associação entre uma acepção e um contexto imediato é a colocação. Na sua definição mais simples, uma colocação é um padrão de co-ocorrência de palavras. No pequeno fragmento de concordância apresentado na subseção anterior, há pelo menos dois casos de co-ocorrência fáceis de detectar, *peça por peça* e *peça de resistência*. Em seu trabalho sobre colocações, Kjellmer (1991) sustenta que o léxico mental humano não é constituído apenas de palavras isoladas, mas também de unidades perifrásticas maiores, as quais tanto podem ser fixas quanto suscetíveis a maior ou menor grau de variação. A noção em si é amplamente reconhecida em psicolingüística e também em lexicografia, mas

sua utilização sistemática como fundamento de análise nem sempre é levada a cabo.

No Caldas Aulete, *peça por peça* é uma das expressões cristalizadas listadas. O sentido a ela associado é de *separadamente, por miúdo, aos bocados, cada coisa de per si*. A definição da acepção é adequada, a despeito da linguagem erudita usada no verbete, na qual poderia ter sido incluída, para simplificar, a expressão *um por um*, já que parece ser a que melhor reproduz o sentido de *peça por peça*, pelo menos se a ocorrência número 5 acima for tomada como parâmetro. Na metodologia da construção do modelo probabilístico, o analista normalmente faz uma busca específica da expressão *peça por peça* no corpus, o que não é problema para a maioria dos programas concordanceadores, nem para os recursos do sistema operacional Linux.

A busca revela alguns aspectos importantes do trabalho de análise lingüística a partir de corpus. Há apenas duas ocorrências da expressão em trinta e dois milhões de palavras. Uma delas é a mostrada acima. Logo, parece necessário expandir a busca para um outro corpus de modo a aumentar o tamanho da amostra e garantir a validade das conclusões. O aumento das dimensões da amostra torna-se ainda mais importante porque a segunda ocorrência apresenta uma ligeira variação de sentido que pode ser decisiva para uma interpretação semântica correta. Abaixo está o exemplo em questão, extraído da edição da Folha de São Paulo de 9 de abril de 1994, seção de esportes.

(8) Porque o ataque foi concebido, peça por peça, em função dos grandes atacantes existentes naquele momento.

À primeira vista, parece que as características do verbo, ou, talvez, do objeto do verbo, ao qual a expressão está vinculada podem fazer com que ocorra uma alteração da interpretação semântica. Na ocorrência número 5 do exemplo de concordância, o objeto de *ficar contando* - isto é, *bonecos, tênis ou camisetas* – denota as peças que seriam contadas uma

a uma, separadamente. Através dos sintagmas nominais plurais, o texto permite que o leitor saiba a que tipo de peça a expressão se refere. Já no exemplo (8), a expressão refere-se ao sujeito da passiva do verbo *conceber*, portanto, em última análise, também ao objeto do verbo. Porém, a palavra *ataque* refere-se a um conjunto composto por peças, os *atacantes*. Parece persistir, mesmo nas expressões cristalizadas, uma variante da distinção entre *parte de um todo* e *unidades de um determinado tipo*, a despeito da estrutura sintagmática. O objeto do verbo modificado pela expressão tanto pode ser interpretado como *o conjunto do qual fazem parte as peças* quanto como o *tipo de peças* a que se refere o texto.

As características do sintagma nominal objeto do verbo a que se liga a expressão apontam para algumas opções de análise que podem ser integradas ao modelo probabilístico para o tratamento da polissemia de **peça por peça**. A seqüência de sintagmas nominais plurais que constitui o objeto de *ficar contando* indica que a interpretação correta da expressão, nesta sentença, é *contar uma por uma peças do tipo definido pelo objeto do verbo*. O objeto formado por um sintagma nominal singular indica uma probabilidade maior de interpretação como *conceber uma a uma as peças do conjunto definido pelo objeto do verbo*. Esta interpretação torna-se muito mais provável se, dentro de uma janela de contexto limitada, for encontrado item lexical que possa ser interpretado como *componentes do conjunto definido pelo objeto do verbo*, como é o caso de *atacantes*. Porém, tendo em vista o pequeno número de ocorrências, uma para cada interpretação, não é possível definir probabilidades com base nestas características do contexto imediato.

A interpretação semântica do verbo talvez seja um caminho mais promissor para a definição do sentido exato de **peça por peça**. Uma vez que o verbo *contar* tenha seu sentido especificado como *enumerar*, parece seguro definir o sentido de **peça por peça** como se referindo a *peças do tipo definido pelo objeto do verbo*. Esta informação, porém, teria, mais uma vez, que ser verificada em pelo menos algumas outras ocorrências antes de ser incorporada ao modelo probabilístico. Trata-

se de um problema de tamanho da amostra que, a despeito das dimensões aparentemente muito grandes de um corpus, tende a ocorrer à medida em que se procura especificar o uso de palavras ou expressões menos freqüentes. A solução adequada é buscar, em outros corpora, mais informações sobre o uso do item em questão. Neste artigo, contudo, a investigação sobre a expressão **peça por peça** será interrompida neste ponto, uma vez que a preocupação do trabalho é de natureza fundamentalmente metodológica.

A segunda expressão cristalizada detectada no fragmento de concordância é **peça de resistência**. Segundo o Caldas Aulete, a expressão **peça de resistência** é usada para se referir a *peça musical que apresenta muitas dificuldades e pode dar a medida do artista* ou a *o forte (aquilo em que uma pessoa excele)*. A segunda definição aproxima-se do sentido observado na ocorrência 8, mas ainda não constitui interpretação suficientemente exata. O sentido da expressão não está relacionado a uma pessoa, mas sim ao sintagma *todas as mesas*, no caso européias, sendo a palavra *mesas* interpretada na acepção de *hábitos alimentares*. Parece mais apropriado interpretar a expressão no sentido de *parte integrante* ou mesmo *parte fundamental*. Talvez o sentido do sintagma nominal seja melhor parafraseado pelo adjetivo *básico*.

Uma busca no corpus produz, desta vez, quatorze ocorrências, um número bem maior que o obtido para a expressão anterior. Três delas dizem respeito ao mesmo texto, justamente aquele da qual foi extraída a ocorrência 8. As duas primeiras ocorrências são repetições do mesmo título, *batata é peça de resistência*, o qual aparece pela primeira vez em um índice, com referência de página, e, posteriormente, como título do artigo em questão. Uma vez que, neste caso, o título define inequivocamente o tópico de que trata o artigo, os leitores já sabem, de saída, que a peça de resistência em questão é um objeto descrito pelo sintagma nominal *batata*. O vegetal e sua utilização como alimento são suficientemente disseminados para dispensar uma definição específica do objeto no texto. Um sistema programado para rea-

lizar a desambiguação de sentidos poderia fazer uso deste tipo de informação contida em títulos, mas isto não será discutido aqui.

As onze ocorrências restantes, no entanto, demonstram que o sentido de *ponto forte* prepondera, uma vez que é a melhor interpretação em oito ocorrências. Apenas uma outra ocorrência parece ter o sentido de *ponto básico*, pelo menos na interpretação do autor deste artigo. As três ocorrências restantes são de difícil interpretação e exemplificam bem algumas das dificuldades típicas do tratamento da polissemia, tanto do ponto de vista cognitivo quanto do computacional. Nos três casos, há duas interpretações aceitáveis e não parece haver meio de levar a cabo a desambiguação de maneira categórica. Os exemplos em questão são mostrados abaixo.

(9) A NWS tem programa leve. Abertura "Leonora no.3", de Beethoven, e o "Concerto para Violino em Ré Maior Op.35", de Tchaikovsky (solista Robert McDuffie), são os destaques da primeira noite. A "Sinfonia no.2", de Brahms, é a **peça de resistência** da segunda.

(10) A **peça de resistência** dos designers, a cadeira, é o tema da exposição que abriu segunda-feira no museu da casa brasileira.

(11) Folha: E a montagem de Beckett?
Fernanda: É uma **peça de resistência**, que mostra o ser humano sendo enterrado, a terra tapando-lhe o nariz, e ele cantando. É aquilo de não sabe pra que vive, mas vive.

O exemplo (9) parece, em uma primeira análise, encaixar-se no sentido diretamente relacionado ao contexto musical mencionado no Caldas Aulete. Porém, uma análise do sentido das expressões *programa leve* e *destaques da primeira noite* acaba por sugerir que peça de resistência na verdade significa o *forte* da segunda noite. No exemplo (10), o sentido de *básico*, ou, talvez melhor, *peça básica*, parece o mais ade-

quado. A interpretação em que a cadeira seria considerada como o *forte* dos designers parece um tanto inadequada, embora não totalmente equivocada. No exemplo (11), a forma de interpretação correta parece ser a literal, isto é, uma **peça** (de teatro) que trata do tema da resistência, e não a acepção da expressão cristalizada, o que pode parecer surpreendente, mas é a melhor solução.

Se forem aceitas as interpretações de *forte* para o exemplo (9) e *básico* para o exemplo (10), obtém-se um total de quatorze ocorrências classificadas em oito casos de acepção *forte*, cinco casos de acepção *básico* e uma interpretação literal, sem tratamento como colocação. Esta última pode ser separada das demais por algumas características do contexto imediato. É a única ocorrência de **peça de resistência** precedida de artigo indefinido e com sentido de **peça de teatro**. O reconhecimento da acepção depende de informações enciclopédicas, como o autor *Beckett* na pergunta precedente, além da desambiguação do sentido de *montagem* e do processamento sintático, que deve detectar a ausência de um sintagma preposicional modificador de **peça de resistência**, ao contrário do que ocorre em exemplos como **peça de resistência do filme** e **peça de resistência da candidatura**.

Todos estes aspectos do processamento apresentam problemas próprios, que ressaltam a importância e a dificuldade de integrar todas estas informações de modo a realizar a desambiguação de sentido com êxito. Na entrevista de Fernanda Montenegro à Folha, de onde o exemplo (11) foi extraído, o tópico global *teatro* precisa estar ativado e integrado ao processamento para que a interpretação de *montagem de Beckett* possa ser utilizada na interpretação semântica. O processamento sintático, da mesma forma, tem que fazer frente a dificuldades como a recuperação de elementos omitidos em casos de elipse e o processamento diferenciado de títulos em relação ao texto corrente. Não obstante, a identificação destas características do contexto imediato é de suma importância para a construção do modelo probabilístico, uma vez que a probabilidade de interpretar **peça**, em **peça de**

resistência, como uma *peça de teatro que tem como tema a resistência* fica associada à presença destas características.

A distinção entre o sentido de *básico* e a acepção de *forte* é difícil de modelar. Quando a *peça de resistência* em questão é parte ou membro de uma lista explicitamente apresentada, ou facilmente inferida, o sentido *forte* parece predominar, sobretudo quando a palavra *parte* está associada ao que é definido como **peça de resistência**. Quando a relação é de natureza mais abstrata, como no caso da *batata* em relação aos hábitos alimentares dos europeus, o sentido de *básico* prepondera. Porém, há casos em que é difícil ter certeza quanto a elementos do contexto que de fato indiquem uma probabilidade maior de um sentido sobre o outro, sobretudo, como seria de se esperar, naqueles casos em que a distinção do sentido, pelo analista, é difícil. Isto sugere que talvez a distinção possa ser abandonada sem prejuízo da interpretação semântica.

Uma vez sistematizado o tratamento das expressões cristalizadas, o próximo passo é procurar definir características do contexto imediato, conforme detectadas no corpus de treinamento, que possam ajudar na escolha do sentido correto, em situações novas, com base nas probabilidades de um determinado sentido ser o adequado, uma vez que uma ou mais destas características estejam presentes. Em última análise, a metodologia é semelhante à utilizada para as expressões cristalizadas, mas os padrões associados aos sentidos são mais variados, e as definições precisam ser mais flexíveis. O analista que procura definir estas características a partir do corpus de treinamento conta com vários recursos computacionais para obter estas informações sem ter que procurá-las manualmente.

A ocorrência 1 do fragmento de concordância acima apresenta pelo menos duas características dignas de atenção. Uma é o sujeito *Tony Meola*, um indivíduo, e a segunda é o adjetivo *decisiva*, associado à **peça**. A primeira é mais difícil de utilizar para fins de análise semântica, mas o adjetivo associado à **peça** na posição imediatamente

adjacente à esquerda tem boas possibilidades de abrir perspectivas de sistematização do tratamento da polissemia de **peça**. Uma maneira comum de iniciar este tipo de investigação é usar os recursos de um programa de manipulação de corpus para verificar que palavras co-ocorrem em número significativo com **peça** nas 3685 ocorrências detectadas no corpus. A janela de contexto utilizada na análise discutida a seguir é de cinco palavras à esquerda e cinco à direita, mas pode ser de até cinqüenta palavras (ver Pedersen 1999).

Conforme esperado, alguns adjetivos pertencentes a um mesmo campo semântico se destacam na quantificação dos "colocados", isto é, das palavras que aparecem nas colocações que incluem **peça**. A combinação *peça decisiva* aparece quatro vezes. Em todas as quatro vezes, é predicativo do sujeito, precedida de um sujeito que é um indivíduo e verbo *ser* como verbo de ligação. Portanto, este padrão é real e pode ser explorado na análise da polissemia de **peça**, já que em todos os quatro casos o sentido é de *parte de um todo*, tipicamente uma equipe esportiva (caso da ocorrência *Tony Meola*) ou grupo que age para obter alguma coisa (caso de uma ocorrência em que a *peça decisiva* é *Suruagy*). Há uma quinta ocorrência em que decisiva aparece na posição direita-4, isto é, há três palavras entre a ocorrência de **peça** e a ocorrência de *decisiva*. Estas ocorrências não adjacentes podem ser úteis na construção do modelo, mas não serão discutidas aqui.

Por ordem de freqüência, o adjetivo *especial* é o que mais vezes ocorre na janela de contexto especificada para esta análise, com 81 ocorrências. Porém, nenhuma delas aparece na posição direita-1, que caracteriza o padrão antes detectado. Embora, numa análise completa das ocorrências de um corpus de treinamento, este tipo de co-ocorrência deva ser investigado, parece improvável obter resultados significativos, em termos do padrão investigado, com este adjetivo, e é muito provável que um número grande destas ocorrências sejam conseqüência de coincidências ou de padrões diferentes não relacionados à polissemia de **peça**. A maioria das ocorrências está a esquerda de peça

(70), e apenas onze à direita. Além disso, não há nenhuma ocorrência em esquerda-1 ou direita-1, e há 57 ocorrências nas posições esquerda-5 (17), esquerda-4 (19) e esquerda-3 (21). A distribuição não caracteriza um padrão de co-ocorrência com a palavra estudada.

O próximo adjetivo da lista, em termos de freqüência, é *importante*, logo seguido de *fundamental*. Há 35 casos de co-ocorrência de *importante* com **peça**, sendo que 22 estão na posição direita-1, 7 em direitá-2 e duas em direita-3, totalizando 31 casos. Como a presença de modificadores, sobretudo intensificadores, é comum neste tipo de combinação, (*peça muito importante, peça mais importante, peça muito mais importante*), o padrão está bem mais claro aqui. Quanto a *fundamental*, há 33 co-ocorrências, das quais 32 estão em direita-1 e apenas uma em direita-2, mostrando que o adjetivo não admite intensificadores normalmente. O padrão, não obstante, está claramente delineado. Da mesma forma, há 27 co-ocorrências de **peça** e *central*, sendo que, em 20 delas, *central* está em direita-1.

A discussão concentra-se nas ocorrências destes adjetivos em posição direita-1, uma vez que as demais exigiriam análises excessivamente longas. Em relação à **peça** *importante*, doze das vinte e duas ocorrências na posição direita-1 são predicativos do sujeito onde a **peça** *importante* em questão é o sujeito da oração, no mesmo padrão de *Tony Meola é uma peça decisiva* apresentado acima. Dois destes casos incluem a palavra **outra,** ao invés de artigo, ou não há determinante. A enorme maioria destes casos (11 em 12) usa o verbo *ser* como verbo de ligação. O sujeito da oração nem sempre é um indivíduo ou pessoa, mas pode ser uma instituição (*a universidade contemporânea*) ou um conceito (*a neutralização*). A relação entre o padrão sintático e a desambiguação do sentido de **peça** não se altera, no entanto, sobretudo se considerarmos o sintagma preposicional subseqüente - geralmente iniciado por *de, em* ou uma de suas contrações - que define o conjunto do qual a **peça** faz parte.

Antes de procurar definir mais claramente o padrão de co-ocorrência em questão, vale a pena explorar mais detidamente os dados

coletados no corpus. As dez ocorrências de **peça** *importante* que não são predicativos do sujeito estão distribuídas da seguinte maneira: quatro casos de adjunto adverbial ligado pela preposição *como* às formas verbais *desponta*, *continua*, *foi recebido* e *adquirido*, tal como em *Mercadante desponta como peça importante do "novo" PT*; três ocorrências na função de aposto, em que a peça importante em questão é o sintagma nominal ao qual está ligado, tal como em *Pat Ewing, por exemplo, peça importante do time, por pouco não comprometeu a vitória...*; duas ocorrências em que **peça** *importante* aparece como sujeito da passiva, tal como em *...outra peça importante de seu programa de governo, a reforma do sistema nacional de saúde, seja aprovada*, na qual a *peça importante* está definida em um aposto; e um caso de objeto da passiva, tal como em *Malan é considerado ainda uma peça importante no Banco Central...*, no qual a **peça** *importante* é o sujeito da oração na voz passiva.

A despeito das variações na estrutura sintática, todos estes casos guardam uma semelhança com a estrutura da oração sujeito-verbo de ligação-predicativo do sujeito. Ainda mais importante, a estrutura do sintagma nominal que contém a palavra **peça** e o adjetivo do campo semântico em questão inclui, na enorme maioria das vezes, um sintagma preposicional que descreve o conjunto a que pertence a peça qualificada como *decisiva* ou *importante*. Isto torna a desambiguação não apenas possível, mas bastante segura, identificando a ocorrência como *parte de um conjunto* e não *unidade de um tipo*. No caso de *peça decisiva*, a taxa é de cem por cento dos quatro casos registrados. No caso de *peça importante*, não é verdade em dois casos dos vinte e dois encontrados. Um destes casos apresenta um problema de processamento textual complexo, mas o segundo é uma variação simples da estrutura do sintagma nominal com uso do locativo *ali*, que faz referência ao conjunto em questão. Nestes dois casos, peça importante aparece como sujeito da passiva, no mais simples, e como aposto a *Lei de Defesa do Consumidor*, no mais complexo. Isto parece indicar que os casos de predicativo do sujeito são mais seguros.

Resta examinar as ocorrências de *peça fundamental* e *peça central*. Nas 32 ocorrências da primeira colocação, o padrão em que a colocação é seguida de um sintagma preposicional que define o conjunto aparece em 30. Uma das duas exceções não constitui, na verdade, quebra do padrão, já que o sintagma preposicional está anteposto e entre vírgulas[8]. Poderia ser dito, portanto, que o padrão se confirma em 31 dos 32 casos. Na segunda exceção, **peça** *fundamental* é objeto direto do verbo *usar* e seguida de dois pontos, um caso único[9]. A identificação da parte de conjunto é feita após os dois pontos, mas o conjunto em si a que pertence tem que ser depreendido do texto precedente. A ausência de exceções nos casos em que a colocação tem função de predicativo do sujeito, os quais constituem 22 das 32 ocorrências, continua sendo confirmada. As demais ocorrências estão distribuídas em: quatro apostos ao elemento que é a peça fundamental; três adjuntos adverbiais, dois ligados por *como* e um ligado por *em*; dois outros objetos diretos, dos verbos *esquecer* e *constituir*, totalizando três casos, e, assim, os dez restantes.

As ocorrências da colocação **peça** *central*, em número de 20, preservam o mesmo padrão das colocações anteriormente discutidas em dezoito dos casos, isto é, a colocação é seguida por um sintagma preposicional que define o conjunto do qual a peça em questão faz parte, caracterizando a acepção. As duas exceções definem um segundo padrão, que é o da oração relativa, na qual o antecedente do pronome relativo é o conjunto, como *em quebra-cabeças que talvez não tenha peça central* e em *processo sucessório cuja peça central continua a ser Getúlio Vargas*. Esta última ocorrência também é um caso de oração que contém uma locução verbal que pode ser caracterizada como verbo de ligação, e,

8 ...nem por isso ficará dispensada a perícia, que, na ação de nunciação, é peça fundamental.

9 Também usa uma peça fundamental: o "foley artist", profissional que "interpreta" os barulhos no ritmo das cenas

portanto, parte integrante do outro padrão definido anteriormente. Porém, como nesta mesma ocorrência, a colocação *peça central* aparece como sujeito em sete dos 14 casos em que a colocação aparece em estruturas com verbo de ligação. A variação não acontece nas demais colocações discutidas. Não resta dúvida, no entretanto, que a distinção entre sujeito e predicativo do sujeito depende fundamentalmente da ordem das palavras, a qual não tem, à diferença dos predicados verbais, onde há objeto, influência na interpretação semântica.

As demais ocorrências do fragmento de concordância apresentam mais dois casos em que **peça** vem seguida de adjetivo, as de número 2 e 4. Na número 2, o sentido de **peça** é de unidade de um tipo. Os adjetivos que aparecem adjacentes à **peça** quando usada neste sentido podem ser rapidamente compilados via programa de manipulação de corpora, uma vez que as acepções tenham sido previamente anotadas. Não será feita análise detalhada destas ocorrências, devido a limitações de espaço, mas parece provável que um número significativo de adjetivos que definem áreas do conhecimento ou classificam obras de arte apareça na posição direita-1. Estes adjetivos podem ser usados para desambiguar ocorrências novas, em um sistema computacional, e também em uma léxico-gramática que fundamente uma teoria lexical de base cognitiva. A ocorrência 4 também faz parte deste grupo de ocorrências na acepção de unidade de um tipo, mas o caráter específico da linguagem jurídica pode determinar um tratamento em separado. A decisão depende do analista e dos interesses específicos do levantamento.

Os usos metafóricos de *peça do dominó* e *peça de uma engrenagem* teriam que ser verificados quanto à freqüência, de modo a determinar se estes usos metafóricos predominam em relação ao uso no sentido literal e que condições do contexto imediato sinalizam no sentido de um ou de outro uso. As ocorrências 3 e 11 apresentam a dificuldade de poderem ser interpretadas tanto na acepção de *parte de um todo* quanto de *unidade de um tipo*. Porém, a presença de uma palavra que define um conjunto no contexto imediato – *aerofólio* e *caminhão*, respectivamente

– pode tornar mais provável a interpretação como *parte de um conjunto*, sobretudo quando usadas com artigo definido, o que, tudo indica, tem influência real. A análise não será integralmente realizada aqui.

Antes de resumir estas informações em um modelo probabilístico único do uso de **peça**, será focalizado o uso no sentido de *peça de teatro*. Esta colocação aparece 56 vezes no corpus, com sentido muito provavelmente invariável de *composição dramática*, ainda que isto não seja garantido, uma vez que não foram todas analisadas. Além disso, como pode ser observado na ocorrência 10, a acepção aparece quando **peça** ocorre antes de um título, marcado por aspas, no caso deste corpus, e, possivelmente, por alguma outra forma de sinalização gráfica nos textos de um modo geral. Uma busca da palavra **peça** seguida de um espaço e um sinal de abre aspas no corpus encontrou 396 ocorrências, praticamente todas elas no sentido de *composição dramática*, embora haja algumas que têm a acepção de *composição musical*. Este tipo de análise a partir dos dados mostra a importância do "diga-me com quem andas e te direi quem és", de Firth (1957), em termos de análise lingüística. A pontuação, os sinais gráficos, os títulos, subtítulos e outros elementos de organização textual influenciam diretamente a interpretação semântica, mas são muitas vezes ignorados como elementos constituintes do sentido.

Ainda dentro da acepção de *composição dramática*, outras "companhias" da palavra **peça** parecem exercer influência determinante, a saber: *elenco da peça*, *peça em cartaz*, *estréia da peça*, *montagem da peça*, *peça de Shakespeare*, *peça de Nelson Rodrigues* e *cenário da peça*. Embora a colocação *peça de teatro* seja em certa medida semelhante, em termos de acepção, à *peça de carne*, isto é, *unidade de um determinado tipo*, aparece muito mais freqüentemente sem o sintagma preposicional que especifica o tipo em questão, uma vez que seu uso está cristalizado. Estes elementos, que também aparecem freqüentemente no contexto imediato quando a acepção é *peça de teatro*, devem ser, portanto, levantados.

Estes elementos aparecem no mínimo sete vezes na posição especificada acima, mas um programa adequado pode detectar a presença

destas palavras em janelas de contexto maiores. O cruzamento destas "companhias" com as informações do corpus de treinamento pré-anotado manualmente em termos de acepções possíveis deve, portanto, resultar em um modelo probabilístico completo, nas bases da amostra resumida apresentada abaixo, que não apenas tem possibilidades promissoras no campo do processamento computacional, mas pode ser capaz de explicar diversos fenômenos do processamento semântico em termos cognitivos. Diante de uma ocorrência da palavra **peça**, tipicamente polissêmica, a ser processada, o caminho poderia ser:

1. Colocação em *peça por peça* ?
    a. Sim: sentido de um(a) por um(a)
        i. Verbo a que se liga ?
        ii. Se for do campo semântico *contar*, a acepção deve ser *unidades do tipo* definido pelo objeto do verbo (sobretudo se SNs plurais) *contadas* (ou semelhante) *uma a uma*
        iii. Se for do campo semântico *conceber*, a acepção deve ser parte do conjunto definido pelo objeto do verbo (ou sujeito da passiva), sobretudo se SN singular, *concebidas* (ou semelhante) *uma a uma*
    b. Não: verificar item 2

2. Colocação em *a peça de resistência* ?
    a. Sim: sentido de *o forte* ou *básico*
        i. Verificar SP adjacente para conjunto que integra
    b. Não: verificar item 3

3. Colocação com adjetivo ?
    a. Sim
        i. Campo semântico (*importante*, *fundamental*)
        ii. SP adjacente define conjunto que integra

b. Sim
   i. Campo semântico (*arqueológico*, *artístico*, *musical*)
   ii. Unidade do tipo definido pelo adjetivo
c. Não: verificar item 4

4. Colocação com SP *de teatro* ?
   a. Sim
      i. Sentido de composição dramática
   b. Não: verificar item 5

5. Seguida de título ?
   a. Sim
      i. Sentido de composição dramática ou musical
   b. Não: verificar item 6

A análise integral do corpus de treinamento acabaria por produzir uma especificação completa do modelo probabilístico nestas bases, isto é, relacionando as ocorrências a seu contexto imediato. Com a utilização de um programa de manipulação de texto mais sofisticado, é possível especificar elementos deste contexto que não estão adjacentes, mas que estão, ainda assim, relacionados às acepções especificadas pelo analista que anota manualmente o corpus de treinamento. Como é fácil concluir, além disso, muitos outros sintagmas preposicionais ligados por *de* a **peça**, como *peça de caminhão* e *peça de reposição*, teriam que ter sua análise detalhada.

Diversos recursos computacionais podem ser utilizados para reduzir a considerável carga de trabalho que este tipo de anotação exige, no caso de utilização em tecnologia da linguagem humana. A experimentação para a verificação de hipóteses de natureza cognitiva, em termos psicolingüísticos, deve ser baseada nas informações extraídas do corpus de treinamento, de modo a permitir uma especificação da influência das "companhias" tanto em aquisição de linguagem quanto

no processamento semântico e textual de adultos. A despeito dos muitos aspectos da construção do modelo probabilístico de peça que foram omitidos, por falta de espaço, espera-se que este tipo de tratamento dos fenômenos polissêmicos tenha se tornado suficientemente claro para o leitor, de modo a tornar-se uma opção metodológica real.

4. Conclusão

Conforme destacado inicialmente, os métodos estatísticos já se encontram bastante disseminados em PLN, a ponto de tornarem-se praticamente matéria de estudo obrigatório para os que se dedicam a pesquisas nesta área do conhecimento. Um dos objetivos deste artigo foi, portanto, procurar recolocar a discussão do uso de métodos probabilísticos no âmbito dos estudos da linguagem em termos atuais, onde o corpus e o computador são parte integrante do processo de investigação. Além disso, foi também levantada a questão da participação da lingüística nos estudos relacionados à ciência cognitiva, onde o debate relativo às capacidades inatas e o raciocínio probabilístico vem sendo redimensionado a partir de concepções como o conexionismo e o processamento paralelo distribuído.

O tratamento da polissemia e do léxico de um modo geral com base em métodos estatísticos, sobretudo em sistemas computacionais construídos com o objetivo de realizar a desambiguação de palavras polissêmicas automaticamente, serviu como referência para a discussão das dificuldades do tratamento da polissemia, enfatizando a construção de modelos probabilísticos com base em corpora de treinamento manualmente anotados. Há muitas outras formas de tratar a polissemia em sistemas computacionais, inclusive a abordagem não-supervisionada, isto é, que não inclui o processamento de anotação manual por analistas humanos. Embora estas metodologias sejam de grande importância, a produção de corpora anotados com informações lingüísticas sempre resulta em um produto útil, uma vez que tenha sido realizada com os devidos critérios.

O exemplo de modelo probabilístico da palavra **peça** apresentado ficou incompleto, devido à complexidade de sua polissemia. Não obstante, espera-se que o objetivo metodológico deste artigo, isto é, demonstrar como este tipo de modelo probabilístico pode ser construído com base em um corpus de grande porte e uso de um computador, tenha sido atingido. Os aspectos técnicos da manipulação de um corpus podem ser facilmente aprendidos, uma vez que os recursos computacionais estejam devidamente instalados. Não se trata de equipamento particularmente poderoso ou caro. Um programa de manipulação de corpus, tal como o WordSmith, e um computador pessoal comum são suficientes para realizar este tipo de análise lexical. A expansão deste tipo de abordagem para todos os níveis de análise lingüística, sem pretensão de eliminar outros enfoques, parece um desenvolvimento desejável para a ciência que estuda a linguagem humana.

5. Referências bibliográficas

ABNEY, S. (1994). Statistical methods and linguistics. IN: Klavens, J. e Resnik, P. *The balancing act: combining symbolic and statistical approaches to language*. Cambridge, Mass.: MIT Press.

BICK, E. (1996). Automatic parsing of Portuguese. IN: *Anais do II encontro para o processamento computacional do português escrito e falado*. Curitiba: SBIA 96.

BRUCE, B.C. (1975). Case systems for natural language. *Artificial Intelligence 6*, 327-360.

CHOMSKY, N. (1957). *Syntactic structures*. Haia: Mouton.

COOPER, R. (1983). *Quantification and syntactic theory*. Dordrecht: D.Rudel.

BUTTON, G., Coulter, J., Lee, J.R., e Sharrock, W. (1997). *Computadores, mente e conduta*. São Paulo: Fundação Editora da UNESP.

CORPUS *Encoding Standard*. http://www.cs.vassar.edu/CES.

CRYSTAL, D. (1995). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

FILLMORE, C.J. (1968) The case for case. IN: Bach, E. e Harms, R. (orgs). *Universals in linguistic theory*. Nova York: Holt, Rinehart and Winston, 1-90.

FILLMORE, C.J. (1977) The case for case reopened. IN: Cole, P. e Sadock, J. (orgs.). *Syntax and semantics 8: Grammatical Relations*. Nova York: Academic Press.

FIRTH, J. (1957). A synopsis of linguistic theory, 1930-55. *Studies in linguistic analysis*. Oxford: Philological Society, 1-32.

GARCIA, H. (1980) *Dicionário contemporâneo da língua portuguesa Caldas Aulete*. Rio de Janeiro: Delta, 3ª. edição brasileira.

GARSIDE, R., Leech, G. e MCENERY, T. (1997). *Corpus annotation*. Londres: Longman.

KJELLMER, G. (1991). A mint of phrases. IN: Aijmer, K. e Altenberg, B. (orgs.), *English corpus linguistics: studies in honour of Jan Svartvik*, Londres: Longman.

LABOV, W. (1970). The study of language in its social context. *Studium Generale*, 23, 30-87.

LINGUISTIC *Annotation*. http://www.ldc.upenn.edu/annotation.

MADDIESON, I (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

MANNING, C. e Schulze, (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

PEDERSEN, T. *A decision tree of bigrams is an accurate predictor of word sense*. http://www.d.umn.edu/~tpedersen.

SAPIR, E. (1921). *Language*. Nova York: Harcourt Brace.

SCOTT, M. (1998) *Wordsmith Tools*. Oxford: Oxford University Press.

SUPPES, P. (1971) Probabilistic grammars for natural languages. *Synthese* 22, 95-116.

WOODS, W.A. (1978) Semantics and quantification in natural language question answering. IN: Yovitz, M. *Advances in computers* (vol. 17). Academic Press.