UM ESTUDO DE EXPRESSÕES CRISTALIZADAS DO TIPO *V+SN* E SUA INCLUSÃO EM UM TRADUTOR AUTOMÁTICO BILÍNGÜE (PORTUGUÊS/ INGLÊS)

Milena Uzeda Garrão
Maria Carmelita Pádua Dias
PUC-Rio

Introdução

A word, in a word, is complicated.
But then what in the world is a word?
(Pinker, 1995:147)

Dentre os variados problemas lingüísticos com os quais um programa de tradução se depara, há uma questão particularmente relevante que é a de reconhecimento e geração de expressões cristalizadas, principalmente de expressões idiomáticas. Isto tem origem, assim nós iremos argumentar, em um dos maiores problemas dentro da teoria lingüística: a questão da delimitação entre sintaxe e léxico no que concerne à definição de itens lexicais com constituição maior do que um vocábulo. Portanto, o objetivo deste estudo foi defender uma visão mais inclusiva de léxico, principalmente no que diz respeito às expressões cristalizadas. Estas são geralmente consideradas como uma exceção ou como simples curiosidades dentro da língua. Por isso, acabam ocupando um papel coadjuvante numa descrição lingüística. Entretanto, como Jackendoff (1997) vai defender para o inglês, e como Gross (1982) já defendia para o francês, estas expressões são muito mais co-

muns do que parecem; conseqüentemente, devem ser equacionadas por um programa de PLN[1].

Programas de tradução automática, em especial, apresentam problemas no tratamento de expressões cristalizadas. A maioria dos sistemas disponíveis no mercado não reconhecem várias ocorrências de grupos de palavras que funcionam como uma unidade. Algumas poucas expressões idiomáticas estão incluídas no léxico de tais sistemas (como "abrir mão"). Outras, porém, de uso freqüente na língua, são tratadas como conjuntos acidentais de palavras, o que resulta em uma tradução ininteligível (como, por exemplo "ter alta", que é vertido para o inglês como "to have high").

Esta pesquisa teve como objetivo prático aventar a possibilidade de o tradutor automático bilíngüe (português-inglês) *Delta Translator*® (1999) vir a traduzir expressões cristalizadas do tipo *V + SN*, particularmente, do tipo *bater + SN*[2].

É importante salientar que não ambicionávamos fornecer conhecimento de mundo para este tradutor automático, visto que um programa de computador nem sempre tem heurísticas a para solução de ambigüidade ou para decidir quais os itens mais adequados para um contexto específico. Isto é, num sistema informatizado é preciso que as descrições sejam as mais detalhadas e refinadas possíveis (Vale, 1998). Nossa proposta foi fornecer tais expressões cristalizadas como uma unidade lexical, e evitar que o programa viesse a traduzi-las como um fragmento sintático.

Como os tradutores automáticos atuais têm a capacidade de incluir informações fornecidas pelo usuário em seu banco de dados, esta proposta não parece pretensiosa, mas compatível tanto com a formatação dos programas de tradução contemporâneos quanto com a necessidade de rever a questão da delimitação lexical. Nossa tentativa evitaria que construções do tipo *bater perna*, uma expressão cristalizada bastante corriqueira no português do Brasil, fosse vertida para o inglês como "*beat leg*", como faz o *Delta Translator*®. Em outras palavras, nossa meta foi possibilitar que este tradutor automático contivesse, em seu banco

de dados, uma construção aparentemente sintática, porém, com um conteúdo lexical.

Para coletar tais expressões cristalizadas utilizamos como fontes os periódicos jornalísticos *Jornal do Brasil*, *O Globo* e *Veja*. Tais fontes foram escolhidas por lançarem mão de recursos idiomáticos da língua mantendo um registro semi-formal: portanto, corroboram a argumentação de que as expressões idiomáticas, de fato, não pertencem a um escopo marginal da natureza lingüística; elas fazem parte do léxico corrente da língua (algumas mais flagrantemente do que outras).

No intuito de facilitar a coleta de dados, utilizamos a versão *on-line* destes periódicos. Cabe ressaltar, contudo, que nossa metodologia de coleta de dados foi qualitativa. Não tivemos como objetivo quantificar as ocorrências das expressões em questão, apenas apresentar um certo número delas para ilustrar e discutir as questões apresentadas neste trabalho.

Em suma, esta pesquisa objetivou contribuir tanto para uma reflexão sobre a noção de unidade lexical, como para um dos domínios da Lingüística Computacional, que é a Tradução Automática (TA, doravante) enquanto ferramenta.

Palavra: termo que foge a definições

Afinal, o que é uma palavra? O gerativista Steven Pinker, no seu livro *The Language Instinct* (1995: 147), antes de se aventurar a explicar este conceito, argumenta que o uso do termo não é cientificamente preciso: *palavra*, ele afirma, pode se referir a duas idéias distintas. Uma delas seria a noção de "átomo sintático", no sentido original de átomo como algo indivisível. Nesta acepção, o termo se refere às unidades da língua que são produtos de regras morfológicas e as quais são indivisíveis através de regras sintáticas; trata-se da palavra morfológica. A segunda acepção, explica Pinker, bastante diferente da primeira, está relacionada a "peda-

ços lingüísticos" arbitrariamente associados a um significado específico; um item da extensa lista que denominamos "dicionário mental". Os sintaticistas Anna Maria Di Sciullo e Edwin Williams (1987) cunharam este último conceito de palavra pelo termo *listema*: a unidade de uma lista memorizada (assim como *morfema* é a unidade morfológica e *fonema* a unidade sonora). É esta noção de palavra que estamos considerando no presente estudo: uma delimitação semântica e não formal da mesma.

Cruse (1986), Gross (1982) e Jackendoff (1997) ratificam a importância da delimitação semântica do léxico. Biderman (1999), compartilhando da mesma visão, justifica:

> …a fonologia e a morfossintaxe ajudam-nos a reconhecer segmentos fonicamente coesos e gramaticalmente pertinentes enquanto formas funcionais; contudo só a dimensão semântica nos fornece a chave decisiva para identificar a unidade léxica no discurso. Assim, no topo da hierarquia, a semântica vem congregar as demais informações de nível inferior para nos oferecer a chave do mistério da palavra. (p. 87)

Tomando como pressuposto esta idéia, propomos, neste estudo, que as expressões cristalizadas devem ser implementadas no léxico computacional de um programa que lide com PLN. Mais especificamente, estas expressões devem ser inseridas no dicionário de um tradutor automático. Esta tentativa, além de ser coerente com uma visão mais abrangente de léxico, resolveria o problema de tradução de uma expressão idiomática como sendo um fragmento sintático, embora ela apresente um conteúdo lexical.

O caso das Expressões Idiomáticas (EIs)

A definição tradicional de EI postula que seu significado não pode ser inferido através dos significados de suas partes. Estas

construções, em sua maioria, demonstram uma invariabilidade típica de unidades lexicalizadas. Portanto, elas necessariamente fariam parte do léxico do falante. Vale (1998) e Biderman (1999) fazem uso do exemplo tradicionalmente apresentado em língua portuguesa — *bater as botas* —para ilustrar o teor de uma EI:

> Verificamos que o significado da expressão não pode ser calculado pela soma dos significados dos seus componentes; ou seja, o seu significado nada tem a ver com o verbo *bater* nem com o substantivo *bota*. A análise tradicional pouco pode fazer nesta frase, pois as propriedades normalmente admitidas pelos verbos transitivos não funcionam nela (Vale, 1998: 132)

> O sentido da seqüência *bater as botas* não é previsível a partir de *bater* (dar pancada, chocar-se com) e de *botas* (tipo de calçado). De fato, temos aqui uma combinatória cristalizada, culturalmente herdada e registrada na memória coletiva com o significado de *morrer*. Por isso, podemos afirmar que ela faz parte do acervo do léxico e não é uma combinatória discursiva qualquer. (Biderman, 1999: 94)

Cruse (1986) argumenta que tal definição — a de que o significado da EI não pode ser inferido através dos significados de suas partes — pode ser lida como: "é uma expressão cujo significado não é resultado dos significados de suas partes quando estas não pertencem a uma EI" (p. 37). Cruse reconhece que tanto a primeira quanto, principalmente, esta última definição são circulares. Ou seja, para aplicar tais definições, devemos saber de antemão distinguir uma EI de uma expressão não-idiomática. Ele sugere, contudo, que é possível definir uma EI não-circularmente, utilizando a noção de constituinte semântico.

Segundo a sua proposta, a EI deve ter duas características: ser lexicalmente complexa — isto é, deve compreender mais de um constituinte lexical — e ser um constituinte semântico único — em

outras palavras, um constituinte que não pode ser segmentado em constituintes semânticos elementares (por exemplo, em *o gato está em cima do tapete*, o sintagma sublinhado é um constituinte semântico da frase, e as unidades *em, cima, do, tapete* são constituintes semânticos elementares da frase). Qualquer expressão que é divisível em constituintes semânticos é chamada de não-idiomática ou semanticamente transparente.[3]

Cruse se utiliza do tradicional exemplo da língua inglesa, "*kick the bucket*"[4], para explicar a particularidade do fenômeno idiomático. Curiosamente, esta expressão é a tradução consagrada de *bater as botas:*

> A razão pela qual '*kick the large bucket'* não seja interpretado idiomaticamente é porque '*bucket'* não carrega significado na EI, portanto '*large'* não exerce na EI sua função modificadora tradicional[5] (Cruse, 1986: 38)

Portanto, algumas das restrições de potencial sintático das EIs têm uma clara motivação semântica. Quando se diz *bater grandes botas* e *pegar no pé esquerdo de alguém*, estas construções não sofrem uma interpretação idiomática porque *botas* e *pé* não carregam um significado na EI; com isso, *grandes* e *esquerdo* não podem exercer sua função normal de modificadores.

No ponto de vista de Cruse, a EI é uma unidade lexical elementar: "embora consista em mais de uma palavra, apresenta uma coesão interna de palavras únicas" (p. 38). Seus componentes geralmente resistem à interrupção e reordenação das partes, como demonstrado nos exemplos 1 e 2, respectivamente:

1a) Depois de muito sofrimento, bateu as botas.
1b) ?Bateu, depois de muito sofrimento, as botas.

1c) Depois de muita discussão, deu o braço a torcer.
1d) ?Deu o braço, depois de muita discussão, a torcer.

2a) O que ele fez foi bater as botas. (deixa a EI fisicamente intacta)

2b) O que ele bateu foram as botas. (não tem leitura idiomática)

No ponto de vista de Cruse (1986), a expressão idiomática é uma unidade lexical elementar: "embora consista em mais de uma palavra, apresenta uma coesão interna de palavras únicas" (p. 38). O autor considera Expressões Idiomáticas, Metáforas Cristalizadas e Colocações como tipos de expressões cristalizadas distintas. Porém, Cruse reconhece, há casos limítrofes. Por esta razão tais distinções não são consideradas relevantes para o domínio do tratamento automático do léxico, como argumenta Santos (1990: 3), quando postula que "as fronteiras entre restrições colocacionais, leituras metafóricas e expressões idiomáticas são difusas e talvez impertinentes para o tratamento automático da língua". Por isso, iremos denominar todos estes fenômenos como expressões cristalizadas.

Delimitando as Expressões Cristalizadas

O artigo de Neves (1999) vai tratar da delimitação das unidades lexicais partindo da investigação do comportamento de construções com verbo-suporte, o qual é contrastado com o de certas construções de formação semelhante. Da introdução do artigo de Neves, podemos esboçar o seguinte vetor:

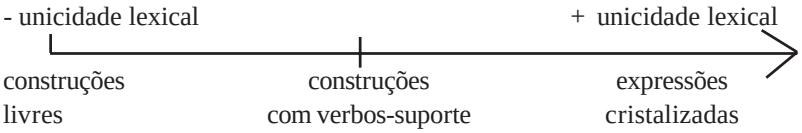

Gráfico 1

Na extrema esquerda, temos combinações com verbos plenos e sintagmas nominais complementos, que são completamente livres (ex: *consolidar a estrada; findar propostas*), onde os dois elementos exercem papéis independentes na estrutura argumental; na extrema direita, temos expressões que constituem um significado unitário, em que "nem mesmo parece ser possível postular um SN em posição de objeto" (Neves, 1999: 99), como *dar um pulo*, *tomar partido*; e entre estes dois graus extremos de construção, há aquelas construções intermediárias, constituídas dos chamados verbos-suporte, que, por sua vez, recebem certo grau de esvaziamento do sentido lexical, porém, semanticamente contribuem para o significado total da construção (*dar um riso; ter confiança*).

Na proposta do presente estudo, utilizamos construções *bater + SN* e as diferenciamos de acordo com o vetor estabelecido anteriormente. Desta forma, pudemos classificar as construções cristalizadas deste tipo e incluí-las no dicionário do tradutor automático *Delta*®.

No intuito de classificar inequivocamente a estrutura dos constituintes de construções como as supracitadas, Neves se apropria dos testes propostos por Radford (1988:90) e os adapta para a língua portuguesa. Segundo o gerativista inglês, os instrumentos mais seguros para determinar a estrutura dos constituintes destas expressões são: a distribuição, a posposição, a coordenação, a intercalação de advérbios, a elipse.[6]

Desta forma, os testes propostos por Radford (1988) são capazes de distinguir as construções livres, as construções com verbo-suporte e as construções fixas, cristalizadas. E isto veio ao encontro das aspirações de nossa análise.

Enquanto as expressões que se encontram na esquerda do vetor são consideradas livres e até imprevisíveis, a distinção entre os dois outros tipos de construção traz uma certa hesitação pelo fato de ambas as estruturas se situarem no domínio da convencionalidade; ou seja, "das estruturas recorrentes que o falante escolhe com reduzida liberdade quanto ao modo de composição" (Neves, 1999: 103).

Neves explica que as construções com verbo-suporte ora se situam mais próximas de construções livres, ora mais próximas de expressões cristalizadas (ou seja, ora mais próximas de um, ora mais próximas de outro extremo do vetor). Elas são compostas de: a) um verbo com determinada natureza semântica básica, que funciona como instrumento morfológico e sintático na construção do predicado; b) um sintagma nominal que entra em composição com o verbo para configurar o sentido do todo, bem como para determinar os papéis temáticos da predicação.

Usamos como exemplo expressões tais como *bater perna* e *bater boca*, além de expressões em que o SN posposto ao verbo expresse um sentimento de falta ou negativo, como *bater uma fome, bater um medo*.

Embora a construção "bater + SN de sentimento de falta" (como *bater uma saudade, bater uma fome, bater uma sede*) ou "bater + SN de sentimento negativo" (como *bater um desespero, bater uma angústia, bater um medo*) não tenham o comportamento exato do que chamamos tradicionalmente de construções com verbo-suporte (como *ter confiança, dar um riso, dar uma olhada*) — pois, na verdade a sua construção sintática seria *bater em alguém uma dúvida/ um medo* — tentamos mostrar que seu comportamento semântico é análogo ao destas construções.

Usamos as seguintes construções de estrutura "bater + SN" encontradas durante a coleta de dados nos periódicos *JB*, *O Globo* e *Veja*, para estabelecer seu verdadeiro estatuto de acordo com o vetor proposto no início deste capítulo : a) *bater perna; b) bater papo; c) bater o pé; d) bater os olhos; e) bater palmas; f) bater as botas; g) bater boca*[7]*, h) bater bola* (que estariam na extrema direita do vetor de unicidade lexical); *i) bater uma dúvida, j) bater o desespero* (que se enquadrariam no centro do vetor); *l) bater a CBF, m) bater a concorrência* (que julgamos pertencerem à extrema esquerda do vetor, ou seja, estariam distantes do estatuto de unicidade lexical). Os exemplos encontram-se na figura 2.

Além de aplicarmos os testes de Neves (1999) para diferenciar as construções cristalizadas das construções com característica de verbo-suporte do tipo *bater + SN*, incluímos, ainda, um último teste que demonstra que alguns dos exemplos que a autora considera como expressões cristalizadas podem admitir a inserção do marcador de freqüência *muito*: é o caso de *bater perna, bater papo, bater boca, bater palmas, bater bola*. Chegamos à conclusão de que as expressões *bater os olhos, bater as botas* e possivelmente *bater o pé* parecem não admitir a inserção do marcador de freqüência. Isto parece estar intimamente relacionado ao perfil semântico destas expressões, que apresentam um aspecto pontual. Mateus et alii (1983: 134) explicam que "o valor aspectual pontual caracteriza enunciados que descrevem eventos em que ocorre a mudança de estado ou transição sofrida por uma dada entidade". Portanto, o perfil pontual dessas expressões parece estar intimamente relacionado ao bloqueio da possibilidade de marcação de freqüência.

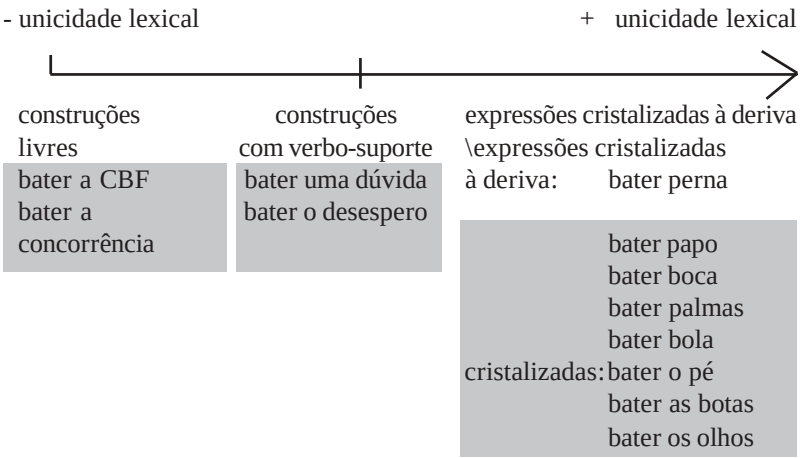

Gráfico 2

Como o propósito prático deste estudo foi viabilizar um tratamento automático das expressões cristalizadas *bater + SN*[8], ou seja, a sua inserção no banco de dados do *Delta Translator*®, para que sejam

tratadas como expressões lexicais e não como fragmentos sintáticos, podemos observar alguns problemas a serem equacionados:

1) Como resolver os casos das expressões cristalizadas à deriva, que tendem a admitir o marcador de freqüência e, em virtude disso, não seriam inequivocamente previsíveis como as expressões cristalizadas o são ?

Para um tratamento automático mais preciso e otimizado das expressões cristalizadas à deriva, incluímos, no tradutor automático, estas expressões com e sem a inserção do marcador de freqüência *muito*.

2) Como resolver o caso da expressão *bater papo* que, por sua vez, além de admitir o marcador de freqüência, admite também a inserção de um quantificador? Esta expressão traz uma dificuldade ainda maior para um tratamento automático, pois admite a inserção de diferentes tipos de marcadores de freqüência e quantificadores como *longo(s), muito(s), vários, três, diversos, etc*. e, portanto, diferentemente das construções cristalizadas à deriva, parece estar ainda mais distante do estatuto de unicidade lexical, pelo menos no que tange a esse quesito.

Para um tratamento automático da expressão *bater papo,* inserimos além do marcador de freqüência *muito*, o quantificador *um*, visto que parece ser o mais utilizado com esta expressão.

As únicas expressões que parecem não trazer problemas para um tratamento automático são as cristalizadas, em cuja combinação não existe nenhuma flexibilidade ou liberdade, sendo, desta forma, inequivocadamente previsíveis. São elas: *bater as botas, bater os olhos, bater o pé*. Estas últimas podem ser tratadas automaticamente como sendo uma unidade lexical, pois tendem a rejeitar a inclusão de qualquer tipo de constituinte.

Inserindo as Expressões Cristalizadas no *Delta Translator*®

Embora o *Delta*® seja um programa bastante sofisticado, o seu léxico computacional não é tão abrangente quanto aparenta. Este é,

atualmente, um dos maiores problemas dentro da Lingüística Computacional. Boguraev e Pustejovsky (1996: 3) ressaltam:

> Independentemente da sofisticação do sistema, seu desempenho deve ser medido em grande parte pelos recursos do léxico computacional associado a ele. Tais recursos viabilizam tarefas como análise lingüística, processamento de texto, sumário de documentos e Tradução Automática… há um enorme número de diferentes classes de palavras que ficam fora do alcance de um dicionário computacional

Dentre as unidades que ficam "fora do alcance de um dicionário computacional" estão as expressões cristalizadas. Abaixo apresentamos as frases com expressões cristalizadas do tipo *bater + SN* que o *Delta®* não conseguia equacionar ao verter para o inglês:

a) Ele *bateu perna* no centro até achá-los.
b) …e passa o dia *batendo papo* com a vizinhança.
c) O *chef bateu o pé*, disse que não cederia.
d) Quem *bate os olhos* nas fotos desta reportagem…
e) Eles *bateram palmas* depois do hino nacional.
f) Dessa forma, eles acabariam *batendo as botas*.
g) Ana *bateu boca* no shopping.
h) Ele *bateu bola* com Paulo César em 1980.

A seguir, o resultado da versão automática:

a) He *beat leg* in the downtown until find them.
b) …and he spends the day *beating crop* with the neighborhood.
c) Chef *beat the foot*, he said that wouldn't give way.
d) Who *it beats the eyes* in the photos of this report…
e) They *beat palmas* after the national anthem.
f) Thus, they would finish *beating the boots*.
g) Ana *beat mouth* in mall.

h) He *beat ball* with Paulo César in 1980.

Claramente, estas expressões não pertenciam ao banco de dados lexical do *Delta*®; portanto, eram solucionadas como um fragmento sintático. Sendo assim, percebemos que um assistente sintático robusto nem sempre resolve os problemas presentes na Tradução Automática. A simples inclusão destas construções em um dicionário de expressões solucionaria estes problemas, como foi implementado neste estudo, cujo resultado se segue[9]:

a) He went around in the downtown until find them.
b) …and he spends the day chatting with the neighborhood.
c) Chef stood fast, he said that wouldn't give way.
d) Who glances at the photos of this report…
e) They clapped hands after the national anthem.
f) Thus, they would finish kicking the bucket.
g) Ana squabbled in mall.
h) He played ball with Paulo César in 1980

Contudo, apesar de ser possível incluir expressões em seu léxico, concluímos que o tradutor automático em questão não possui uma heurística para a inferência verbal. Isto quer dizer que, embora ele reconheça todas as ocorrências do verbo *bater*, teríamos que inserir todas as possibilidades de ocorrência deste mesmo verbo quando o editamos dentro de uma expressão. Isto, no entanto, requer um trabalho manual incompatível com a sofisticação do sistema. Portanto, o programa não está preparado para lidar com construções cristalizadas encabeçadas por verbos, mesmo que estes já estejam em seu Dicionário de Palavras. Esta parece ser a grande falha do programa. Acreditamos, contudo, que uma pesquisa direcionada em lingüística computacional pode equacionar essa falha em prol de uma tradução mais completa e fluente.

Conclusão

Durante o desenvolvimento do presente estudo, algumas idéias conclusivas puderam ser estabelecidas como:

1) A visão de que o critério semântico é o decisório para a delimitação da unidade lexical permite uma visão mais inclusiva de itens lexicais e tenta dar conta daquilo que Gross (1982) e Jackendoff (1997) julgam como uma enorme falha dentro dos estudos lingüísticos: o não tratamento científico de expressões cristalizadas. Esta perspectiva lexical mais abrangente pautou o decorrer de nossa investigação.

2) Levamos para o desenvolvimento do presente trabalho a idéia de que o léxico não contém apenas estruturas simples, e especialmente a noção de que as expressões cristalizadas não fazem parte do "lixo lingüístico", e sim, pertencem ao conhecimento lingüístico do falante (de acordo com Jackendoff, 1997). Tomando como pressuposto esta idéia, propusemos que tais expressões devessem ser implementadas no léxico computacional de um programa que lide com processamento de linguagem natural. Em outras palavras, estas expressões devem ser inseridas no dicionário de um tradutor automático. Esta tentativa, além de ser coerente com uma visão mais abrangente de léxico, resolveria o problema de tradução de uma expressão idiomática como sendo um fragmento sintático.

3) Demonstramos como é possível aplicar testes que verifiquem o estatuto de cristalização destas expressões. Introduzimos um teste complementar que parece indicar que a semântica da expressão, mais especificamente o valor aspectual da mesma, interfere na sua soldadura. Em outras palavras, verificamos que as expressões com um perfil aspectual pontual tendem a bloquear a inserção de marcador de freqüência; o mesmo não parece ocorrer com as expressões com perfil durativo. Em suma, o valor semântico da expressão em um ambiente sintático-semântico específico parece determinar o seu estatuto de unicidade lexical.

4) Detectamos a dificuldade de inserção de expressões cristalizadas no *Tradutor Automático Delta*®, particularmente, pela sua inefi-

ciência de inferência verbal em expressões. Argumentamos que o trabalho manual necessário para a edição de todos os tempos verbais em que a expressão pode ocorrer é incompatível tanto com a sofisticação do programa quanto com a otimização de tempo que um tradutor automático deve oferecer ao tradutor humano. Aliada a esta dificuldade estaria a necessidade de aplicação da bateria de testes para a certificação do estatuto de cristalização das expressões. Portanto, concluímos que a solução mais coerente seria otimizar esta metodologia para ser aplicada por lingüistas computacionais programadores de tradutores automáticos antes mesmo de estes programas chegarem ao mercado. Defendemos que uma metodologia baseada nos testes de Neves poderia servir como suporte para o desenvolvimento de um dicionário automático mais abrangente e, conseqüentemente, para um programa de tradução automática que não converta uma expressão lexical como um fragmento sintático.

5) Finalmente, destacamos a importância do tradutor automático como ferramenta e ratificamos que de forma alguma consideramos que nossa metodologia de delimitação de expressões cristalizadas resolveria a maior parte dos problemas de TA e gradativamente substituiria o trabalho humano. O que consideramos de suma importância é a resolução de algumas questões lingüísticas que otimizaria a capacidade de tradução de programas especializados. Na verdade, uma revisão do léxico computacional equacionaria muitos problemas típicos de uma ferramenta automática.

Notas

1. Processamento de Linguagem Natural

2. Expressões como *bater as botas, bater perna, bater boca, bater bola, bater o*

martelo, *bater ponto*, dentre outras, foram incluídas no dicionário de um tradutor automático bilíngüe (port/ing) para que o sistema não equacionasse as expressões de forma sintática e literal. Portanto, o programa deixou de verter a expressão *bater as botas* como "*beat the boots"* e passou a fazê-lo por "*kick the bucket*". Contudo, como estávamos lidando com a interface entre o usuário e o programa, não aplicamos nenhum tipo de formalização destas expressões.

3. A maioria das EIs coincidem com expressões semanticamente transparentes (como *bater as botas*, "*kick the bucket"*); aquelas que não apresentam esta transparência (como *tirar de letra*, *"by and large"*) são chamadas de *EI assintáticas*.

4.Este exemplo é também utilizado por Gibbs (1995), numa abordagem de semântica cognitiva, e Jackendoff (1997), numa abordagem gerativa.

5. As traduções de textos estrangeiros no artigo, quando não indicados, são de autoria de Milena Garrão.

6. Para uma observação detalhada dos testes, consultar Neves, 1999 ou Garrão 2001: cap. 4.

7. Os exemplos e), f) , g) foram formulados pela autora.

8. Como o nosso objeto de estudo são as expressões cristalizadas, aquelas com verbo-suporte não serão inseridas no léxico computacional do tradutor automático em questão, embora seja importante ressaltar que constituem um fenômeno importante de uma língua e, portanto, seu tratamento automático é desejável.

9. A inconsistência sintática das frases traduzidas pelo programa (regência verbal, por exemplo) não foi objeto do nosso estudo. Hoje em dia, o grau de aceitação de tradutores automáticos é medido pela quantidade de pós-revisão requerida. Um programa cujo índice de revisão posterior é menor do que 20% (uma correção a cada cinco palavras) é considerado aceitável. (ver Alfaro & Dias, 1998 e Garrão, 2001).

Bibliografia

ALFARO, C & M.C.P. DIAS (1998). "Tradução Automática: uma ferramenta de auxílio ao tradutor". In *Cadernos de Tradução* n° 3. Centro de Comunicação e Expressão:GT de Tradução. Universidade Federal de Santa Catarina.

BAR-HILLEL (1964). *Language and Information. Selected essays on their theory and application*. Massachusetts: Addison-Wesley Publishing Company.

BASÍLIO, M. (1999a). "Questões Clássicas e Recentes na Delimitação de Unidades Lexicais". In Basílio, M. (org.) *Palavra* n° 5. Rio de Janeiro, Departamento de Letras da PUC: 9 -18.

BIDERMAN, M.T. (1999). "Conceito Lingüístico de Palavra". In Basílio, M. (org.) *Palavra* n° 5. Rio de Janeiro, Departamento de Letras da PUC: 81-97.

BOGURAEV, B. & J. PUSTEJOVSKY (1995). "Issues in Text-based Lexical Acquisition". In Boguraev, B. e Pustejovsky, J (orgs.) *Corpus Processing for Lexical Acquistion*. Cambridge, Massachusetts: MIT Press.

CRUSE, D. A. (1986). *Lexical Semantics*. Cambridge, Inglaterra: Cambridge University Press.

DI SCIULLO, A-M, & E. WILLIAMS (1987). *On the Definition of Word*. Cambridge, Massachusetts: MIT Press.

GARRÃO (2001). *Tradução Automática: ainda um enigma multidisciplinar*. In Pereira. J. (org.) Anais do V congresso Nacional de Lingüística e Filologia. Instituto de Letras da UERJ, Rio de Janeiro.

GIBBS (1994). *The Poetics of Mind*. New York: Cambridge University Press.

GROSS, M. (1982). "*Une Classification des phrases 'figées' en français*". Revue Québecoise de Linguistique, 11:151-185.

JACKENDOFF (1997). *The Architecture of the Language Faculty*. Cambridge, Massachusetts: MIT Press.

_____(1998). "What's in the Lexicon ?" Resumo de conferência apresentada no *Utrecht Congress on Storage and Computation in Linguistics*. Universiteit Utrecht: Holanda.

MATEUS, M.H.M. (1995). "Tradução Automática: um pouco de história". In Mateus, M.H & Branco, A. H. (orgs.) *Engenharia da Linguagem*. Faculdade de Letras da Universidade de Lisboa, Lisboa: Ed. Colibri. 115-120

MATEUS, M.H.M. et alii (1983). *Gramática da Língua Portuguesa*. Coimbra: Livraria Almedina.

NEVES, M.H.M. (1999). "A delimitação das unidades lexicais: o caso das construções com verbo-suporte". In Basílio, M. (org.) *Palavra* n° 5. Rio de Janeiro: Departamento de Letras da PUC. 98-114.

PINKER, S. (1995). *The Language Instinct*. New York: Harper Perennial.

RADFORD, A. (1988). *Transformational grammar: a first course*. Cambridge, Inglaterra: Cambridge University Press.

SANTOS, D. (1990). "Lexical gaps and idioms in Machine Translation», Hans Karlgren (org.), *Proceedings of COLING'90* Vol 2. Helsinki. 330-335.

VALE, O. A. (1998). "Sintaxe, Léxico e Expressões Idiomáticas". In Brito A. N. & Vale, O. A. (orgs.) *Filosofia, Lingüística e Informática: aspectos da linguagem*, Goiânia: UFG.127-137.