Divergência Interna e Externa na Classificação de Erros em Inglês como Língua Estrangeira.

Fernando José Rodrigues da Rocha

O presente trabalho constitui-se numa apresentação parcial de um tema desenvolvido no V Congresso da Associação Internacional de Linguística Aplicada. Seu objetivo é o de, através da detecção das áreas de divergência na classificação de erros em Inglês como língua estrangeira, contribuir para uma metodologia da Análise de Erros onde a interferência da subjetividade dos pesquisadores não venha por em jogo a fidedignidade dos resultados obtidos.

O experimento realizado baseou-se em um corpus composto de 135 composições escritas por alunos brasileiros de língua inglesa, como parte do "Lower Cambridge Exam". O total de 2007 frases foi introduzido em um computador e suas partes automaticamente referenciadas. Um programa especial imprimiu este material num formato adequado a uma enquete linguística. Esta consistiu em solicitar a 14 professores de linguística ou de inglês das universidades inglesas de Reading, Birmingham e Edimburgo para testar a aceitabilidade das frases do fragmento do corpus que lhes foi confiado. Aquelas julgadas compreensíveis, mas não aceitáveis a um falante culto deveriam ter um tratamento específico. Elas deveriam ser corrigidas e os erros encontrados classificados em: (1) sintático, (2) léxico-semântico ou (3) morfológico, incluindo-se a grafia. Em casos de co-ocorrência de erros numa mesma palavra, os dois códigos respectivos deveriam ser colocados.

Obteve-se através da enquete um corpus de 1.800 frases contendo 3.238 palavras erradas e, por conseguinte, codificadas. Cada erro foi então analisado linguisticamente e classificado dentro de 42 descritivas, com números variáveis de sub-categorias, perfazendo um total de 354 códigos. Novamente, em casos de justaposição de erros numa mesma palavra, códigos distintos foram atribuidos.

Uma vez o fichário atualizado, isto é, quando todos os dados foram registrados em fita magnética, o computador forneceu mediante um programa de seleção, uma listagem na qual os códigos de descrição linguística dos erros foram agrupados dentro dos fragmentos analisados por cada um dos corretores. Deste modo, pode-se constatar não somente as áreas de divergência entre os diversos corretores (divergência externa) mas também as áreas de flutuação de cada um deles individualmente (divergência interna).

Na tabela a seguir o grau de divergência interna foi obtido através do cálculo da percentagem de corretores, dentre os 14,que atribuiram mais de um código de tipo de erro para um mesmo fenômeno linguístico, ou seja, para um mesmo código de descrição linguística do erro.

Para o estabelecimento da divergência externa, foram calculados os coeficientes de variação através das fórmulas seguintes:

$$\bar{x} = \frac{\Sigma Pi}{n} \qquad S = \frac{\sqrt{\Sigma (Pi - \bar{x})^2}}{n-1}$$

$$cv = \frac{S}{\bar{x}}$$

sendo o grau de divergência externa o contrário do coeficiente de variação (1- cv).

Dentre os resultados obtidos, destaca-se a relação dos dez casos onde o maior índice de divergência externa foram encontrados, acompanhada dos índices de divergência interna e da distribuição dos códigos de erros atribuidos pelos corretores (o código zero indica erro detectado, mas não codificado).

TABELA 1

Graus de divergência na classificação de erros

Descrição do erro	Distribuição dos códigos de erro				grau de divergência	
	/ 0	/ 1	/ 2	/ 3 /	externa	interna
1.Substituição Subst/adj		33.3%	33.3%	33.3%	1,00	7.1%
2.Separação de palavras		41.6%	16.8%	41.6%	0,71	7.1%
3.Omissão de determinante	0.7%	44.8%	41.8%	13.5%	0,67	7.1%
4.Forma verbal inexisterte		46.1%	15.4%	38.5%	0,57	14.3%
5. N.dodetermin.:sing/plural	10.0%	40.0%	50.0%		0,52	0,0%
6. N.do substant.:sing/plural		37.0%	15.8%	47.2%	0,52	14.3%
7.Substituição det/det		44.4%	33.3%	22.2%	0,40	14.3%
8.Substituição adv/conj		31.3%	25.0%	43.7%	0,37	7.1%
9.Substituição prep/prep	14.3%	47.0%	28.4%	14.3%	0,28	50.0%
10.Substituição passado simples/presente	4,4%	54.0%	4.4%	35.0%	0,00	14.3%

Com base nos dados acima apresentados foi concluído que, mesmo se preparada por um grupo de pesquisadores, nenhuma taxonomia de erros em termos de níveis da língua pode ser considerada isenta de subjetividade. Isto porquanto as noções de sintaxe, léxico, semântica e morfologia são passíveis de interpretações ambíguas. Constatou-se que a imprecisão das fronteiras destas áreas é tal, que dois ou mais códigos foram atribuidos à cerca da metade (49.8%) das 354 categorias descritivas examinadas pelos corretores. Por outro lado, é também importante verificar-se que as restantes categorias descritivas (51-1%) às quais somente um código de erro foi atribuido têm características especiais. A maioria delas (57.4%) é composta por estruturas que ocorrem somente uma vez no corpus, deixando, desta forma, nenhuma margem para eventuais flutuações. Em outras palavras, o índice de divergência aumenta à medida em que a frequência de cada código descritivo cresce,isto é, quanto mais uma estrutura errada aparece no corpus,mais possibilidade ela tem de ser classificada de modo divergente. Este truismo aparente tem suas consequências. A partir dele pode se inferir que a não-divergência na classificação de erros, segundo os níveis linguísticos clássicos, é primordialmente um produto da baixa frequência dos itens e não oriundo de um consenso geral entre os corretores.

Foi constatado que em certas áreas não há uma tendência nítida quanto à classificação de erros em Inglês como língua estrangeira. A tabela 1 mostra que, no caso da substituição de um substantivo por um adjetivo, um número igual de ocorrências foi classificado como pertencentes aos grupos 1, 2, e 3 o que significa que as opiniões dos corretores dividiu-se em três partes equitativas. No quarto caso, em que formas verbais não existentes na língua deveriam ser classificadas, uma divisão em duas facções foi observada. Uma optando pelo nível sintático, enquanto a outra preferiu considerar o erro como pertencente ao nível morfológico.

Nos demais casos delineia-se uma tendência em direção a um ou outro código. No entanto, um outro tipo de ocorrência merece ser chamado à atenção: quando um determinante foi substituido por outro determinante com valor lexical e semântico distinto, 14.3% dos corretores detectaram o erro, mas aparentemente não lograram decidir-se quanto ao código a lhe atribuir e, assim sendo, preferiram não codificar, deixando o espaço em branco.

A análise dos graus de divergência interna mostra que foi a classificação da substituição de uma preposição por outra, independentemente da regência verbal, que causou maiores problemas à decisão dos corretores. Metade destes codificou, em várias reprises, de forma distinta o mesmo tipo de erro.

Foi também observado que, exceto o caso da substituição de preposições, todos os demais têm algo a ver com a forma das palavras, também que é o uso indevido de determinantes (número, omissão e substituição) que causam maior dificuldade à classificação. Os substantivos (substituição por adjetivos e mudança de número) e os verbos (formas inexistentes na língua e troca de tempos) constitutem o segundo maior foco de problemas quanto à diversidade de classificação, não sendo levadas em conta a freqüência de suas ocorrências no corpus.

Como consideração final sugere-se que mais atenção seja dada à questão da validade e utilidade da classificação de erros em língua estrangeira dentro de níveis descritivos que podem ser postos em questão e, sobretudo, tornar infrutífera a comparação dos resultados de pesquisas distintas.