

Capítulo 4

O corpus CONDIV e o estudo da convergência e divergência entre variedades do português

Augusto Soares da Silva

Prendemos apresentar o corpus CONDIV, em construção, e perspectivar o seu desenvolvimento como instrumento fundamental para o estudo da convergência e divergência entre as variedades europeia e brasileira do português. O CONDIV foi compilado no âmbito do projecto de investigação “Convergência e divergência no léxico do português” (2004-2006), centrado na questão diacrónica da convergência/divergência lexical entre o português europeu e o português brasileiro nos últimos 50 anos e, secundariamente, na questão sincrónica da estratificação lexical das duas variedades. Parte do corpus está disponibilizado na Linguateca, fazendo parte do projecto AC/DC (Santos e Sarmiento, 2002), e espera-se que brevemente o restante também aí se possa encontrar.

O CONDIV compreende actualmente textos de três domínios – futebol, vestuário e saúde – e está estruturado na base de três variáveis: (i) geográfica (Portugal vs. Brasil), (ii) diacrónica (1950-1970-1990/2000) e (iii) estilística (jornais e revistas > etiquetas, Net fóruns > Net *chats*). Os materiais foram extraídos de três fontes: (i) jornais de desporto e revistas de moda e de saúde dos primeiros anos das décadas de 50, 70 e 90-2000; (ii) linguagem da Internet de conversação electrónica de IRC ou *chats*; e (iii) etiquetas de roupas de lojas de vestuário. Todos os materiais de (i) e (iii) foram extraídos manualmente. Os materiais do português brasileiro provêm de São Paulo e Rio de Janeiro. A extensão actual do corpus é de 5 milhões de palavras do registo formal (jornais e revistas) e 15 milhões do registo informal (*chats* e etiquetas).

O objecto de análise é a variação onomasiológica que envolve sinónimos *denotacionais* – porque são estes os que melhor revelam a própria existência e a competição de variedades de uma língua – e a sua base empírica consiste em largos milhares de observações do uso de sinónimos alternativos que designam um mesmo conceito/referente. Na fase actual, foram estudados 43 conceitos nominais dos campos lexicais do futebol (21 conceitos e 183 termos) e vestuário (22 conceitos e 264 termos).

Os resultados da investigação sociolexicológica realizada (Silva, 2005) indicam que as duas variedades divergem claramente no vocabulário do vestuário (e a distância actual entre padrão e subpadrão é aí maior na variedade brasileira), mas convergem ligeiramente no vocabulário do futebol. Os mesmos resultados apontam para a existência de mais mudanças na variedade brasileira. É também a variedade brasileira a que manifesta uma maior permeabilidade aos estrangeirismos.

Prendemos prosseguir a investigação estendendo-a em três direcções: (i) ainda no domínio lexical, incluir outros campos lexicais e passar das palavras *de conteúdo* às palavras *funcionais*, particularmente as preposições; (ii) em direcção ao domínio gramatical, analisar variáveis não-lexicais, sobretudo sintácticas e morfológicas; e (iii) ampliar e refinar a situação estratificacional de cada variedade, acrescentando mais variáveis estilísticas, em ordem ao estudo dos indicadores de substandardização. Para isso, é essencial aumentar quantitativa e qualitativamente o corpus existente. Dada a escassez de textos em suporte informático que cumpram o critério diacrónico do projecto e a escassez ou mesmo falta de

subsídios para a investigação, lançamos um desafio à Linguateca no sentido da convergência de projectos e recursos.

O quadro teórico da presente investigação é o da linguística cognitiva. Três razões justificam a opção pela linguística cognitiva: (i) a sua orientação *recontextualizante* (reintegrando as diferentes formas de *contexto*, excluídas pelos modelos autonomistas, particularmente o generativista), para o *significado* (incluindo o significado social) e *baseada-no-uso* (origem da própria variação linguística); (ii) a importância dada à flexibilidade e à variação semasiológica e onomasiológica (Silva, 2006a); e (iii) a assunção de que não só a capacidade para a linguagem se fundamenta em capacidades cognitivas gerais, como também todas estas capacidades são cultural e socialmente situadas. Mais especificamente, a nossa investigação insere-se no âmbito da emergente sociolinguística cognitiva (Kristiansen e Dirven, 2008; Silva, 2006b, 2008), inevitavelmente implicada por aqueles princípios. A sociolinguística cognitiva está bem colocada para ajudar a resolver duas tensões: (i) a tensão entre o *cognitivo* e o *social*, integrando no programa cognitivista a variação intralinguística (Geeraerts, 2005; Bernárdez, 2005); e (ii) a tensão entre o *cognitivo* e o *empírico*, optando pela metodologia de corpus (ou dados experimentais) e por técnicas quantitativas capazes de analisar a natureza *multivariacional* do uso linguístico (Geeraerts, 2006; Gonzalez-Marquez et al., 2007). Neste âmbito, a presente investigação apoia-se na concepção e nos métodos quantitativos da investigação desenvolvida por D. Geeraerts e sua equipa para o neerlandês da Holanda e da Bélgica (Geeraerts et al., 1999).

Para medir a convergência e a divergência entre variedades linguísticas são utilizados dois métodos quantitativos: a medida de uniformidade (U) entre as variedades e a medida do impacto de determinado traço (A) nessa uniformidade. A medida U é a medida fundamental. Envolve duas noções específicas: *perfil onomasiológico* ou conjunto de sinónimos denotacionais usados para designar determinado conceito ou função, diferenciados pela sua frequência relativa, e *uniformidade* ou medida da correspondência entre dois conjuntos de dados, definidos em termos de perfis onomasiológicos. Por exemplo, a uniformidade de um conceito entre duas amostras, em que uma contém 6 ocorrências do termo A e 4 do termo B e a outra 3 ocorrências de A e 7 de B, resulta do número de pares comuns de nomeação desse conceito (7 pares), sendo portanto $U = 70\%$. Este resultado obtém-se somando as frequências relativas mais pequenas de cada termo alternativo: 30% de A e 40% de B. Tecnicamente, a uniformidade de um conceito é calculada pela seguinte fórmula:

$$U_Z(Y_1, Y_2) = \sum_{i=1}^n \min(F_{Z,Y_1}(x_i), F_{Z,Y_2}(x_i))$$

Isto é, a uniformidade U para um conceito Z entre duas amostras Y_1 e Y_2 equivale à soma \sum dos mínimos das frequências relativas F do termo x nos perfis onomasiológicos de Z em Y_1 e Y_2 . O símbolo x_i representa os diferentes termos x_1 a x_n usados nas amostras

Y para designar o conceito Z . Quando estão em causa vários conceitos, a uniformidade média é calculada em termos de média ponderada:

$$U'(Y_1, Y_2) = \sum_{i=1}^n U_{z_i}(Y_1, Y_2)G_{z_i}$$

A uniformidade U' para um conjunto de conceitos Z entre duas amostras Y_1 e Y_2 equivale à soma dos valores- U dos Z s ponderados pela frequência relativa G de Z dentro do conjunto total de Z s. Convergência e divergência entre duas variedades exprimem-se em aumento e diminuição de U/U' , respectivamente. Os cálculos ponderados (U' , A') são mais representativos, porque têm em conta a frequência relativa dos diferentes conceitos e termos em análise. A título de exemplo, a tabela 4.1 apresenta os resultados de U e U' do perfil de “avançado” no sub-corpus de jornais de desporto e relativamente a um total de 90.202 observações do uso dos termos de futebol seleccionados.

FUTEBOL	P50		B50		P70		B70		P00		B00	
atacante	101	8,8	119	36,6	50	13,6	208	73,8	42	9,7	658	96,2
avançado	820	71,6	3	0,9	175	47,4	0	0,0	240	55,4	0	0,0
avante	0	0,0	159	48,9	0	0,0	31	11,0	0	0,0	23	3,4
dianteiro	220	19,2	22	6,8	74	20,1	2	0,7	38	8,8	0	0,0
forward	1	0,1	17	5,2	0	0,0	0	0,0	0	0,0	0	0,0
ponta-de-lança	3	0,3	5	1,5	70	19,0	41	14,5	113	26,1	3	0,4
AVANÇADO	U = 16,9		U' = 0,6		U = 28,8		U' = 0,8		U = 10,1		U' = 0,4	

Tabela 4.1: Frequência (absoluta e relativa) e uniformidade U e U' do perfil onomasiológico de “avançado”.

Como extensões da investigação realizada, são analisados dez perfis onomasiológicos preposicionais e três perfis onomasiológicos construcionais. Os perfis preposicionais, restringidos ao mesmo contexto sintagmático, de forma a satisfazer a condição de sinonímia denotacional, incluem casos como *falar de/sobre/acerca de/em*, *precisar/necessitar de/()*, *ansioso de/para/por*. Os perfis construcionais incluem construções com verbos causativos e perceptivos seguidos de complemento finito ou infinitivo e construções com os adjectivos atributivos *verdadeiro*, *falso*, *bonito*, *lindo*, *recente* em posição posposta ou anteposta. A anotação do CONDIV, feita pelo analisador sintáctico automático PALAVRAS (Bick, 2000), permite a pesquisa imediata das ocorrências das referidas regências e construções, incluindo as mais complexas, como as diversas construções causativas e perceptivas seguidas de complemento infinitivo. Uma hipótese sociolinguisticamente relevante é a de que as palavras funcionais e as construções sintácticas se comportam em termos de variação linguística diferentemente dos itens lexicais de conteúdo. Os resultados obtidos, embora em número ainda reduzido, permitem confirmar a hipótese de uma divergência mais acentuada entre as duas variedades nacionais em relação a variáveis funcionais e sintácticas.