

Descrição superficial do corpus

Diana Santos

Maiο 1992

(Esta descrição não pretende substituir a descrição das fontes, publicada em 1990 em *Documentação do Grupo Científico IBM-INESC* e reeditada neste volume.)

O que o corpus contém

A escolha das frases que constituem este corpus não obedeceu a quaisquer critérios temáticos nem sintácticos, apenas se evitaram textos literários (são disso excepção as primeiras 100 frases, de que constituem a maioria, quer como originais quer como traduções) e frases com erros¹. Como as frases foram digitadas manualmente, todos os erros ortográficos detectados foram corrigidos, mantendo-se contudo inalterados os outros problemas, por ser muito menos provável que fossem da responsabilidade dos autores do corpus (veja-se, por exemplo, as frases 123, 182, 259, 304, 353, 362, 380, 393, 509, 519, 521, 535, 629). É sintomático que a maioria destes erros são devidos a falta ou uso incorrecto de pontuação.

Alguns casos que não seriam cobertos pelo conceito de frase numa gramática tradicional constituem entradas do nosso corpus, por se encontrarem precisamente misturadas, em textos reais, com frases propriamente ditas. Assim, 213 consiste numa referência bibliográfica, 240 exprime uma definição em formato de glossário, 271, 594 e 625 foram provavelmente encontradas como títulos ou legendas de figuras, e as frases 598 a 605 são itens duma lista num texto com características técnicas. 287 e 412 parecem ser reproduções de uma caixa de jornal. Em todos estes casos, e em mais alguns (por exemplo 194, 284, 285, 314, 337, 340, 411, 456, 610, 611, 616, 636), as “frases” não seguem o modelo sintáctico de oração principal com verbo finito, não possuindo sequer um verbo na maioria dos casos.

A linguagem oral também não foi excluída do corpus, e encontra-se em frases de entrevistas publicadas, ou de textos com evidente intenção de a reproduzir: veja-se por exemplo as frases 13, 14, 216 a 222, 236 a 238, 400, 417, 418, 423, 457, 628. Assim, até o calão tem o seu lugar.

Além disso, observámos algumas palavras que não se encontram em dicionário (para esta afirmação, usámos o *Dicionário Prático Ilustrado*, Porto: Lello & Irmão, 1990) e que passamos a apontar: “enculturação”, 8, “astigmática”, 12, “maçados”, 15, “patine”, 24, “rodeo”, 27, “paragramáticas”, 52, “desvelamento” e “fundacional”, 53, “baudelairiana”, 54, “ecologicamente”, 64, “balética”, 103, “tarifários” e “telecópia”, 112, “listados”, 127, “multiequipamento”, 129, “automatização” e “monitorização”, 130, “multicritério”, 138b, “barthesiano” e “desajustamento”, 144, “endurance” e “jogging”, 157, “absidóla”, 158, “ameiada”, 160, “rudimentaridade”, 165, “tirados”, 166, “dançares”, 167, “stress”, 177, “hipotermias”, 182, “sedia”, 183, “multifacetado”, 184, “pontilista”, 186, “expressionistas”, 188, “videotex”, 208, “marketing”, 211, “fundamentalismo”, 223, “militância”, 228, “suite”, 230, “autoconsumo”, 278, “mútuas”, 287, “fármaco”, 290, “andebol”, 292, “perspectividades” e “projectividades”, 299, “computacional”, 304, “tecnólogo”, 318, “morfossintáctica”, 322, “interdisciplinaridade”, 324, “subdomínios” e “corpus”, 326, “aldeamento”,

¹Um outro corpus, de menor dimensão, foi coligido exactamente para frases com erros.

334, “sarcasta”, 351, “repegar”, frase 358, “organizativas”, 387, “zumbidora”, 404, “complementarmente” e “ambientais”, 408, “software” e “hardware”, 410, “proposicionais”, 435, “optimamente”, 475, “recálculo”, 588, “supermagnético”, 600, “disposicionais”, 622, “conjuntural”, 634, “fractais”, 638.

Por outro lado, outras palavras encontram-se decididamente em sentidos distintos daquelas em que figuram no dicionário: ainda exemplificando com a mesma fonte, observámos “baixas”, 62, “concreto”, 97, “indicativos”, 125, “pontualmente”, 244, “pontos”, 390, “copiar”, 570 e “resolução”, 645.

De notar que nas listas acima não se encontram palavras compostas (unidas por hífen), nem todas aquelas que de uma forma ou outra se encontrem marcadas no texto como especiais, através do uso de maiúsculas ou de aspas. Os exemplos referem-se pois a palavras simples cuja utilização pelos falantes se fez sem consciência de que essas palavras (ou usos das mesmas) ainda não tinham sido “oficializadas” pelo lexicógrafo.

Tratamento prévio do corpus

Além da correcção de erros ortográficos, possivelmente não exaustiva, mencionada acima, convém fazer alguns comentários em relação à forma como o corpus aparece impresso nesta colectânea.

Algumas siglas correspondentes a nomes de empresas foram alteradas de forma a não serem reconhecidas; por outro lado, e devido a algumas vicissitudes no processamento automático do corpus, os algarismos que aparecem no texto podem não corresponder aos originais.

Mantivemos contudo o uso de maiúsculas das fontes, assim como tentámos conservar os sinais de pontuação utilizados.

Finalmente, os números de frase e a translineação das palavras (geralmente incorrecta) não faz parte do corpus em si (em que cada frase se encontra numa única linha) e apenas aparecem na versão impressa que aqui publicamos. Também o formato interno dos caracteres acentuados, e de algumas marcas de pontuação, é diferente.

Durante a escrita destas linhas, detectámos um erro na separação do corpus por linhas, nomeadamente na frase 138, que de facto corresponde a duas frases distintas e nem sequer provenientes da mesma fonte. Quando tal for necessário, as duas frases serão identificadas como 138a e 138b.