CORPUS DE FALA TRANSCRITO − TRANSCRIÇÃO ORTOGRÁFICA

Campo transcrição

Na fase de transcrição dos dados das fitas sonoras gravadas para computadores de grande porte, a transcrição ortográfica não seguiu, fielmente, o sistema de sinais escritos do alfabeto latino utilizado pela língua portuguesa. Isso se explica pela intenção de corrigir as inadequações decorrentes do fato de os computadores *mainframes* disponíveis na época não disporem, na sua cadeia de caracteres, de sinais diacríticos utilizados, na ortografia, para a acentuação gráfica, e pela tentativa de corrigir, ao menos em parte, as inadequações entre o sistema ortográfico e o sistema linguístico falado – algumas delas resultantes da supressão, quase total, dos chamados acentos diferenciais que distinguiam palavras homógrafas.

Procuramos, dessa forma, dar à transcrição ortográfica o máximo de funcionalidade, para que a cada diferença de segmento na transcrição ortográfica correspondesse uma diferença de sentido. Assim sendo, a transcrição ortográfica empregada, desprovida, na medida do possível, de ambiguidades, facilita, em muito, a aplicação de métodos por processamento automático. Por exemplo, os dados, na maneira como foram codificados, permitem a identificação e a definição paradigmático-sintagmática dos fonemas e alofones do *corpus* – permutando segmentos da transcrição fonética que apresentam uma única diferença num mesmo ponto da cadeia de caracteres fonéticos, verifica-se se se trata de fonemas distintos, quando essa única diferença se apresenta como suficiente para provocar a mudança da palavra na transcrição ortográfica, ou se se trata de alofones de um mesmo fonema, quando não ocorre alteração na transcrição ortográfica.

Nas novas Bases, em PC, dado que os sistemas atuais de processamento de dados contam, na sua cadeia de caracteres, com os sinais diacríticos utilizados, na ortografia, para a acentuação gráfica, seguimos o sistema de sinais escritos do alfabeto latino utilizado pela língua portuguesa, mantendo convenções para a distinção de palavras homógrafas e as demais anotações, pelas vantagens já explicitadas. Trata-se de tarefa que exigiu tempo e atenção.

Seguem os procedimentos utilizados nas transcrições ortográficas:

- emprego de recursos para, na medida do possível, desfazer os homógrafos nocionais, a fim de evitar-se, no processamento, uma mesma entrada para itens lexicais distintos, tendo em vista o tratamento dos *corpora* por programas de análise linguística (tratamento automático de textos). Os homógrafos foram, pois, identificados como palavras distintas. Para desfazer os homógrafos resultantes da supressão dos acentos de valor diferencial, usamos o acento circunflexo nas vogais de timbre fechado. Por exemplo: *acôrdo, almôço, apêlo, cêrca, colher, comêço, confôrto, desespêro, destempêro, emprêgo, enderêço, entêrro, êrro, esfôrço, estrêla, exagêro, gôsto, govêrno, interêsse, ôlho, pêso, refôrço, refrêsco, retôrno, sossêgo, trôco* (substantivos) x *acordo, almoço, apelo, cerca, colhêr, começo, conforto, desespero, destempero, emprego, endereço, enterro, erro, esforço, estrela, exagero, gosto, governo, interesse, olho, peso, reforço, refresco, retorno, sossego, troco* (formas verbais);

- exclusão do hífen ou traço-de-união que liga pronomes enclíticos a verbos, para que sejam tratados distintamente, constituindo itens lexicais diferenciados, dada a possibilidade de as partículas pronominais átonas (ou clíticos) assumirem a posição proclítica, na qual não são ligadas ao verbo, graficamente, por meio do hífen. Pelo critério morfossintático-semântico, identificam-se, nesses casos, dois segmentos, pertencentes a classes gramaticais distintas, com funções e significados diferentes. Exemplos: *punha se*, *utilizar nos*. Em função disso, fizeram-se alterações quando se tratava de infinitivos – utilizá-lo para utilizar o;

- representação do *corte de palavra* e da *gaguez*, respectivamente, por / e (...), juntamente com os signos de pontuação, conforme codificação apresentada a seguir, na referência ao campo *pontuação*;

- transcrição ortográfica das interjeições, dos cliques e de outras realizações fônicas, que veiculam, fora da simples enunciação, informações complementares relevantes para a boa compreensão de um diálogo – como *alegria, hesitação, surpresa, desejo, chamamento, incredibilidade, receio, interrogação* etc –, variações essas de sentido expressas pelos signos de pontuação que acompanham a representação das emissões: *ah, ahn, eh, éh, ih, oh, oh, ó, uh, uhn, ay, chi, chiu, fí, hah, heh, huh, há, ham, hum, hara, hein, thu, oy, po, pu, puta, putis, puxa, uai, ué, opa, pomba, ora, nem, nossa, olha, olhe, sabe, tsi (clique), tsu (clique).*

A transcrição ortográfica – como a fonética – de todas as produções sonoras dos informantes, inclusive *cortes de palavra, gaguez, interjeições, cliques e outras emissões denotadoras de hesitação na enunciação,* procuram reproduzir, graficamente, certos dados informativos do código falado, tanto os *linguísticos* – como os movimentos de entonação – quanto os *extralinguísticos* – explícitos no contexto situacional.

Campo pontuação

Através de códigos, representamos as pausas, entonações e outras informações contextuais características do código falado, conforme codificação exposta a seguir.

Tabela − Códigos de Pontuação da Transcrição Ortográfica

Códigos	Sinais de pontuação	Códigos	Sinais de pontuação
01	**.**	39	**?...—**
02	**:**	40	**),**
03	**...**	41	**—.**
04	**,**	42	**—;**
05	**;**	43	**(...)...**
06	**!**	44	**!;**
07	**?**	45	**, (...)**
08	**?,**	46	**?:—**
09	**?;**	47	**,—**
10	**?...**	48	**?, (...)**
11	/ (corte de palavra)	49	**—,**
12	**(...)** (gaguez)	50	**?(**
13	**?!**	51	**/(...)**
14	**!...**	52	**/;**
15	**—**	53	**/—,**
16	**...—**	54	**?—;**
17	**?—**	55	**!—,**
18	**:—**	56	**?).**
19	**?...,**	57	**!:**
20	**!,**	58	**—...**
21	**!...,**	59	**/?**
22	**/,**	60	**!—;**
23	**!—**	61	**/.**
24	**(...),**	62	**...;**
25	**/—**	63	**...—;**
26	**?:**	64	**? (...),**
27	**(**	65	**...:—**
28	**)**	66	**).**
29	**?)**	67	**?...—**
30	**;(...)**	68	**...:**
31	**...,**	69	**);**
32	**(...)—**	70	**...—,**
33	**—(...)**	71	**!:—**
34	**! (...)**	72	**(...)—,**

Códigos	Sinais de pontuação	Códigos	Sinais de pontuação
35	**? (...)**	73	**—:**
36	**!...—**	74	**;—**
37	**?—,**	75	**?),**
38	**...,—**	76	**(( ))**

O emprego de tantos códigos de pontuação, em que os signos de pontuação – ponto, ponto de interrogação, ponto de exclamação, vírgula, ponto e vírgula, dois pontos, reticências, parênteses, travessão –, aparecem isolada ou simultaneamente, dois ou mais, foi um outro recurso de que nos valemos, para representar as pausas, as diferentes entonações e outras informações contextuais características do código falado. Observamos que a pausa real – pausa efetivamente realizada pelo informante, assinalada, na transcrição fonética, pelo código 01 no campo *juntura sílaba final* – nem sempre coincide com um signo de pontuação, representado no campo *pontuação* da transcrição ortográfica. Isso porque, para o emprego dos signos de pontuação, na falta de uma gramática da língua oral que nos guiasse nessa tarefa, combinamos a orientação ditada pelas regras de pontuação da gramática normativa com a tentativa de transcrever certos dados do código falado carentes no código escrito.