

DESENHO E ANÁLISE DE UM CORPUS DO PORTUGUÊS CONTEMPORÂNEO

Maria Tereza Camargo Biderman
Universidade Estadual Paulista (UNESP)

Um corpus da língua escrita do Português Contemporâneo (1950 a 1990) constitui a base do **Dicionário de Frequências do Português Contemporâneo** cuja versão preliminar concluímos recentemente. A colega Profa. Guiomar F. Calçada da USP foi nossa colaboradora nesta pesquisa; contamos ainda com a ajuda de dois engenheiros eletrônicos e informáticos: Paltônio Daun Fraga e Raphael Biderman.

Este corpus compõe-se dos seguintes subcorpora, ou gêneros:
literatura romanesca (romances e contos): 1.394.855 palavras
literatura dramática: 620.386 palavras
literatura técnico-científica: 1.223.605 palavras
literatura jornalística: 1.458.174 palavras
literatura oratória: 442.172 palavras¹.

Devo lembrar que o termo *palavra* não está sendo usado com rigor técnico, mas refere tão somente cada uma das formas ocorrentes na base textual e mais precisamente "o espaço entre dois brancos", que é o critério com que opera o computador. De fato, o computador só conta diferentemente se a sequência de caracteres estiver unida por hífen como em *lagarta-rosada*, *lambe-lambe*, *milk-shake*, *mina-d'gua*, *obra-prima*, *tamanduá-bandeira*, *ultra-som*. Foi feita a frequência bruta das formas deste corpus de 5.139.196 palavras. Esse banco de dados estatísticos constitui-se no fundamento do dicionário de frequências.

Para poder avaliar os resultados quantitativos estabeleci os parâmetros indicados a seguir:

frequência 1 a 10 : baixa frequência
frequência 11 a 80 : frequência média
frequência acima de 80 : alta frequência

Lembro que o especialista em Estatística Léxica, Paul Rivenc, na pesquisa realizada em Portugal para a seleção do vocabulário do *Português Fundamental* [inquérito da frequência] estabeleceu uma fórmula para calcular o limiar de frequência [= 40]. Como esta pesquisa utilizou um corpus de 1 milhão de ocorrências, no presente caso, o limiar 80 parece razoável para a identificação das altas frequências.

Nesta versão preliminar do **dicionário de frequências** fizemos várias constatações a respeito do léxico contemporâneo que serão comentadas neste documento. Tais comentários serão direcionados precipuamente para a análise das baixas frequências; quanto às altas frequências, quero apenas apontar algumas evidências. Como em outros trabalhos anteriores, meus e de outros especialistas que trabalharam com lexicostatística, constatamos que as palavras gramaticais têm frequência muito elevada (artigos, pronomes pessoais, demonstrativos, preposições, conjunções, verbos auxiliares, sobretudo ser, estar, ter). Não pretendo examinar aqui e agora o resultado relativo a tais lexemas. Contudo, já apresentei uma pequena conclusão sobre o vocabulário mais frequente em comunicação feita no Congresso Internacional sobre a Língua Portuguesa, realizado em Lisboa em abril de 1994 – **Vocabulário Fundamental, Cultura e Sociedade**.

Uma constatação surpreendente para mim foi a de que os estrangeirismos, especialmente anglicismos, são quase irrelevantes no conjunto geral do corpus. Na língua oral e na comunicação diária no Brasil, temos muitas vezes a falsa impressão de que o inglês está avançando maciçamente sobre o português, a ponto de desfigurá-lo irremediavelmente. Não é o que evidenciam estes dados quantitativos do léxico português. Como afirmei anteriormente numa comunicação feita no Congresso da ALFAL no México em 1993: " Numa massa de dados lexicais tão grande [5 milhões de ocorrências] o peso dos anglicismos parece neutralizado pela maciça presença do vocabulário vernáculo. Colocando de outra forma: parece que são alguns domínios técnicos como a informática, que abusam dos anglicismos sobretudo na linguagem oral. Creio ainda que a mídia tende a constituir uma caixa de res-

sonância exagerada para tal fenômeno. De fato, os publicitários e jornalistas se deixam influenciar enormemente pelos valores culturais americanos; ora, a influência dos meios de comunicação de massa sobre nós é intensíssima, esbatendo-se assim nossa clareza de visão com respeito aos usos linguísticos." [*Inferências lexicográficas*,..., cf. bibl.].

Infelizmente a arquitetura do corpus foi inadequada. Embora a representatividade das diversas modalidades da língua escrita fosse satisfatória, a composição de cada subcorpus deixou a desejar. Para a literatura romanesca faltaram obras importantes e nem sempre foram as obras mais representativas de um autor as que compõem nosso corpus. Isso ocorre com E. Veríssimo (falta **O Tempo e o Vento, Incidente em Antares** e foi incluído **Fantoches**) Jorge Amado (faltam **Gabriela, Cravo e Canela, Dona Flor e seus Dois Maridos e Velhos Marinheiros**), Autran Dourado, Lígia Fagundes Telles (só está **Ciranda de Pedra**). Inversamente, alguns autores e obras deveriam ser excluídos, a despeito de sua importância literária. É o caso especificamente de Guimarães Rosa (aqui com **Sagarana e Ave, Palavra**), J.C. Carvalho (**O Coronel e o Lobisomem**), Mario Palmério (**Vila dos Confins**) e Edgard Lopes (**Travessias**). Guimarães Rosa, autor conhecido e celebrado por sua fantástica criatividade lexical, possui um estilo muito desviante da norma; possui um vocabulário riquíssimo e originalíssimo, porém, não é representativo do léxico comum do português. Vou apontar muitos exemplos mais abaixo. Em geral esses autores atestam o vocabulário mais idiossincrático do corpus. Outra inclusão inadequada foi a de **Boca do Inferno** de A. Miranda, que relata biografia romanceada de Gregório de Matos. De fato, uma história situada no século XVII traz à baila um vocabulário não contemporâneo.

Os dados estatísticos resultantes revelam claramente o desequilíbrio no desenho do corpus, sobretudo nas baixas frequências, pois o vocabulário reflete o recorte de universo feito.

Por outro lado, houve desequilíbrio também no subcorpus técnico-científico, sobretudo porque estão aí incluídas algumas obras excessivamente específicas sobre assuntos pontuais. É o caso de alguns textos da coleção *Primeiros Passos*, tais como: **O que é benzeção, O que é capoeira** e também: **Escara, problemas de hospitalização**. Os resultados estatísticos refletem essa arquitetura desequilibrada. A palavra *benzeção* ocorreu 14 vezes no corpus (mais de 5 milhões), 12 vezes

só na obra referida. *Capoeira* ocorreu 103 vezes no corpus, 47 só na obra referida, ou seja, quase a metade das ocorrências. *Escara* ocorreu 12 vezes no corpus, todas elas na obra citada. Assim as palavras *benzeção*, *capoeira* e *escara* totalizam uma frequência maior do que sua importância real no vocabulário comum da língua. Embora tivéssemos planejado obras de divulgação técnico-científica e não obras de nível mais teórico, é preciso cuidado com estas coleções de obras de divulgação; de fato, alguns tópicos tratados em tais coleções são genéricos [cf. *O que é direito*, *O que é justiça*, *O que é mercadoria*.] Outras, porém, são muito específicas como acima apontei. Neste gênero técnico-científico também são muito especializados temas como: **Fundamentos da Farmacologia**, **Grupos Vegetais**, **As Pedras Preciosas**, **Toxicologia Clínica e Forense**, **Antibióticos na Clínica Diária**. Talvez seja muito difícil encontrar obras de divulgação nas áreas de ciências biomédicas e biológicas, assim como em botânica, mineralogia, etc. De qualquer forma estas obras foram por demais específicas e as idiosincrasias lexicostatísticas o revelam claramente como apontarei mais adiante.

No estudo já citado, apresentado no México, **Inferências lexicográficas, linguísticas e pedagógicas de um dicionário de frequências**, comentei algumas formações lexicais com prefixos como *anti-*, *re-* e o prefixóide *não-*. Embora não tenha examinado com grande acurácia os dados relativos a neologismos formados a partir de prefixos, enquanto elaborava o dicionário, notei constantemente a presença de formas prefixadas inusitadas, que aparecem sob a pena de autores literários, se não considerarmos o caso conspícuo dos neônimos técnico-científicos. Constatei, por exemplo, a presença de criações lexicais com formas prefixais inusitadas e raras, ao invés das formas simples (não-prefixadas) do léxico comum:

a- *arrelampado* (PR)², *arrepelar* (SA, AVE), *arrepetir* (TG), *arresolver* (SA), *arrespeitar* (DE e R), *arreunidos* (SA);

con- *concolor* (SA), *consabido*, *consóror* (VID), *convales* (SA), *conviajar* (SA);

contra- *contramalhado* (AVE), *contramentir* (SA), *contrapiar* (SA);

des- *desalongar-se*, *despulo*, *desterrestre*, *desvago*, *desvoltar* (essas 5 palavras em AVE); *despiolhar*, *desvontade* (em TR) e *desvontadoso* (CL); *desacontecer*, *desassar* (SA) *desavançar* (TR), *desberganhar* (SA), *descomidação* (TR), *descomparecer* (CL),

desconhecimento (PR), *desdar* (TR), *desdemente* (AVE), *desencarar* (TR), *desensofrido* (CL), *desexplicar* (AVE), etc.

O número destas criações é significativo e muito evidente sobretudo na obra roseana. Parece que E.Lopes imita G.Rosa quanto a criações lexicais, especialmente neste aspecto.

Outro fato digno de registro: algumas palavras de frequência baixa, porém, com ocorrências superiores a 4/5 são registradas apenas numa única obra ou quase. Veja-se:

almiranta (nau -): só em BOI
alcouce: 10 em BOI³
almandina: 8 em PEP
cádmio: 19 ocorrências em TC e 1 em CF
coclear: 5 em ANT
colchicina: 11 ocorrências, 10 em TC
conina: 6 em TC
cormo: 5 em GV
cormófitas: 5 em GV
coojornal: 26 ocorrências em ANB
coríndon: 19 em PEP

Dei apenas alguns exemplos mas este fato é notório. É verdade que não se pode falar em leitmotiv, pois a frequência destas palavras não é alta. Antes julgo que essa característica aponta para um fenômeno léxico-semântico que há muitos anos já detectei quando elaborava minha tese de doutoramento (**Análise Computacional de Fernando Pessoa**. Ensaio de Estatística Léxica. USP, 1969). Notei então no vocabulário dos três principais heterônimos de Fernando Pessoa algo de semelhante. Certos vocábulos só apareciam em Alberto Caieiro, outros só em Ricardo Reis, e outros só em Álvaro de Campos. Pode-se quase afirmar que um tema ou tópico de certa forma já seleciona seu vocabulário específico; e isso é tanto mais verdadeiro quando se trata de vocabulário científico. A propósito: outro caso típico é o de nomes de pratos e comidas, documentados quase exclusivamente na obra **À moda da casa da amizade – 745 receitas testadas e aprovadas** (MCA). Também notei o mesmo fenômeno quando estudei os arquivos do **Português Fundamental** no Centro de Linguística da Universidade de Lisboa (1986/1987). Aí também termos relativos a comidas e culinária só apareciam em receitas de cozinha. Lembro-me que a palavra *clara* (clara de ovo) só ocorria neste contexto.

A esmagadora maioria das palavras raras, incomuns ou técnicas são *substantivos*. Contudo, tem-se eventualmente algum adjetivo, ou mais raramente ainda um verbo. Exemplos:

acontecete, adj: 1 (TR)	desassar, v: 2 (SA)
acorcundado, adj: 1 (SA)	desavançar, v: 1 (TR)
azunhar, v: 1 (FR)	desberganhar, v: 1 (SA)
cadeirar, v: 1 (PR)	desbridante, adj: 1 (ESC)
corável, v: 1 (BC)	descomparecer, v: 1 (CL)
concóide, adj: 2 (PEP)	desdeixar, v: 1 (CBC)
contramalhado, adj: 1 (AVE)	desengurdir, v: 1 (CF)
contramentir, v: 1 (SA)	desenxergar, v: 1 (SA)
contrapiar, v: 1 (SA)	desexplicar, v: 1 (AVE)
conviajar, v: 1 (SA)	desfechante, adj: 1 (CF)
contravir, v: 1 (MAR)	desfeliz, adj: 2 (SA)
contuso, adj: 1 (TR)	desfrejar, v: 1 (SA)
cornibaixo, adj: 1 (AVE)	desganar, v: 1 (TR)
corombó: 1 (SA)	desimportar, v: 1 (CBC)
corográfico, adj: 1 (AVE)	desinvocar, v: 1 (SA)
corricar, v: 1 (SA)	desladear, v: 1 (AVE)
crapudo, adj: 1 (AVE)	desmedusar-se, v: 1 (AVE)
crocaz, adj: 1 (AVE)	desreprimir, v: 1 (REA)
desalongar, v: 1 (AVE)	destamanho, adj: 1 (SA)
desamontar, v: 2 (SA)	desterrestre, adj: 1 (AVE)

Vê-se nesta amostragem que G. Rosa contribui para a maioria dos vocábulos desta lista, o que prova mais uma vez que ele é um grande inovador e criador de palavras novas na língua. Por fugir dos usos normais na comunidade dos falantes, não se recomenda como fonte ideal para documentação e/ou abonação de dicionários da língua.

Excepcionalmente alguma outra categoria registra forma rara ou obsoleta. É o caso de *Inhor* como forma de tratamento, que ocorreu em *A Grande Estiagem* de Gondim Filho (LD), e em *Os Sete Pecados Capitais*, contos de Guimarães Rosa (LR).

Às vezes as ocorrências de palavras raras e inusitadas são causadas por referências literárias, ou por citações.

Isso ocorre com *almotacé*, utilizada ao fazer-se referência à obra e personagens de Machado de Assis; idem com a ocorrência de *azinha* na citação de provérbio por E. Lopes em *Travessias* (TR):

... porque a tempo e azinha, o ferro e o pau são medicina.

Também em *O Pagador de Promessas*, Dias Gomes introduz a palavra *aruande* ao citar os versos de uma cantoria:

Mestre Do Coro
(canta)
Sinhazinha que vende aí?
vendo arroz do Maranhão
meu sinhô mando vendê
na terra de Salomão
aruande
camarado.

Coro:
e e
aruande
camarado

Fruito, forma arcaica de *fruto*, aparece numa paráfrase que O. Lessa faz de Camões (BH): "Inês de Castro, posta em sossego, de seus anos colhendo o doce *fruito*". E *fermosa*, forma também arcaica é citada por G.Rosa num entrenó em Sagarana:

Teus olho tão singular
Dessas trancinhas tão preta
Quero morrer em teus braço
Ai *fermosa* Marieta

Também curiosa é a presença inesperada e até mesmo esdrúxula de palavras como *confrada*, *confreira*, *consóror* no corpus. Ocorreram num texto jornalístico da revista *Vida Doméstica* da década de cinquenta, em que o autor de artigo/ coluna sobre a língua – "Nos domínios do idioma" – discute a esdrúxula questão do feminino de *confrade*. Vai comentando as informações dos dicionários mais respeitados naquela época – o Caldas Aulete, o Moraes, o Cândido de Figueiredo e outros – bem como as lições de linguistas e filólogos como Gonçalves Viana e Botelho do Amaral.

Todos esses comentários pretendem explicar o fato de que não só devemos ter grande cuidado com os textos (ou obras) selecionados para compor um corpus, como também as passagens e fragmentos selecionados. A única exceção para essa cautela seria a alternativa da utilização de um corpus gigantesco, mais ou menos da ordem de 100 milhões de ocorrências. Este procedimento evitaria a interferência indevida de recortes peculiares do universo discursivo e consequentemente do vocabulário correlato.

Registrei a pálida presença de dados sobre a natureza brasileira (fauna e flora) – quase nada – evidenciando a necessidade de incluirmos obras/ textos sobre esse domínio.

Há um enorme volume de *hapax legomena* (frequência 1) e também de frequências muito baixas como 2, 3; e mesmo 4 ou 5. Essas lexias contribuem para a totalização de muito mais da metade de todas as entradas do dicionário, ou seja, a esmagadora maioria do vocabulário do corpus. E aí predominam os hapax. Isso indica que ocorre, em alta escala, um uso idiossincrático da língua extremamente evidente numa consulta sistemática ao *dicionário de frequências*. Isso confirma observações já feitas por todos os especialistas em Estatística Léxica em línguas em que essas pesquisas vêm sendo feitas de longa data: francês, espanhol, inglês [cf. bibliografia, sobretudo os *frequency dictionaries* das línguas românicas elaborados por Juilland et alii]. Em conferência feita na UNESP, Campus de Araraquara, no **Simpósio de Lexicologia, Lexicografia e Terminologia** (outubro de 1994), no trabalho intitulado *Processos de Formação do Léxico Português Contemporâneo [variedade brasileira]* Sandmann apresentou dados sobre a variedade e pletora das criações individuais em todos os níveis (língua oral e escrita, de letrados e analfabetos), mostrando que esse é um vezo, ou um dom de todos os falantes.

Neste corpus os hapax legomena são típicos de dois gêneros: literatura romanesca e literatura técnico-científica. No caso do subcorpus LR (romances e contos), a causa evidente é que os escritores são os usuários da língua que maior preocupação têm com usar as potencialidades oferecidas pelo sistema linguístico para criar palavras novas com fito artístico e estilístico na sua ânsia de criar algo "novo", buscando evitar vocábulos que consideram desgastados pelo uso. O caso mais típico é Guimarães Rosa. Contudo, para que uma palavra entre para o

grande fluxo do uso linguístico na comunidade dos falantes é preciso que outros usuários passem a servir-se dela, o que raramente ocorre com essas criações idiossincráticas. A seguir, darei uma pequena amostra destes neologismos literários, indicando com uma sigla o nome da obra. Fiz uma análise exaustiva das letra A e C, que são as maiores do dicionário [A: 1.116.871 megabytes; C: 1.535.488 megabytes], bem como da letra D, que é uma das maiores.

adágamo, sub: 1 OE	camueca, sub: 2
agargalado (a), adj: 1 SA	canardo, sub: 1 AVE
aíva, adj: 1 TR	cangussu,adj (onça cangussu): 1 SA
ajoujado, adj: 1	chouto: 5 SA: 2, CBC: 2
ajoujo, sub: 1 SA	cocléia, sub: 2 SA
alaranjo, sub: 1 SA	colverde, sub: 1 AVE
alastrim, sub: 1 CL	cona (s), sub: 3 BOI
alerão (alerões), sub: 1 CF	concolor (es): 2 SA
alestado, adj: 1 CF	concriz: 2 PR
aleta, sub: N:1; SA: 1	coninho, sub: 2 BOI
aligator, sub: 1 AVE	congado: 1 SA
aloendro (s), sub: 1 SA	congro: 1 AVE
alvoroto, sub: 1 TR	conhecença: 1 PR
arcanjado, adj: 1 BH	conquilho: 1 AVE
arejador, adj: 1 TR	conselhagem: 1 CL
astuciador (a), adj : 1 SA	conselheirismo: 1 CL
atafona, sub : 5 CL	contuso,adj: 1 TR
atoinha, adv.: 4 SA	coraleira: 1 SA
atro, adj: 1 AVE	cormorão: 1 AVE
atuarial, adj: 2 ANI	cornibaixo: 1 AVE
azo, sub: TR	cornimboque: 1 SA
bafagem: SA	corrubiana: 1 AVE
baeteira: AVE	crimbamba: 1 CF
babirussas: 1 AVE	crocaz: TR
cacaió(s), sub: 1 ML	culapada: AVE
cafraria, sub: 1 TR	croticos: SA
caguinxo, sub: 4 SA	crusta: SA
calanga, sub: 1 CF	curiboca: SA e TR
calão (ões), sub: 1 SA	curimatá: SA
cambia: 1 AVE	dispautério, sub: 1 TR

Outro gênero que abunda em palavras raras na língua comum é a linguagem técnico-científica. Nesse caso pela especificidade de que se reveste o processo de nomeação nos diversos domínios do conhecimento. Aliás, a maioria das palavras deste domínio são substantivos. De fato, é por causa desse universo conceptual que o vocabulário de qualquer língua moderna de civilização tende a ampliar-se e a agigantarse. Segue uma pequena amostra de termos técnico-científicos que só ocorrem nesta modalidade de discurso, ou quase exclusivamente nela.

arreflexia: 1 TC
 arseniacal: 1 TC
 arsenopirita: 1 PEP
 arsina: 2 TC
 autoralismo: 1 FI
 aventurina: 1 PEP
 aventurino: 1 PEP
 azapetina: 1 TC
 cabimicina: 1 ANT
 caimbozeira: 1 CA
 calcemia: 1 NFN
 calcitonina: 1 NFN
 californita: 1 PEP
 condricte: 1 GAN
 conoscópio: 2 PEP
 cordofônio: 1 FN
 crisofíceas: 1 GV
 cristaluria: 1 ANT
 cotrimoxazol: 1
 crocidolita:
 dibenzociclo: 1 TC
 dibenzociclo-heptadieno: 2 TC
 dibenzociclo-heptatrieno: 2 TC
 diazepam: 1 TC
 diazinon: 1 TC
 dicumarin: 1 TC
 dicumarol: 2 TC
 difenadione: 1 TC
 digaleno: 2 TC
 digitoxina: 2 TC
 dionina: 3 FF

Obras como **Toxicologia Clínica e Forense (TC)**, **As Pedras Preciosas (PEP)**, **Antibióticos na Clínica Diária (ANT)**, **Grupos Vegetais (GV)**, **Grupos Animais: Embriologia dos Cordados (GAN)**, **Fundamentos da Farmacologia (FF)** concorrem para a presença da maioria de hapax legomena e termos raros. É possível que isso se deva a uma escolha inadequada dos textos, mas também é possível que se trate de áreas por demais específicas cujo vocabulário realmente só seja utilizado nesses domínios.

Um caso, à parte, e curioso, no interior deste gênero (LT) é o das denominações de tribus indígenas e/ou suas línguas, bem como de elementos destas culturas. É significativo o número de entradas lexicais extraídas quase exclusivamente da obra **Introdução à Antropologia Brasileira (IA)**: aipi, ajagua, ajajeni, aje, araguaco, arari,arauá, arauaca, aravaque, ariti, aroá, aroço, aroaqui, arua, atorayo, aturati, atunai augutgê, caberre, caiurucrê, cuicuro, cujijeneri, culino, cumanaxó, cunibo, cupaná, cuati, cuaracy, craó, crenaque, crengê, cretaxó, cricatagé, cotoxó, covene, crangê, curina, tagé, uabijana, uacaica, uaimaré, uainamari, ualiperi-dákeni, uariperi-dakeni, uariparevi, uarequema, uariquena, uaree, uacaica, uchikrin, etc. Outro dado interessante é que várias dessas denominações não são exclusivas, isto é, existem variantes para o nome do povo. Pode ser que se trate de problemas relativos à transcrição de um vocábulo de uma língua para outra diferente como o português; nesse caso as variantes seriam decorrência das imprecisões na interpretação de sons/fonemas da língua indígena de partida. Eis alguns casos: *arauá*, *arauca* ou *aravaque*; *aroá*, *aroaco*, ou *aruaqui*; *cângite*, *cangiti*, ou *cangutu*; *copixó* ou *copoxó*; *crangê* ou *crengê*. São numerosos estes casos. Citei alguns como exemplo.

Os outros três gêneros são menos marcantes que os dois já referidos. A linguagem jornalística parece constituir uma média dos demais tipos de discurso. São incomuns as idiossincrasias no jornalismo; entre outros fatos, algumas palavras incomuns documentadas aí advieram da inclusão de coletâneas de crônicas publicadas regularmente num dado jornal (Folha, Estado, Globo, etc.). Pode-se identificar até um laivo de pretensão literária por parte do cronista/jornalista. Na linguagem dramática, algumas das idiossincrasias ficam por conta de vocábulos da linguagem chula e obscena, que outros gêneros evitam geralmente. Isso é característico de obras como **Navalha na Carne** de Plínio Marcos, **Ópera do Malandro** de Chico Buarque. Dois outros

casos peculiares são: o **Pagador de Promessas** de Dias Gomes e o **Teatro** (4 peças) de Guilherme de Figueiredo. A primeira peça contém termos regionais ausentes do resto do corpus e a segunda, palavras raras relativas à cultura grega, em virtude de seus temas.

Embora esta análise preliminar ainda seja insatisfatória, já permitiu um primeiro diagnóstico sobre o léxico do português contemporâneo (língua escrita), bem como sobre quais seriam os melhores procedimentos para desenhar um corpus linguístico a fim de compor uma base textual para análise da língua.

Lista das siglas

LR = literatura romanesca; LT = literatura técnica; LJ = literatura jornalística.

ANB [LR]: J.E. Veríssimo, O ANALISTA DE BAGÉ, 1982.

ANT [LT]: A.L. Fonseca, ANTIBIÓTICOS NA CLÍNICA DIÁRIA, 1984.

AVE [LR]: L.G. Rosa, AVE, PALAVRA, 1970.

BOI [LR]: A. Miranda, BOCA DO INFERNO, 1989.

CBC [LR]: A.Bosi, O CONTO BRASILEIRO CONTEMPORÂNEO, 1977.

CF [LR]: P. Nava, CHÃO DE FERRO, 1976.

CL [LR]: J.C. Carvalho, O CORONEL E O LOBISOMEM, 1978.

DE [LR] : D. Trevisan et alii, OS DEZOITO MELHORES CONTOS DO BRASIL, 1968.

ESC [LT]: M.C. Campedelli, R.R.Gaidzinski, ESCARA, PROBLEMAS DE HOSPITALIZAÇÃO, 1987.

FR [LR]: O.G.R. Carvalho, FICÇÃO REUNIDA, 1981.

GV [LT]: E.A. Brito, GRUPOS VEGETAIS, 1982.

PEP [LT]: R.R. Franco, J.E.S. Campos, AS PEDRAS PRECIOSAS, 1965.

PR [LR]: A.Suassuna, A PEDRA DO REINO, 1967

REA [LJ]: REALIDADE, 1969.

SA [LR]: J.G. Rosa, SAGARANA, 1951.

TC [LT]: H.R. Alcântara, TOXICOLOGIA CLÍNICA E FORENSE, 1985.

TG [LR]: J.Amado, TOCAIA GRANDE, 1984.

TR [LR]: E. Lopes, TRAVESSIAS, 1980.

VID [LJ]: VIDA DOMÉSTICA, 1955.

Notas

- ¹ A seleção e coleta dos textos foi feita pelo colega da UNESP, Prof. Francisco da Silva Borba. Também foi esse professor quem conseguiu recursos da FAPESP e do FUNDUNESP para a compra de equipamento computacional e para o pagamento dos digitadores e revisores.
- ² As siglas referidas neste trabalho [como esta: PR] encontram-se alistadas e identificadas ao fim deste texto.
- ³ Conferir o que disse acima sobre a inclusão indevida desta biografia romanceada de Gregório de Matos situada no século XVII.

Bibliografia

- BIDERMAN, M.T.C. " *Inferências lexicográficas, linguísticas e pedagógicas de um dicionário de frequências*", X Congresso Internacional da ALFAL, Veracruz, México, 1993.
- Português Fundamental. Vocabulário e gramática*. Lisboa, Instituto de Investigação Científica, Centro de Linguística da Universidade de Lisboa. Tomo I, 1984.
- Português Fundamental. Métodos e Documentos*. Tomo I e II. Lisboa, Instituto de Investigação Científica, Centro de Linguística da Universidade de Lisboa, 1987.
- RIO-TORTO, G.M. *Formação de Palavras em Português*. Coimbra, 1993 (tese de doutoramento).
- SANDMAN, A.J. *Formação de palavras no português brasileiro contemporâneo*. Curitiba, Icone Editora, 1988.