ANÁLISE LEXICOMÉTRICA DAS NARRATIVAS DE SANTARÉM

Abdelhak ***Razky***
Universidade Federal do Pará

- **RESUMO:** *O objetivo desse artigo é enfatizar o papel e a importância dos métodos estatísticos no processamento e exploração de quantidades maiores de textos. Uma das ferramentas tradicionais usadas neste trabalho é o programa Tact. Através de uma segmentação das formas lexicais de 54 narrativas orais da cidade de Santarém, tentamos identificar algumas características lingüísticas tendo como embasamento um dicionário de freqüência das palavras e "concordances".*

- **PALAVRAS-CHAVE:** *Lexicometria; Lexicologia; Narrativas Orais.*

- **ABSTRACT:** *The object of this article is to emphasize the role and importance of statistical tools in the processing and exploration of large amount of texts. One of the traditional tools used in this work is the program Tact[1]. Through a segmentation of 54 oral narratives of the city of Santarém, situated in the north west of Brasil, into lexical forms, we try to identify some linguistic characteristics based on word frequency and and concordances.*

- **KEY WORDS:** *Lexical Statistics; Lexicology; Oral Narratives.*

Introdução

A emergência do computador a partir dos últimos vinte anos levou lingüistas de diferentes áreas a se preocupar com números maiores de dados. Analisar quantidades de dados lexicais não é mais um obstáculo para a interpretação, já que, após o levantamento do *corpus*, a fase de cálculo e comparação é facilitada pelo computador.

Os instrumentos de lexicometria ajudam o pesquisador a estabelecer características do discurso, seja político, religioso, narrativo ou de outro gênero.

---

[1] Tact software de gerenciamento de "Concordances" e colocações (collocations) da Universidade de Toronto Canada.

Neste artigo, procuramos descrever algumas características lingüísticas das narrativas orais contadas por moradores do município de Santarém, situado no Baixo Amazonas, Estado do Pará.

Métodos lexicométricos

A comparação quantitativa produz índices pertinentes sobre as estratégias do discurso e sobre as ideologias subjacentes (ver, por exemplo, o discurso dos sindicatos citado por Boyer (1996, p. 183). A lexicometria é uma pesquisa sobre o específico, tendo este mais importância do que o especificado. Boyer (ibid) confirma isto: «Les raisons d'être des mots l'emportent en intérêt sur leur être de raison. Bref le spécifique prime le spécifié».

De fato, o exame de *corpus* maiores através de métodos sistemáticos permite, se não garante, acesso a evidências de qualidade que antes não estavam disponíveis. Em lexicografia, sobretudo, a diferença é notável entre dados coletados pelo computador e dados coletados pelo leitor humano que sofre a influência da intuição no processo de decisão sobre o que pode ser incluído num dicionário. Sinclair[2] (1991, p. 4) observa que lingüistas se preocupavam mais com a intuição e confiavam mais nas suas intenções sobre o texto do que no texto. A pesquisa era feita mais sobre a intuição do que sobre a linguagem.

Kucera & Francis (1967) foram alguns dos pioneiros na criação de *corpus* de maior importância para o inglês.

[2] «Students of linguistics over many years have been urged to rely heavily on their intuitions and to prefer their intuitions to actual text where there was some discrepancy. Their study has been more about intuition than about language.»

Corpus e instrumentos

O *corpus* que serviu para este trabalho consiste em 48 narrativas coletadas e transcritas grafematicamente pelos bolsistas e pesquisadores do programa integrado IFNOPAP[3].

O *corpus* é composto de 22.597 palavras. A narrativa menor é constituída por 150 palavras, a maior por 1479 palavras.

O programa TACT usado neste artigo foi desenvolvido nos laboratórios do Quebec, Canadá. É um programa de uso simples para o neófito. Não pretendemos nesse trabalho adotar procedimentos robustos da estatística lingüística, apesar de estarmos consciente da importância de outras ferramentas de analise lexicométrica, como os programas d'André Salem (Lexicloud) dos laboratórios de St. Cloud (Paris — França), de Etienne Brunet (Nice — França), de Camlong (Stablex) e Max Reinert (Alceste) (Toulouse — França), e dos programas exploratórios dos laboratórios da Suíça (EDA) e de outros.

Nosso objetivo é mais simples, pois visamos a mostrar a possibilidade de repertoriar e contabilizar conjuntos de caracteres para análises morfossintáticas, lexicais, discursivas e interpretações objetivas a partir de resultados de segmentação e concordâncias (concordances) lexicais que pertencem a discursos quantitativamente maiores.

As freqüências lexicais foram categorizadas em quatro níveis

- > 100
- > 60 < 100
- > 30 < 60
- > 10 < 30

Os gráficos apresentados a seguir representam algumas das características do discurso narrativo de Santarém (Pará).

[3] «O Imaginário nas Formas Narrativas Orais Populares da Amazônia Paraense» coordenado pelos professores Maria do Socorro Simões e Christophe Golder.

Freqüências lexicais das narrativas orais de Santarém
Categorias gramaticais

A lista de freqüência indica que aś primeiras palavras são as categorias gramaticais. Os itens na faixa superior a 200 são:

Item	Freqüência
> o	1237
> e	733
> que	715
> a	525
> ele	523
> de	406
> aí	486
> ela	347
> eu	317
> um	311
> quando	247
> pra	244
> lá	242
> foi	238
> uma	222
> era	220
> com	207

Formas verbais

O processo de lematização[4] (lemmatization) resulta nos seguintes lemas[5] (lemmas):

Chegar
chegamos. 5
chegando. 8
chegar. 11
chegaram. 19
chegassem 1
chegava 15

ser
era 220
foi 154
eram. 11
éramos. 1

Fazer
faz 12
fazem 1
fazemos 3
fazendo 5
fazer 32
fazia 10

[4] O processo de agrupamento de formas lexicais em lemas (lemmas).

[5] Lema refere-se à noção de «palavra», como confirma Sainclair (1995, p. 173): «A lemma is what we normally mean by 'word'. Many words in English have several actual word-forms... for example the verb to give [lemma] has the forms give, gives, given, gave, giving, and to give».

chegavam. 1
chego 2
chegou. 59
cheguei 11

Começar
começaram 13
começava. 3
começavam 2
comecei 1
começo. 2
começou 30

Correr
correndo. 15
correr. 5
correram. 3
correu. 15
corria. 8
corríamos 1

Dormir
dormia. 3
dormindo. 12
dormir. 20
dormirem. 1
dormiu. 3

ir
vou 36
foi 74

Ter
tinha 149
tinham. 8
tiveram 2
tivesse 4
teve. 27
tenho 9

estar
..está. 49
estava. 143
estavam 22
tenho 9

Dar
dar 8
dava. 15
dei 2
deu 31

Dizer
diz 14
dizem 9
dizendo 3
dizer 8
dizia 20
disse 155
disseram. 12
diziam. 7

faziam. 1
farei 1
fez 16

Passar
passava 9
passavam. 1
passei. 6
passou. 29
passamos. 1
passando. 7
passar. 17

Sair
saí 4
saía. 5
saíam 2
saída 2
saído 2
saímos. 1
saímos. 2
saindo. 2
saio. 2
sair. 19
saíram. 10
saísse. 1
saiu. 37

A freqüência das formas verbais lematizadas ilustradas no gráfico 1

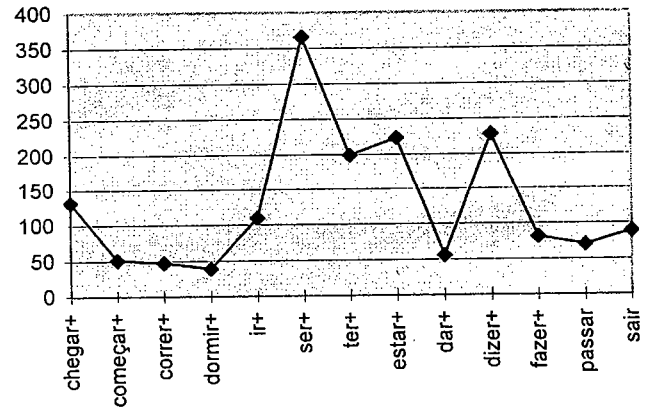


Gráfico 1

O gráfico 2 indica a predominância de formas verbais no pretérito perfeito, seguidas pelas formas verbais no pretérito imperfeito e no infinitivo.

Os verbos ser, estar e ter são produtivos e não apenas em português. Os verbos «être, avoir», em francês[6], e «to be, to have», em inglês, são os mais freqüentes nessas línguas. As formas que seguem são os verbos dizer, ir, chegar, fazer, sair, começar, dar, passar, dormir, correr.

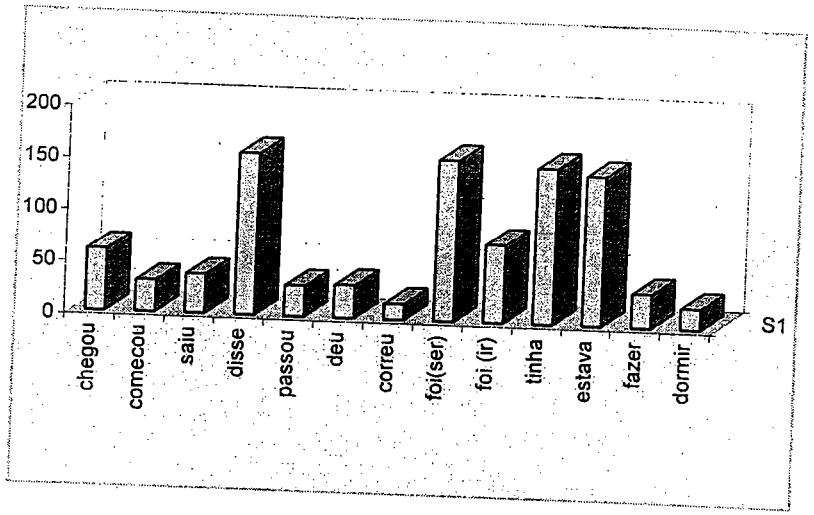


Gráfico 2

Advérbios de localização no espaço

O gráfico 3 indica um emprego considerável em termos estatísticos do advérbio 'lá', comparado a «aqui» e «ali»:

[6] Lembramos aqui uma das primeiras experiências de aplicação de freqüências lexicais na elaboração do francês fundamental (Français Fondamental) em 1956 para o ensino do francês como língua estrangeira. A freqüência mais alta é dos verbos «être» (14.083) e avoir» (11.552) (Gougenhein, 1967).

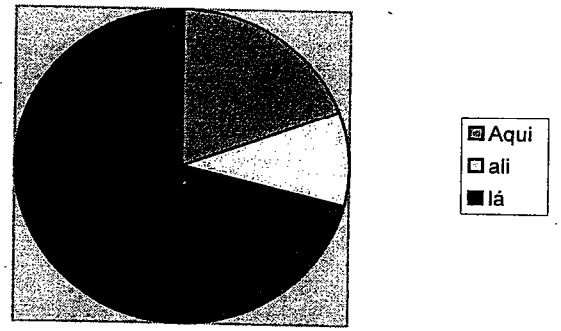


Gráfico 3

Pronomes
Masculino ou feminino, singular ou plural?

O pronome da terceira pessoa masculina 'ele', no gráfico 4, ocorre mais do que o pronome 'ela'. Há uma diferença quantitativa entre os pronomes no singular e no plural. No entanto, a presença da primeira pessoa singular é bem marcada.

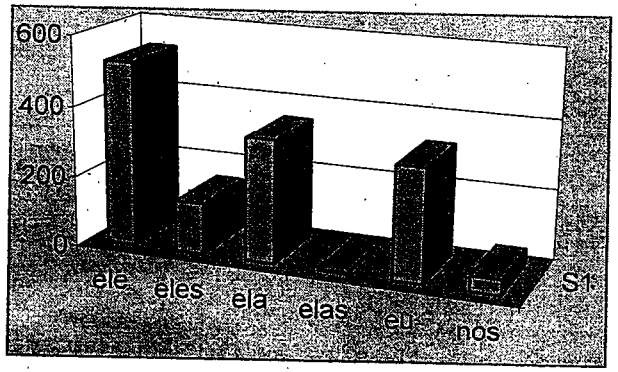


Gráfico 4

Pronomes demonstrativos

As variações dos pronomes demonstrativos nas narrativas orais indicam uma freqüência de 149 no uso do lema 'isso+' (isso, essa, isto, esta, essa essas). O lema 'aquilo+' (aquilo, aquela, aquelas, aquele, aqueles) foi o segundo mais freqüente.

aquela. 49	**isso. 48**
aquelas 8	isto. 1
aquele. 41	**essa. 36**
aqueles 3	essas 12
aquilo. 28	**esse. 49**
	esses 3
dessa 14	
dessas. 2	**nessa 22**
desse 15	nessas. 2
desses. 3	**nesse 13**
	nesta 1
	nestas. 1
	neste 1

Pronomes e Formas de tratamento

As narrativas orais de Santarém são caraterizadas pelo emprego da forma 'a gente' mais do que o do pronome 'nós'. O pronome 'você' predomina sobre o pronome 'tu' (gráfico 5). Outras marcas de oralidade são as variações de (pra/para) e (pro/para o) (gráfico 6).

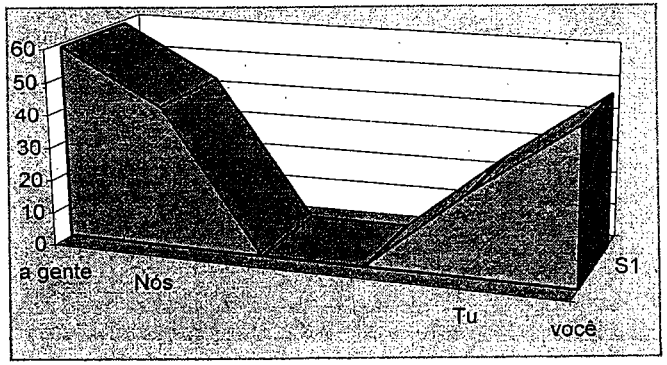


Gráfico 5

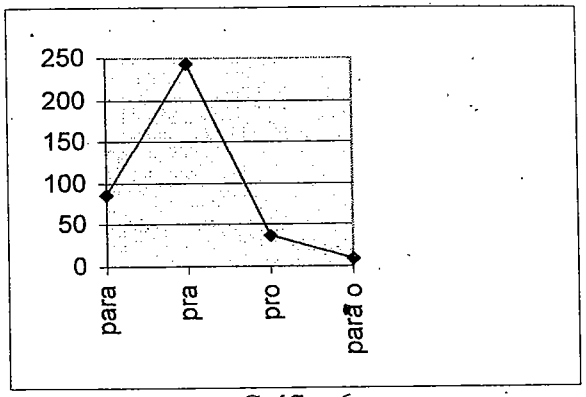


Gráfico 6

Marcadores conversacionais

O marcador 'né', no gráfico 7 abaixo, representa 86% dos marcadores. Eis alguns contextos onde aparece o 'né'. "Concordances" refere-se à realização contextualizada:

né (157)

(3) | Isso aconteceu na Vila do Curuaí, >né. Que uma noite, uma moça,
(4) enquanto ela | dormia, veio um homem, >né, todo de branco..., e
(4) né, todo de branco..., e levou, >né, bem pra perto do... do
(5) do... do rio, | bem pra beira do rio, >né, que a casa dela ficava bem
(6) | Então ele levou a moça pra lá, >né. E quando foi de manh?, a
(7) não encontrava. Foi achar a moça nua, >né, lá na beira do rio, e...
(8) ela acordou, que era um... um rapaz, >né, que parece que | gostava
(9) né, que parece que | gostava dela, >né. Que tinha abusado dela. | Só
(10) dela. | Só que um... curador, >né, desses velhos do interior,
(13) Ele apareceu e quis tirar o calç?o, >né, do João. | Só que o João
(14) Só que o João conseguiu fugir, >né. Mas dizem, né, que al
(14) João conseguiu fugir, né. Mas dizem, >né, que alguns não conseguiam
(15) escapar | não, da mão do boto, >né | Mas, até que uma certa noite,

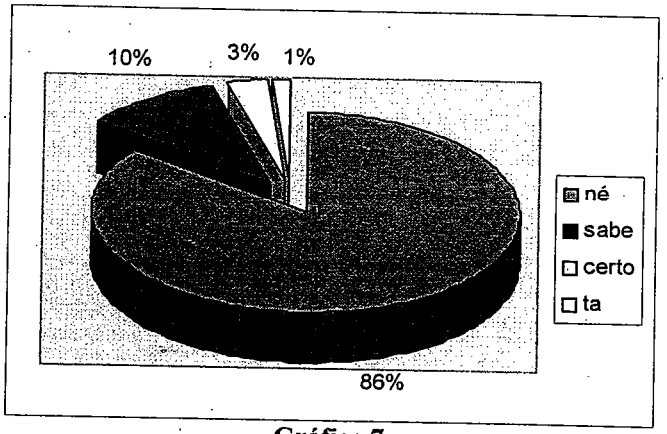


Gráfico 7

Os itens "Tarde, manhã, madrugada e noite"

O item 'noite' aparece em uma grande parte das narrações, pois a noite é um momento privilegiado, onde acontecem histórias que refletem o imaginário dos contadores. Eis aqui um trecho dos contextos "concordance" de 'noite':

noite (52)

(3) na Vila do Curuaí, né. Que uma >noite, uma moça, enquanto ela
(16) to, né? | Mas, até que uma certa >noite, antes do seu Manuel,
(60) extraordinária. | | Era uma >noite, que eu saí para pescar...
(287) mas não acreditavam! Nessa >noite, eles ouviram uns
(301) transformava em gente e passava a >noite com as mulheres. Foi um
(318) Ai, milha filha, nasceram... toda >noite cho- | ravam,
(334) quarto... | Quando foi na quinta >noite, ela já estava quase com
(1593) Eu não sabia de nada. | De >noite, o companheiro dormiu
(1754) | Eles ficaram lá dormindo a >noite toda. Quando foi ao |
(1756) porque não pude- | ram dormir à >noite. Aí, eles levaram comida
(1933) era, | mais ou menos, meia >noite. Estava claro, a lua
(1940) peixe de terçado lá embaixo,à | >noite. Aí, eu fui lá embaixo
(1948) começou aquilo. Isso era | toda >noite. Aí, começou me dar assim
(2079) Aí, ele | saiu aquelas hora da >noite pra vir21 chamar gente.
(2131) na praia. | Quando foi numa >noite, eu escutei ele brigando
(2143) meu, cunhado do meu pai, de >noite, | amando ele pra beber.

O gráfico indica as freqüências altas do lema 'noite' em relação aos lemas 'manhã', 'tarde', e 'madrugada'.

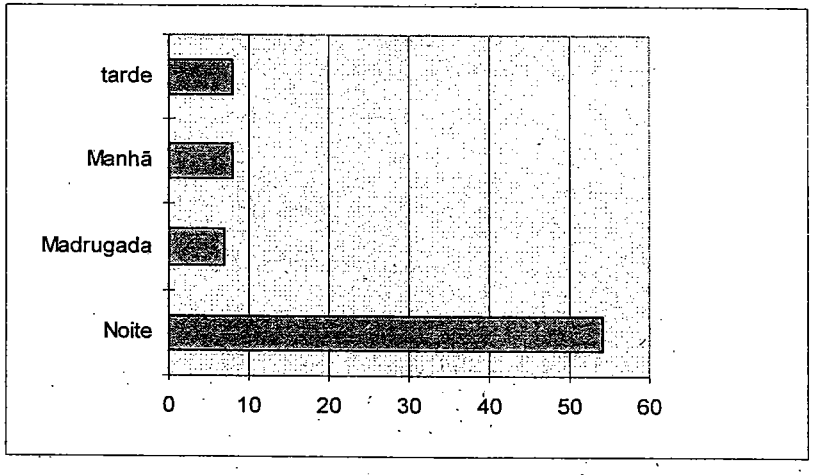
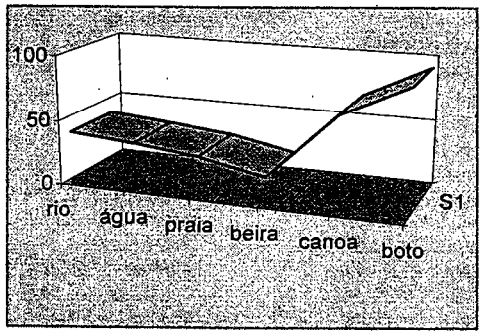


Gráfico 8

Campos semânticos do 'boto'

As histórias do boto são estatisticamente as mais representativas do imaginário de Santarém. Essas histórias estão diretamente ou indiretamente ligadas a campos semânticos, como 'canoa', 'rio', 'água', 'praia' e 'beira'.



Lista das freqüências

Freqüência > 100

> o 1237
> e 733
> que 715
> a 525
> ele 523
> de 406
> aí 486
> ela 347
> eu 317
> um 311
> quando . . 247
> pra 244
> lá 242
> foi 238
> uma 222
> era 220
> com 207
> na 202
> da 200
> do 186
> é 176
> no 163
> né 157
> disse 155
> se 152
> tinha 149
> eles 144
> estava 143
> mas 132
> em 120
> já 113
> os 111

Freqüência >60 <100

> muito 95
> boto 94
> assim 93
> dele 86
> para 85
> então 83
> mais 81
> dela 77
> meu 71
> minha 71
> canoa 69
> ficou 69
> ia 68
> me 68
> só 67
> m 66
> aqui 65
> dia 63
> as 62

Freqüência > 30 < 60

> gente 59
> vai 56
> casa 54
> porque 53
> chegou 59
> noite 52
> mesmo 51
> aquela 49
> esse 49
> está 49
> bem 48
> isso 48
> outro 48
> embora 47
> bicho 38
> coisa 38
> homem 38
> vinha 38
> onde 37
> pro 37
> rio 37
> saiu 37
> essa 36
> mulher 36
> olha 36
> vou 36
> água 33
> ali 33

> pai 47
> por 47
> você 47
> cobra 46
> até 45
> tudo 44
> nós 43
> veio 43
> como 42
> grande 42
> pessoa 41
> tem 41
> aquele 40
> cima 39
> depois 33
> fazer 32
> foram 32
> nada 32
> seu 32
> vez 32
> deu 31
> filha 31
> também 31
> começou . . . 30
> dois 30
> meio 30
> moça 30
> viu 30

Freqüência >10 <30

> caiu 17
> cara 17
> das 17
> entrou 17
> horas 17
> jeito 17
> levar 17
> longe 17
> mato 17
> passar 17
> pé 17
> pedaço 17
> à 16
> bonito 16
> caminho . . . 16
> correndo . . . 16
> fez 16
> ir 16
> nem 16
> ninguém . . . 16
> perto 16
> pode 16
> quem 16
> tirar 16
> ah 15
> ch 15
> chegava . . . 15
> correu 15
> dava 15
> caiu 17
> cara 17
> das 17
> entrou 17
> horas 17
> pode 16
> quem 16
> tirar 16
> ah 15
> ch 15
> chegava . . . 15
> correu 15
> dava 15
> desse 15
> fiquei 15
> grito 15
> matar 15
> menino 15
> nunca 15
> parece 15
> peixe 15
> pelo 15
> zagaia 15
> apareceu 14
> companheiro . 14
> criança 14
> crianças 14
> curupira 14
> diz 14
> fora 14
> índio 14
> levou 14
> lhe 14
> mata 14
> nas 14
> olho 14
> olhou 14
> outros 14
> quer 14
> podia 13
> ramal 13
> senhor 13
> todos 13
> velho 13
> amigo 12
> às 12
> boa 12
> caçar 12
> cada 12
> chamar 12
> comer 12
> contou 12
> dá 12
> deixou 12
> deus 12
> disseram 12
> dormindo . . . 12
> embaixo 12
> essas 12
> faz 12
> fica 12
> ficar 12
> filhos 12
> frente 12
> meia-noite . . . 12
> mesa 12
> peguei 12
> porco 12
> sua 12
> trás 12
> caça 11
> chegar 11
> cheguei 11

> jeito 17
> levar 17
> longe 17
> mato. 17
> passar. 17
> pé. 17
> pedaço. 17
> à 16
> bonito. 16
> caminho . . . 16
> correndo. . . 16
> fez 16
> ir. 16
> nem 16
> ninguém . . . 16
> perto 16
> repente 14
> sabia 14
> terra 14
> toda. 14
> uns 14
> andando 13
> anos. 13
> barulho 13
> começaram . . 13
> dessa 13
> ei. 13
> fogo. 13
> negócio 13
> nesse 13
> padre 13
> pegar 13
> comigo. 11
> d'água. 11
> daí 11
> dona. 11
> eram. 11
> festa 11
> ficando 11
> índios. 11
> jo. 11
> muita 11
> parecia 11
> ponte 11
> ter 11
> umas. 11
> vi. 11
> vira. 11

Conclusão

Em vez de privilegiar uma abordagem teórica, preferimos mostrar a possibilidade de se levantar hipóteses sobre aspectos diferentes do discurso, narrativo no nosso caso, a partir de freqüências e "concordances". O uso de ferramentas estatísticas permite a segmentação de quantidades maiores de dados facilitando interpretações e generalizações baseadas primeiro nos dados para evitar o risco de adaptar os dados à teoria.

Na área de lexicologia, elaborar dicionários e glossários especializados, se torna mais eficiente adotando métodos lexicométricos a partir de bases de textos maiores.

As características lingüísticas não se manifestam unicamente nas freqüências altas. No CD-Rom[7] de Santarém, elaboramos um pequeno glossário amazônico baseado nas freqüências baixas.

[7] O Cd-Rom de Santarém, que faz parte do projeto «Multimídia: educação e cultura», coordenado por Socorro Simões e Abdelhak Razky, foi apresentado no congresso internacional da Brasa, em Washington, EUA, novembro 1997.

REFERÊNCIAS BIBLIOGRÁFICAS

BOYER, Henri. *Sociolinguistique: territoire et objets*. Delachaux et Niestlé: Paris, 1996.

GOUGENHEIM, Georges. *L'élaboration du Français fondamental (1er. Degré)*, Paris: 1967.

Sinclair, John. *Corpus, Concordance, Collocation*. New York: Oxford University Press, 1995.

SIMÕES, Maria do Socorro, GOLDER, Christophe (coord.). *Santarém Conta*. Belém: Cejup, 1995.

FOSSAT, Jean-Louis, PEYTAVI, Jean Marc, RAZKY, Abdelhak. Du Dictionnaire en Machine au Dictionnaire à Lemmatisation Dynamique de L'Occitan. *Cahiers d'Etudes Romanes*. Toulouse, v. 2, n. 7, 1995.

NORMAS PARA PUBLICAÇÃO DE ARTIGOS

A **Revista MOARA** aceita propostas de artigos. Todas as colaborações são submetidas à Comissão Editorial, a quem cabe a decisão final sobre sua publicação. A Comissão reserva-se o direito de sugerir ao autor modificações de *forma*, com o objetivo de adequar os artigos às dimensões da revista ou a seu padrão editorial e gráfico.

PREPARAÇÃO DOS ORIGINAIS

Os trabalhos, obrigatoriamente *originais*, devem ser enviados em ***DISQUETE*** (cada artigo deve ter no máximo *dez páginas*), digitados em computador versão IBM (recente), usando-se programa *Word for windows* (fonte 12 em *Times New Roman*; espaçamento simples).

Ao disquete, apor uma etiqueta contendo o ***nome do(a) autor(a)***, o ***título do trabalho*** e o ***programa utilizado***.

Observação: o disquete não será devolvido a(o) autor(a), que deve manter seu arquivo para as modificações sugeridas pelos pareceristas.

APRESENTAÇÃO

A apresentação dos trabalhos deve obedecer à seguinte seqüência:

a) **Cabeçalho do artigo** (primeira folha no alto)

– Título (e subtítulo se necessário em português e *inglês* ou *francês*)

– Nome(s) do(s) autores, na ordem direta:

Ex.: *Célia **Brito***

– Filiação institucional – local de atividade de cada um dos autores, colocado abaixo dos seus nomes.

Ex.: *Célia **Brito***
Universidade Federal do Pará

– No rodapé da página poderão ser apresentadas informações sobre o trabalho e menção de auxílios institucionais se for o caso.

b) **Resumos** (antecedendo o texto)

Síntese do conteúdo do trabalho com um máximo de 150 palavras, redigida de acordo com a NB-88, da ABNT. Os resumos em ***português*** e ***inglês ou em português*** e ***francês*** devem ser acompanhados de ***três palavras-chave*** (em português e inglês ou em português e francês).

c) **Texto**

O texto sempre que possível deve obedecer à seguinte divisão: *introdução*, *desenvolvimento do tema*, com as divisões a critério do autor e *conclusão*.

d) **Notas** (não bibliográficas)

Devem ser colocadas no rodapé das páginas. As remissões para o rodapé devem ser feitas por números arábicos, na entrelinha superior.

e) **Citações Bibliográficas**

As citações no texto deverão ser feitas de duas maneiras:

- sobrenome do autor em caixa baixa seguido da data de publicação e da página quando for necessário, entre parênteses.

Ex.: Segundo Saussure (1990, p.13), "a Lingüística tem relações bastante estreitas com outras ciências";

- sobrenome do autor em caixa baixa, data da publicação e da página, quando for o caso, tudo entre parênteses.

Ex.: "A Lingüística tem relações bastante estreitas com outras ciências" (Saussure, 1990, p.13).

As citações devem ser feitas como segue:

- *um autor:* Bosi (1993);
- *dois autores:* Simões & Golder (1995);
- *três ou mais autores:* Bastos et al. (1981);
- *se for citada mais de uma publicação do mesmo autor com o mesmo ano, usa-se alínea:* Pinto (1990a), Pinto (1990b), etc.;
- *para as citações indiretas usa-se a expressão "apud" (citado por).* No texto: J. M. Costa ap. Freitas (1980). *Na referência bibliográfica deve constar apenas a obra consultada;*
- *obras sem autoria*: Manual de Teoria... (1985).

f) **Referências Bibliográficas**

Lista em ordem alfabética das obras citadas no texto. As referências devem vir localizadas imediatamente após o texto. Devem ser feitas conforme o tipo de publicação, obedecendo à seguinte ordem dos elementos:

- ***Livros e outras monografias***

Ex.: TARALLO, Fernando. *A pesquisa sociolingüística*. São Paulo: Ática, 1985.

- ***Parte de obra (capítulos, fragmentos, volumes)***

Ex.: GOMES, Severo. Informática e soberania. In: BENKOUCHE, Rabah, (org.). *A questão da informática no Brasil*. São Paulo: Brasiliense, 1985. 167p. p.30-36.

- ***Artigo de Periódico***

Ex.: GOMES, Sonia Pedrosa, ALOIA, Miriam. *Referências bibliográficas: algumas sugestões*. Boletim Abdf. Brasília, v.6, n.2, p.21-31, abr./jun.1983.

- ***Artigo de jornal***

Ex.: JOB, Fernando. Munique está em festa. *O Liberal*. Belém, 19 set 1990, p.4, cad.1.

- ***Trabalho de Congresso ou similar (publicado)***

Ex.: TARGINO, Maria das Graças. Bibliotecas universitárias e prestação de serviços: a irreverência do óbvio. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 16, 1991. Salvador, Anais... Salvador: APBED, 1991, v.1, p.400-405.

g) **Ilustrações**

As **figuras** (desenhos, gráficos, mapas, esquemas, organogramas, fórmulas, etc.) com suas legendas devem ser claramente legíveis. Devem indicar: autor, título abreviado e sentido da figura. Legenda das ilustrações, nos locais em que aparecerão as figuras, numeradas consecutivamente em algarismos arábicos e iniciadas pelo termo FIGURA. As **tabelas** serão encabeçadas e citadas como tabela, com título auto explicativo, colocado acima da mesma.

*** Importante: ***todos os trabalhos devem ser revisados por seus autores antes de serem submetidos à avaliação***.