

A WEB PARA UM LEVANTAMENTO DE FREQUÊNCIA

Claudia Maria XATARA (UNESP-São José do Rio Preto)¹

RESUMO: Tanto para a elaboração de material lexicográfico concernente às expressões idiomáticas, como para sua sistematização no ensino e aprendizagem de línguas estrangeiras, deveria ser previamente tratada a questão da frequência e uso dessas expressões. É necessário, pois, que se determine um limiar de frequência para se observar os idiomatismos em uma base textual suficientemente representativa da linguagem coloquial, registro por excelência dessas unidades fraseológicas. Assim, revela-se muito pertinente a utilização da medida de ocorrências *per million words* e da web que se impõe como a maior base textual a que se pode ter acesso.

RÉSUMÉ: Pour l'élaboration de matériel lexicographique concernant les expressions idiomatiques, ainsi que pour sa systématisation dans l'enseignement et apprentissage de langues étrangères, il devrait être traité au préalable la question de la fréquence et usage de ces expressions. Il faut donc qu'on détermine un seuil de fréquence pour observer les expressions idiomatiques dans une base textuelle suffisamment représentative du langage courant, registre par excellence de ces unités phraséologiques. Ainsi, il se révèle très pertinent d'utiliser la mesure d'occurrence *per million words* et le Web qui s'impose comme la plus grande base textuelle à laquelle on peut accéder.

1. Introdução

Primeiramente, antes de tratar do problema da frequência e da utilização da web como base textual, especificarei o tipo de item lexical cuja frequência está em questão. Refiro-me a um fraseolismo da língua geral: as expressões idiomáticas (EIs).

Após trabalhos recolhidos ao longo dos últimos 20 anos sobre as EIs, não mais se questiona que os idiomatismos não devam ser tratados como um problema marginal na pesquisa lingüística ou no ensino/aprendizagem da língua materna e de uma língua estrangeira. Ao contrário, confiáveis estudos da léxico-gramática, cujo axioma fundamental elege a frase como unidade mínima de significação, já comprovaram empiricamente que essas expressões são tanto ou mais numerosas que as construções livres (GROSS, 1994).

No que concerne a uma abordagem propriamente lexicográfica das EIs, destaca-se o paradoxo da necessária inclusão maciça de idiomatismos nos dicionários gerais de língua, o que levaria ao dobro de seu volume, e a seleção tão restritiva dessas unidades como subentradas (SALKOFF, 1995). Para deixar a questão ainda mais lacunar, em se tratando de perspectiva contrastiva, quase não se encontra material lexicográfico bilíngüe especial em nosso país, isto é, dicionários que façam determinados recortes no léxico da língua geral para se ocupar em detalhes das unidades lexicais pertencentes a esses diferentes recortes. Mesmo no cotejo com o inglês, não há uma produção muito significativa de obras de referência em português brasileiro, mas a elaboração de dicionários português-francês-português é ainda bem menos representativa.

Assim, dei seqüência a um projeto lexicográfico por ocasião do doutorado (XATARA, 1998) que acabou se concretizando na publicação de um dicionário fraseológico. Essa publicação, o *Dicionário de provérbios, idiomatismos e palavras francês-português / português-francês*, o **PIP** (XATARA & OLIVEIRA, 2002), reúne, além dos idiomatismos, 1103 provérbios e 3500 palavras eróticos-obscenos, unidades fraseológicas abordadas em pesquisas trienais que realizei anteriormente. Entretanto, o que permanece como preocupação central em minhas investigações são as EIs. Isso se reflete também nas atividades de orientadora na graduação e pós-graduação (FALCÃO, 2002; RIOS, 2003; RIVA, 2003 – três ex-mestrandos com bolsa FAPESP, todos atualmente engajados em projetos de doutorado sobre a mesma temática). No Brasil, a questão foi também destacadamente colocada por alguns estudiosos, como Tagnin (1987, 1989), Lodovici (1989), Roncolato (1996, 2001), Fornicola (1999), Ortíz Alvarez (2000), Vale (2000) e Caramori (2000).

¹ xatara@ibilce.unesp.br

Nenhum desses estudos nem tampouco as pesquisas realizadas no exterior explicitaram, entretanto, a frequência dos idiomatismos repertoriados na linguagem contemporânea, nem apresentou sistematicamente contextos nas propostas de seu tratamento lexicográfico. Mesmo o **PIP**, que teve o cuidado de excluir expressões regionais e em desuso, valeu-se apenas de fontes secundárias, ou seja, 39 outros dicionários. Mas para a produção oral e/ou textual, seria de grande valia que a informação sobre a frequência constasse na microestrutura.

O problema, porém, que resta ainda resolver de modo mais consensual entre os lexicógrafos, e com maior cientificidade, é o limiar de frequência que garanta, ou exija, a presença de uma determinada unidade léxica em um dicionário, visto que, na verdade, a frequência é condicionada por diversos fatores, como o meio social, a situação, as preferências pessoais etc (MESSELAAR, 1988).

Há mais de 30 anos, a análise estatística de Carroll, um dos pesquisadores que auxiliou na elaboração do primeiro dicionário de frequência do português (DUNCAN, 1972), estabeleceu o parâmetro SFI (*Standard Frequency Intermediate*), padrão que determina como 20 o limiar mínimo de ocorrência para uma palavra ser considerada frequente. Mas ainda aqui se trata de lexias simples ou compostas, não de lexias complexas.

Felizmente, contudo, pode-se valer na atualidade dos estudos de Colson (2003), que esclarece o seguinte à página 47:

Corpus linguistics offers a useful standard for measuring idiom frequency: the number of occurrences per million words (PMW). Many verbal idioms of English, French and Dutch (and probably of all European languages) correspond to a frequency of less than 1 PMW, i.e. their occurrence in a corpus is inferior to one in a million words. It is worth nothing that a corpus of a one million words roughly corresponds to the text of ten novels!

Esse autor traz ainda a informação que, nos bancos de *corpora* eletrônicos existentes, 70% das EIs frasais têm frequência inferior a 1 PMW e muitos idiomatismos nem são aí detectados. Então, ele conclui:

Idioms as a whole turn out to be very frequent in any given text. This does not mean, however, that a specific idiom will appear frequently, even in large corpora. Linguists or students looking for many contexts of the same idiom have to use corpora of billions of words. The Web is presently the only gigantic corpus that can stand up to that goal. (COLSON, 2003, p. 59)

Reforçando seu posicionamento, Colson (no prelo) continua afirmando: “*Although the World Wide Web does not offer a perfect solution to research on set phrases, its huge size and its multilingual character will probably make it an unbeatable competitor for traditional corpora.*”

Também Sinclair (1991, p. 19) já havia indicado que “*the study of collocation which requires very large corpora to secure sufficient evidence for statistical treatment*”.

Por sua vez, Berber Sardinha (2004) refere-se a Kilgariff e a Renouf quando diz que a web já é considerada, por importantes especialistas, um *corpus* que reúne a maior e mais variada quantidade de textos.

De fato, as bases textuais conhecidas não registram grande quantidade de textos coloquiais, que são a maior fonte para se observar as EIs em uso. Mesmo programas de gerenciamento de bases textuais, como WordSmith Tools, Folio Views ou Hyperbase, não são eficazes na identificação de unidades lexicais complexas. Abaixo segue um trecho do quadro de Berber Sardinha (2004) sobre as dimensões dos principais *corpora* eletrônicos do português brasileiro, todos muito longe dos bilhões de palavras necessárias para um levantamento de frequência com lexias complexas:

<i>Corpus</i>	<i>Palavras</i>	<i>Composição</i>	<i>Localização</i>
- Banco de Português	233 milhões	PB, escrito e falado	PUC/SP
- Usos do Português	200	PB, escrito	UNESP/Araraquara
- PB Contemporâneo	100	PB, escrito e falado	UNESP/Araraquara

Sabe-se que, para validar uma pesquisa e favorecer a verificação de hipóteses em uma amostragem da língua suficiente, o *corpus* escolhido deve ser representativo e adequado aos propósitos da pesquisa, mas todo corpus é “uma amostra de uma população de tamanho desconhecido – a linguagem em uso” (Berber Sardinha, *op. cit.*, p. 326). Então, além de muito material escrito que denota a linguagem cotidiana corrente

não fazer parte de qualquer *corpus* e de grande parte dos textos ditos “comuns” parecer reproduzir características da modalidade oral, a dimensão da web corresponde mais adequadamente tanto para a demonstração do idiomatismo em um contexto real, quanto para acrescentar importantes informações concernentes à significação e uso de cada EI. Assim interessam sobretudo textos de propaganda, entrevistas e conversas, além de todos os que circulam na imprensa.

Tudo, portanto, vem corroborar a utilização da web como base textual, ainda que não represente um conjunto controlado de textos (ou seja, um agrupamento sistemático de textos exploráveis por uma máquina, tendo sido preparados, codificados e armazenados de acordo com regras predefinidas) e que as informações encontradas pela rede não revelem fontes totalmente fidedignas, podendo ser temporárias, conter imprecisões ou mesmo erros ortográficos.

Considerando isso, realizei uma pesquisa de frequência dessas unidades especiais em português do Brasil, na web, um ambiente amplo, (em termos de línguas européias, ela tem armazenado de 1 a 50 bilhões de palavras) e heterogêneo, com extraordinária diversidade de tipos de textos e níveis de linguagem, para analisá-las pragmaticamente, observando os contextos em que ocorrem (COLSON, 2003). Essa pesquisa forneceu um novo inventário de EIs em português que me serviu de ponto de partida para os trabalhos efetuados durante o estágio de pós-doutorado no Laboratório de Análise e Tratamento da Língua Francesa (ATILF), na Universidade de Nancy 2 - França, em 2005.

Quanto ao servidor para gerenciar as buscas, o Google impõe-se por sua incomparável capacidade de procurar em mais de 4,28 bilhões de páginas de texto. Ele assegura, por exemplo, uma pesquisa em mais de 4500 jornais, revistas e agências de notícias *on-line*. Essas estatísticas são assustadoras, porém mais assustador é saber que isso não chega a corresponder nem a 1% da Internet somente, que possui bilhões de catálogos de lojas, informações detalhadas, relatórios oficiais e bancos de dados só acessados quando se preenche um formulário (**Super Interessante**, 2004). De qualquer modo, essa espantosa quantidade de páginas ultrapassa em muito o número total de palavras de quaisquer bancos textuais.

O maior entrave na busca pelas EIs, entretanto, continua sendo o modo de chegar a elas: normalmente tenta-se reduzir o idiomatismo ao núcleo mais fixo para procurar considerar as pequenas variações nas EIs de distribuição restrita ou no caso das EIs verbais.

O ideal, porém, seria poder unir uma base textual com a dimensão da Internet e uma ferramenta como a utilizada na base FRANTEXT, que possui um sistema de verbatização, isto é, um programa que reconhece as formas conjugadas, os plurais, os femininos, e aumenta assim o desempenho da busca, especialmente das EIs verbais. Ou ainda como o programa GlossaNet, um concordanceador que tem a particularidade de identificar EIs verbais em todas as formas conjugadas e em todas as variantes do verbo eventualmente possíveis. Esse serviço automático é gratuito e seria perfeito, se não estivesse ainda tão limitado, pois, em língua portuguesa, armazena apenas todas as edições de um jornal brasileiro (*O Dia*).

Na verdade, essa pesquisa não é menos válida por não se servir de um concordanceador adequado, pois o interesse maior está em poder ilustrar o uso de uma EI com um contexto e não com uma concordância, embora não haja qualquer dúvida que obter concordâncias de uma base textual facilita, sistematiza e organiza a própria escolha do contexto adequado.

2. Levantamento da frequência dos idiomatismos

Primeiramente, considerando-se, de acordo com Colson (2003), que a frequência na web é medida por PMW, determinou-se que as EIs tidas como frequentes para os seus respectivos falantes nativos são as que ocorrem a cada milhão de palavras.

Em segundo lugar, com base nos estudos reunidos por Grefenstette (2000, 2004), por Evans et al. (2004) e pela União Latina (2006), pode-se estimar que há na internet cerca de 360 milhões de páginas em francês, sendo 200 milhões em francês da França (a quarta língua da web após o inglês, o japonês e o alemão), contra 70 milhões de páginas em português (56 milhões em brasileiro e 14 milhões em português lusófono e outras variantes).

Como normalmente cada EI ocorre uma vez em cada página da web, definiu-se o limiar de frequência em 56 ocorrências para o português e 200 para o francês.

Essas ocorrências equivalem, na realidade, aos resultados oferecidos por intermédio do buscador Google, que na prática funciona como um gerenciador de texto. Evidentemente tomou-se o cuidado de se excluir do número total de resultados os resultados duplicados. E para se obter a computação léxica desejada, a pesquisa em curso priorizou as páginas do Brasil para se restringir ao português brasileiro. O mesmo foi feito no Google francês, priorizando-se as páginas da França.

O procedimento de busca que se averigou mais eficaz foi digitar o núcleo da EI entre aspas e às vezes, para evitar uma busca muito restrita, substituiu-se um termo no meio desse núcleo por um asterisco. Assim, pôde-se digitar:

- “comer o pão que o diabo amassou”, uma expressão completa, que deve constar como entrada do dicionário, e se obteve 252 ocorrências.

- “o pão que o diabo amassou”, omitindo restringir as ocorrências à forma infinitiva do verbo, e se triplicou o número: 793 ocorrências, com as formas flexionadas *comeu, comem, comendo, ter comido* e com outros verbos variantes (*aceitar, amargar, passar, sofrer*).

- “o pão que o * amassou”, que indica uma suspeita de outra UL possível, e se chegou a 836 ocorrências, pois *capeta, demônio, demo* e *ele* também são variantes.

Conclui-se, portanto, que essa EI não tem distribuição única, mas restrita.

Quanto ao uso das EIs, foram analisados os contextos em que apareceram as ocorrências, para se chegar ao tipo de situação e ao perfil do usuário que permitem o emprego adequado de cada EI.

O procedimento de análise foi por amostragem: todos os contextos a cada 2, 5, 10 ou 50 páginas buscadas pelo Google, dependendo do número de ocorrências. Então, para cada idiomatismo foi correlacionado um coeficiente de distribuição, ou seja, foi observada qual sua proporção de ocorrência nos diferentes gêneros textuais. Evidentemente foram desconsiderados contextos que revelaram a lexia complexa procurada empregada em sentido denotativo (como no caso da ambígua “colocar em pratos limpos”), pois nesse caso não se trata de EI.

Essa análise dos contextos foi decisiva para confirmar a adequação das equivalências já propostas no **PIP** ou para indicar a necessidade de novas equivalências, uma vez que sobretudo na fraseologia há muitas dificuldades de tradução devido à não-isomorfia, ou seja, recorte diferente da realidade em sistemas lingüísticos diferentes.

A escolha do contexto-exemplo também foi feita com cuidado para se selecionar, dentre todas as ocorrências contextuais mostradas, o melhor contexto melhor que ilustrasse e atestasse as especificações de seu uso. Dessa forma, dado a EI “enxugar gelo”, abonações como: “Muitas vezes o combate ao crime assemelha-se ao ato de enxugar gelo.” (Folha OnLine – Pensata – Gilberto Dimenstein – 30/08/2001) Acesso em 10/03/04.

Pareceram-nos pouco elucidativas. Mas outros exemplos como: “(...) o grande trabalho da PM é apenas enxugar gelo, porque a arma é apreendida e volta à mão dos criminosos.” (www.polmil.sp.gov.br:8080/imprensas.nsf/0/c9013cdb364279ec032569d7006bee92?OpenDocument) Acesso em 11/03/04.

Podem explicitar melhor o sentido de “atividade contínua e quase sem efeito”.

3. Organização do dicionário eletrônico

Somente após se determinar quais idiomatismos poderiam ser de fato considerados frequentes, chegou-se à nomenclatura para um projeto de dicionário eletrônico especial, francês-português/português-francês.

Para a apresentação microestrutural dos verbetes, propôs-se inserir:

- indicação da categoria gramatical de cada locução;
- uma definição parafrásica de cada expressão, na língua da entrada;
- marcas diassistemáticas (indicadas sempre que necessário: marcas diastráticas (culto ou vulgar, visto que coloquial é o nível por excelência dos idiomatismos), diafásicas (formal, informal), diatextuais (jornalístico, administrativo, literário) etc (segundo BLANCO, 2001);
- contexto extraído de texto da www;
- indicação de um conceito-chave a que remeta o sentido da EI
- sistema de remissivas (EIs sinônimas e antônimas);
- equivalências na segunda língua

A seguir, exemplos de verbetes de EIs com frequência atestada na web.

EXEMPLO DE VERBETE COM TODAS AS INFORMAÇÕES PREVISTAS :

abaixar a cabeça loc. verb. « aceitar com resignação, não reagir » O PT não pode nem ser atrevido o suficiente para desafiar tudo, nem pode ser muito dócil e *abaixar a cabeça* diante de tudo. (www.terra.com.br/istoe/1798/1798vermelhas.htm ; acesso em 30/03/04) SUBMISSÃO

Expressões similares: *colocar o rabo entre as pernas* [vulgar]; *enfiar o rabo entre as pernas* [vulgar]

Expressão oposta: *erguer a cabeça*; *levantar a cabeça*

Equivalentes em francês: *baisser la tête*; *courber la tête*; *courber l'échine*; *plier l'échine*

EXEMPLO DE VERBETE COM MAIS DE UMA ACEPTÃO :

de pernas pro ar loc. adv.

1. « sem fazer nada, na ociosidade » [sujeito: pessoa] Cheguei em casa cansada [...] afinal 2 semanas *de pernas pro ar* quebram o clima escolar. (epic.weblogger.terra.com.br/200308_epic_arquivo.htm ; acesso em 25/11/04) ÓCIO

Expressão similar: *de pernas pro alto*

Equivalente em francês: *les doigts de pieds en éventail*

2. « bagunçado, desorganizado » [sujeito: coisa] Minha vida está *de pernas pro ar*, uma bagunça geral! (www.agentepink.blogspot.com.br; acesso em 25/11/04) DESORDEM

Equivalente em francês: *sens dessus dessous*

EXEMPLO DE VERBETE DE EIS OPOSTAS:

com o pé direito loc. adv. « com sorte » Todo mundo quer começar o ano *com o pé direito*. (www.cremesp.com.br/?siteAcao=Revista&id=131 ; acesso em 10/09/04) SORTE

Expressão oposta: *com o pé esquerdo*

Equivalentes em francês: *du bon pied*; *du pied droit*

com o pé esquerdo loc. adv. « com azar » Não se pode afirmar que começou *com o pé esquerdo*, mas com certeza não é um jogo que figuraria entre os melhores. (www.pernambuco.com/diario/2003/07/09/info14_0.html ; acesso em 10/09/04) AZAR

Expressão oposta: *com o pé direito*

Equivalentes em francês: *du mauvais pied*; *du pied gauche*

EXEMPLO DE VERBETE DE EI EQUIVALENTE:

du mauvais pied loc. adv. « avec malchance ou mauvaise humeur » Depuis quelques semaines, Jacques, 28 ans, avait pris l'habitude de se lever *du mauvais pied*. Il s'en voulait d'être irritable à l'égard de Martine [...] (psychologuebraem.free.fr/sexualite.htm ; accès le 14/04/05) MALCHANCE

Expression similaire : *du pied gauche*

Expressions opposées : *du bon pied* ; *du pied droit*

Équivalent portugais : *com o pé esquerdo*

4. Informatização dos dados

A proposta é caracterizar este dicionário eletrônico como um produto interativo, pois possibilitará ao usuário fazer suas próprias seleções, escolher como chegar à informação que deseja e desenvolver uma leitura não-linear por isso. Para tanto, ao informático do ATILF, M. Etienne Petitjean, foi fornecido conteúdos textuais para todos os tipos de “objetos” deste dicionário, ou seja, tudo aquilo sobre o que o dicionário possibilitará associações em forma de hyperlinks. Com base nisso, o informático proporá códigos que permitirão identificar os objetos determinados.

Alguns exemplos de objetos previstos:

- perspectiva semasiológica:

- critério de busca por EI (quando o consulente conhece a forma canônica da EI): digita-se ou se seleciona a EI procurada na barra alfabética de entradas e se obtém o verbete da expressão.

- critério de busca por componente da EI (consulta mais ampla): digita-se ou se seleciona qualquer unidade léxica (UL) que possa estar contido em uma EI e se obtém uma lista de expressões com aquela mesma palavra.

- perspectiva onomasiológica:

- critério de busca por UL que possa constar nas paráfrases ou equivalentes não-idiomáticos intralingüísticos (quando o consulente sabe apenas um conceito possivelmente manifestado por EIs): digita-se qualquer UL restrito ao quadro das paráfrases ou se seleciona o conceito procurado na lista de conceitos, para se obter todas as EIs que se referem àquele conceito.

- critério de busca por componente dos temas contidos nas EIs (quando o consulente quer localizar todas as EIs que trazem um elemento em relação de hiponímia): digita-se qualquer UL restrito ao quadro dos temas ou se seleciona diretamente o tema procurado na lista de temas.

- classificação lexical por tema

5. Considerações finais

A utilização da web como *corpus* comprovou sua eficácia na determinação da frequência de EIs, um dos tipos de fraseologismo bastante usual, mas de menor ocorrência nas bases textuais convencionais que não priorizam a inserção de textos coloquiais, justamente a maior fonte de idiomatismos.

Levando-se em conta essa frequência, pôde-se chegar a uma nomenclatura suficientemente confiável para ser indicada como usual na contemporaneidade das línguas francesa e brasileira.

Todos os resultados dessa pesquisa estarão em breve disponíveis *on-line* no site do ATILF (www.atilf.fr).

6. Referências bibliográficas

BERBER SARDINHA, T. **Linguística de corpus**. Barueri, SP: Manole, 2004.

BLANCO, X. Dictionnaires électroniques et traduction automatiques espagnol-français. **Langages**, septembre 2001.

CARAMORI, A. P. **É o bicho**: é bestial dicionário de expressões idiomáticas no domínio dos animais com equivalências em italiano e respectivas listas temáticas. São Paulo, 2000. Dissertação (Programa de Pós-graduação em Língua e Literatura Italiana) Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo.

COLSON, J-P. Corpus linguistics and phraseological statistics: a few hypotheses and examples. In: BURGER, H., HÄCHI BUHOFER, A., GRÉCIANO, G. (eds.). **Flut von texten – vielfalt der kulturen**. Ascona 2001 zu Methodologie und kulturspezifik der phraseologie. Baltmannsweiler: Schneider Verlag Hohengehren, p. 47-59, 2003.

COLSON, J-P. The World Wide Web as a corpus for set phrases. In: BURGER, H., DOBROVOL'SKIJ, D., KÜHN, P., NORRICK, N. (eds.). **Phraseologie / Phraseology**. Berlin, New York: Mouton de Gruyter, no prelo.

- DUNCAN JR., J. C. **A frequency dictionary of Portuguese words**. 1972. Dissertation, Stanford University.
- FALCÃO, P. C. S. **A tradução para o português de expressões idiomáticas em inglês como nomes de animais**. São José do Rio Preto, 2002, 107 p. Dissertação (Mestrado em Lingüística) – IBILCE, Universidade Estadual Paulista.
- FORNICOLA, M. V. **Expressões idiomáticas da língua francesa e respectivas formas equivalentes em língua portuguesa: tratamento léxico-semântico**. São Paulo: 1999. 254 p. Dissertação (Mestrado em Semiótica e Lingüística Geral) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.
- GREFENSTETTE, G., NIOCHE, J. Estimation of english and non-english language use on the www. Proc. **RIAO 2000**, Content-Based Multimedia Information Access, pages 237-- 246, 2000.
- GROSS, M. Constructing Lexicon-grammars. In: ATKINS, B., T., S., ZAMPOLLI, A. (eds.). **Computational approaches to the lexicon**. Oxford: Oxford University Press, 1994.
- KILGARIFF, A., GREFENSTETTE, G. Web as Corpus. In: **Computational Linguistics**, Volume 29, Number 3, 2003.
- LODOVICI, F. M. M. **Elementos constitutivos dos idiomatismos no português do Brasil**. São Paulo, 1989, 262 p. Dissertação (Mestrado em Ciências Humanas) - Pontifícia Universidade Católica.
- ORTÍZ ALVAREZ, M. L. **Expressões idiomáticas do português do Brasil e do espanhol de Cuba: estudo contrastivo e implicações para o ensino do português como língua estrangeira**. 2000. Tese (Doutorado em Lingüística Aplicada: Ensino/Aprendizagem de Segunda Língua e Língua Estrangeira) - Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas.
- RIOS, T. H. C. **Idiomatismos com nomes de partes do corpo humano**. São José do Rio Preto, 2004. Dissertação (Mestrado em Lingüística) – IBILCE, Universidade Estadual Paulista.
- RIVA, H. C. **Dicionário onomasiológico de expressões idiomáticas**. São José do Rio Preto, 2004, 187 p. Dissertação (Mestrado em Lingüística) – IBILCE, Universidade Estadual Paulista.
- RONCOLATTO, E. **Estudo contrastivo das expressões idiomáticas do português e do espanhol**. Assis, 1996. Dissertação (Mestrado em Letras: Filologia e Lingüística Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista.
- RONCOLATTO, E. **Expressões idiomáticas do português do Brasil e do espanhol da Colômbia: análise, classificação e equivalências**. 2001. Tese (Doutorado em Letras: Filologia e Lingüística Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Assis.
- SALKOFF, M. On using the lexicon-grammar in a bilingual dictionary. **Lexiques-grammaires comparés et traitements automatiques**. Montréal: UQAM, 1995.
- SINCLAIR, J. **Corpus, concordance, collocation**. Oxford : Oxford University Press, 1997.
- SUPER Interessante, São Paulo, n. 201, p. 54-63, 2004.
- TAGNIN, S. E. O. **Levels of conventionality and the translators task**. São Paulo, 1987. Tese (Doutorado em Letras) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.
- TAGNIN, S. O. **Expressões idiomáticas e convencionais**. São Paulo: Ática, 1989.
- VALE, O. A. **Expressões cristalizadas do português do Brasil: uma proposta de tipologia**. Araraquara, 2000. Tese (Mestrado em Lingüística e Língua Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista.

XATARA, C. M. **A tradução para o português de expressões idiomáticas em francês.** Araraquara, 1998, 253 p. Tese (Mestrado em Lingüística e Língua Portuguesa) - Faculdade de Ciências e Letras, Universidade Estadual Paulista.

XATARA, C., OLIVEIRA, W. L. **Dicionário de provérbios, idiomatismos e palavrões francês-português / português-francês.** São Paulo: Cultura, 2002. 363p.