Pandaemonium germanicum 9/2005, 393-416

O trabalho de tradutor como fonte para constituição de base de dados

Bibiana Teixeira de Almeida*

Abstract: The purpose of this article is to report on the work carried out during the research project "O trabalho de tradutor como fonte para a constituição de base de dados" (The translator´s work as a source for the constitution of a database). Through the restoration, organization and digitalization of the personal glossary and part of the books containing the translations made by the deceased public translator Gustavo Lohnefink, this research project intends to construct a digital database of German – Portuguese technical terms (for the language pair), which could then be used by other translators. In order to achieve this purpose, a specific methodology had to be developed, which could be used as a starting-point for the treatment and recovery of other similarly organized data-collections.

Keywords: Translation; Terminology; Digitalization of Database.

Zusammenfassung: Ziel des vorliegenden Beitrags ist es, über die im Laufe des Forschungsprojekts "O trabalho de tradutor como fonte para constituição de base de dados" (Die Arbeit des Übersetzers als Quelle für die Erstellung

* Bibiana T. de Almeida (E-mail: bibiana.almeida@gmail.com) é aluna do curso de pós-graduação em tradução do CITRAT/FFLCH/USP e desenvolveu o projeto "O trabalho de tradutor como fonte para a constituição de base de dados", como bolsista da FAPESP, sob a orientação do Prof. Dr. João Azenha Jr.

von Datenbanken) durchgeführte Arbeit zu berichten. Neben der Aufarbeitung, Digitalisierung und Verarbeitung des persönlichen Glossars des verstorbenen beeidigten Übersetzers Gustavo Lohnefink und eines Teils der von ihm übersetzten Texte hatte das erwähnte Forschungsprojekt das Hauptziel, eine elektronische, zweisprachige Datenbank (Deutsch – Portugiesisch) von Fachwörtern zu erstellen. Diese kann dann anderen ÜbersetzerInnen zur Verfügung gestellt werden. Zur Erstellung der Datenbank wurde eine spezifische Methodologie entwickelt, die als Ausgangspunkt für die Ver- und Aufarbeitung von ähnlich organisierten Datenbeständen benutzt werden kann.

Stichwörter: Übersetzung; Terminologie; Digitalisierung von Datenbanken.

Resumo: O presente artigo visa a relatar o trabalho desenvolvido durante o projeto de iniciação científica "O trabalho de tradutor como fonte para a constituição de base de dados". Através da documentação, recuperação e digitalização do glossário pessoal e de parte do acervo de livros de traduções do falecido tradutor público Gustavo Lohnefink, o referido projeto de pesquisa teve por objetivo principal constituir uma base de dados digital de termos técnicos para o par de línguas alemão – português e colocá-lo à disposição de tradutores(as). Para tanto, foi necessário o desenvolvimento de uma metodologia específica, aqui relatada, que poderá servir como ponto de partida para o tratamento e a recuperação de acervos semelhantes.

Palavras chave: Tradução; terminologia; digitalização de base de dados.

1. Introdução

No Brasil, todas as traduções de documentos oficiais de/para órgãos públicos devem ser feitas por tradutores juramentados, aprovados, reconhecidos e credenciados pelo governo. Considerados documentos oficiais, os textos produzidos por esses tradutores são depositados nas Juntas Comerciais de cada estado brasileiro.

A Junta Comercial do Estado de São Paulo (JUCESP) é a guardiã dos textos produzidos pelos tradutores juramentados desse estado, não apenas dos que ainda estão na ativa, mas também dos tradutores já aposentados ou falecidos. Isso faz do acervo da JUCESP uma valiosa coleção, pois dele

constam traduções produzidas desde o século passado e que, portanto, documentam uma diversidade de informações não apenas de cunho lingüístico, mas também de vários aspectos da história do estado de São Paulo (o desenvolvimento econômico e técnico-tecnológico, a imigração, etc).

Infelizmente, como não há uma sistemática de recuperação desse acervo e os textos guardados pela JUCESP não sofrem processo de conservação ou restauração, a sua utilização não apenas como fonte para consulta, mas também como fonte de pesquisa, torna-se mais difícil com o passar do tempo, pois o meio físico (papel, tintas) de registro sofre constante deterioração, o que inviabiliza cada vez mais o seu aproveitamento.

2. O projeto de iniciação científica como ponto de partida

O projeto de iniciação científica "O trabalho de tradutor como fonte para a constituição de base de dados" propôs-se a formar um banco de dados digitais de termos técnicos para o par de línguas alemão-português por meio da documentação, digitalização e recontextualização dos fichários que compõem o glossário elaborado pelo tradutor público Gustavo Lohnefink[1]. O principal objetivo deste trabalho foi a recuperação do fichário trilíngüe (alemão – português – inglês) elaborado pelo tradutor, bem como dos livros depositados por ele e, após seu falecimento, por sua família junto à JUCESP (Junta Comercial do Estado de São Paulo).

As primeiras tentativas de organização do acervo apontaram lacunas de informações que, do ponto de vista científico, não conferiam idoneidade aos dados. Além disso, uma primeira avaliação mostrou que os registros teriam que passar por um processo de triagem e complementação para que pudessem ser reaproveitados como base de dados para a constituição de glossários bilíngües ou para sua disponibilização para outros usuários. Consideradas as especificidades do *corpus*, sobretudo no que respeita ao seu

---

[1] Embora fosse interessante fornecer neste artigo algumas informações biográficas sobre o tradutor cujo trabalho é objeto desta pesquisa, infelizmente não se teve, desde o início do projeto, acesso a essas informações, de forma que elas não serão aqui disponibilizadas.

vínculo com uma instituição pública, fez-se necessário o desenvolvimento de uma metodologia específica que, ao mesmo tempo, desse conta de tais especificidades e pudesse ser replicada na recuperação de outros acervos organizados de forma similar.

O trabalho foi realizado em três etapas, descritas na Tabela 1 a seguir:

Tabela 1: Plano geral da pesquisa

	Fases	**Duração prevista**
1ª Etapa: Recuperação e digitalização do fichário	1) Descrição	2 meses
	2) Transposição para o ambiente MS Access®	
	3) Classificação temática dos registros	
2ª Etapa: Recuperação, digitalização e preparação dos textos	1) Descrição	6 meses
	2) Digitalização: escaneamento e aplicação de *software* de reconhecimento de caracteres	
	3) Preparação dos textos para amostragem	
3ª Etapa: Cruzamento dos dados das etapas anteriores	1) Projeção do fichário sobre os textos	4 meses
	2) Recontextualização dos termos	

3. As etapas da pesquisa em detalhe

3.1 Primeira etapa: Recuperação e digitalização do fichário

A primeira etapa da pesquisa envolveu os trabalhos de descrição e transposição do acervo de fichas de Gustavo Lohnefink para o ambiente MS Access®.

O fichário do tradutor em questão, tal como doado à Universidade de São Paulo pela ATPIESP (Associação dos Tradutores Públicos e Intérpretes Comerciais do Estado de São Paulo), é uma estrutura fechada, dotada de uma organização interna própria, que possui um início e um fim. De modo geral, cada ficha contém um termo em português com sua respectiva tradução para o inglês e/ou alemão. As fichas, organizadas alfabeticamente, da letra A até a letra Z, estão armazenadas em três arquivos de metal (aqui chamados de AM1, AM2 e AM3), quatro caixas de madeira (CM1, CM2, CM3 e CM4) e duas caixas de papel (CP1 e CP2).

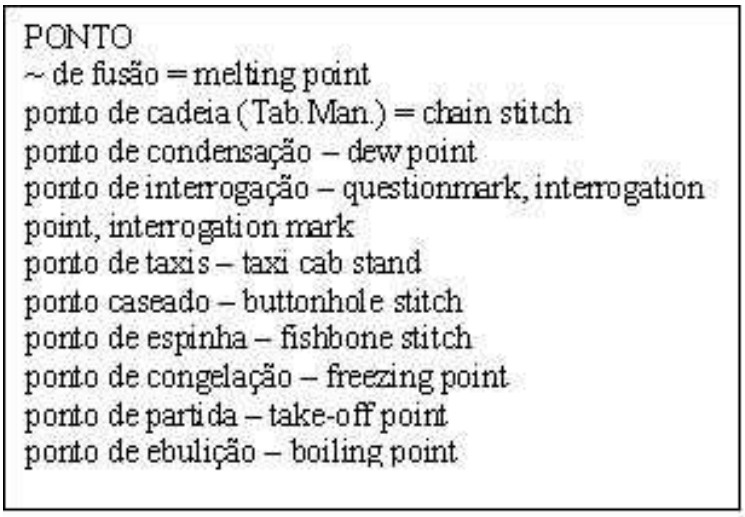
A Tabela 2 fornece um panorama sucinto da organização do acervo de fichas.

Tabela 2: Informações gerais sobre o acervo de fichas

Fichário	*AM1*	*AM2*	*AM3*	*CM1*	*CM2*	*CM3*	*CM4*	*CP1*	*CP2*	Total
Registros	2714	3136	2809	1486	1226	1359	1101	650	274	14751
Letras	A-B	C	D-I	J-N	O-P	P-R	S-T	T-Z	Q-X	

A estrutura central, segundo a qual o acervo de fichas está organizado, recebeu o nome de fichário principal. Nele estão registrados termos em português com suas respectivas traduções para o inglês e/ou para o alemão. Entretanto, no verso de parte das fichas que compõem esse fichário há outros registros, alguns dos quais seguem o mesmo modelo do fichário principal e outros que compõem subgrupos de informações distintas: fichas com referências a outros registros (entradas) dentro do fichário; fichas manuscritas com nomes de plantas e animais, com nomenclatura científica e traduções para um ou mais dos três idiomas mencionados anteriormente; fichas datilografadas com nomes de plantas e animais, com nomenclatura científica e referências bibliográficas; fichas de catalogação de peças musicais, discos e personalidades; fichas com um termo em inglês e sua tradução para o português (exclusivamente no fichário CP2) e fichas com vários termos arrolados sob o escopo de uma palavra-chave, normalmente vários substantivos compostos acompanhados de suas respectivas traduções.

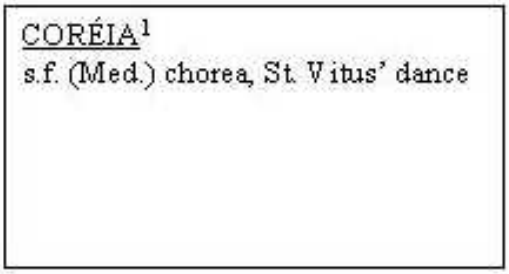
Figura 1: Ficha com substantivos compostos derivados de um mesmo substantivo


PONTO
~ de fusão = melting point
ponto de cadeia (Tab.Man.) = chain stitch
ponto de condensação – dew point
ponto de interrogação – questionmark, interrogation point, interrogation mark
ponto de taxis – taxi cab stand
ponto caseado – buttonhole stitch
ponto de espinha – fishbone stitch
ponto de congelação – freezing point
ponto de partida – take-off point
ponto de ebulição – boiling point

A sistematização do fichário exigiu que fosse deixada de lado grande parte dessas informações que constituíam, por assim dizer, outros fichários dentro do fichário principal. A decisão de se excluírem do *corpus* da pesquisa esses subgrupos pautou-se pelo propósito de o trabalho se concentrar num subgrupo de termos técnicos em alemão. Por essa razão, dos registros encontrados no verso das fichas, foram digitalizados apenas aqueles que seguiam o mesmo modelo dos registros-padrão que constituíam o fichário principal.

No registro ou ficha-padrão constam o termo em português e suas respectivas traduções para o inglês e o alemão, sempre nesta ordem. Em sua forma mais simples, esse tipo de registro contém apenas o termo em português e a sua tradução (ou traduções) para um ou ambos os idiomas. Entretanto, o fichário também contém uma série de fichas enriquecidas por observações do tradutor. Tais anotações fornecem informações sobre a morfologia da palavra, a área terminológica da qual ela faz parte, a acepção utilizada, o registro de linguagem no qual se enquadra etc[2]. Veja-se o exemplo a seguir:

Figura 2: Exemplo de registro enriquecido por anotações



[2] Uma variação particularmente interessante da ficha-padrão traz, acompanhando a tradução dos termos, definições teóricas dos mesmos. Embora tenham sido encontradas sempre nos registros dos versos das fichas e estejam todas em inglês, essas informações foram integralmente digitalizadas, pois poderão ser úteis para outras pesquisas que utilizem os registros do tradutor para o par de línguas português-inglês. Vale ressaltar aqui que a quantidade de livros e de termos do fichário cujo idioma é o inglês supera com certa vantagem o correspondente em alemão, indicando que aquele foi o idioma mais produtivo para o tradutor. Assim, existe nos dados não utilizados uma grande quantidade de informações disponíveis para pesquisas que contemplem o par de línguas inglês-português.

A inscrição *s.f.* classifica morfologicamente o termo em português como substantivo feminino. A inscrição *(Med.)* classifica a área terminológica à qual o termo pertence (neste caso, a Medicina). O número *1* que acompanha o termo em português indica a existência de outras acepções diferentes daquela registrada na ficha para a palavra, mesmo que elas não estejam registradas no fichário do tradutor. Essa classificação numérica indica que a palavra foi consultada em uma obra de referência (provavelmente um dicionário) e que o tradutor selecionou o significado mais adequado para o contexto com o qual estava trabalhando, deixando de lado outras possíveis acepções.

Concluída a descrição do fichário, procedeu-se à sua transposição para o ambiente MS Access®, onde foram criados três diferentes arquivos de banco de dados: Fichas-GLohnefink.mdb, Verso-Fichas-GLohn.mdb e Fichas-CP2.mdb. O primeiro arquivo acolheu os registros que compunham o fichário principal; no segundo arquivo foram inseridos os registros que seguiam o mesmo padrão do fichário principal, mas que estavam anotados no verso das fichas do mesmo; o terceiro arquivo, por sua vez, acolheu o glossário contido na caixa de papelão CP2.

Foram contabilizadas 14.475 entradas (excluídos os 274 registros contidos na caixa CP2). A Tabela 3 apresenta os números de fichas digitalizadas separados de acordo com os idiomas contidos nos registros.

Tabela 3: Número de registros contidos no banco de dados do MS Access ®, por idioma.

Idioma	*Português - Alemão*	*Português - Inglês*	*Português – Inglês + Alemão*	***Total***
Registros	4.253	7.842	2.380	**14.475**

Para a constituição do *corpus* a ser utilizado na pesquisa, foram excluídos os registros exclusivos para o par de línguas português – inglês, restando, portanto, um total de 6.633 registros (4.253 registros do par português – alemão somados aos 2.380 registros trilíngües). Após a eliminação das entradas duplicadas, o número total de registros passou a ser 6.534, o *corpus* efetivamente utilizado na realização deste trabalho de pesquisa.

O processo de transposição das fichas para o ambiente eletrônico foi um trabalho bastante técnico que, a despeito do dispêndio de tempo, não apresentou em si dificuldades do ponto de vista metodológico. Para a digitalização dos registros, consideradas sempre as características específicas do fichário em questão, procedeu-se a uma adaptação mais simplificada do modelo de ficha terminológica proposto por AUBERT (1996). A adaptação inicial das fichas traduziu-se na reestruturação dos campos, a fim de que fossem abrigadas com mais facilidade as informações e anotações constantes dos diferentes tipos de ficha encontrados. Foi, portanto, durante o trabalho de digitalização que se definiram as adaptações necessárias para adequar o modelo de ficha desenvolvido por AUBERT (1996) às características do acervo de maneira a preservar ao máximo sua estrutura, sem descaracterizá-lo. O esforço constante pela simplificação, sem perda do rigor, conduziu à atual configuração dos campos da ficha terminológica utilizada na digitalização do fichário de G. Lohnefink, exemplificada na Figura 3.

Figura 3: Registro do acervo de G. Lohnefink no ambiente MS Access®

01-LP	02-Ocorrência-LP:	
pt	pele	
03-Termo-LP:		
pele		
04-Padronizado	05-Fonte	06-Ano
	CM2	
07-Uso (Pt p/ NC)	08-Morfologia	09-Sintaxe
10-Contexto		
"Legendas "R": 11 facilmente inflamável 23/24/25 venenoso ao ser aspirado, deglutido ou em contato com a pele. "(P-L93P125b)		
11-Sinônimos (Refs Fich)	12-Unitermos (Outrs Refs)	13-LC1
		en
14-Termo-LC1:EN		
skin		
15-LC2		
al		
16-Termo-LC2:AL		
Haut; Fell		
17-Equivalência (NC p/ pt)	18-Área	19-Sub-área
	Anatomia	
20-Tema	21-Data	22-Documentador
		G.Lohnefink
23-Revisor		
24-Definição		

Parte dos campos da ficha foi preenchida com informações provenientes dos textos do tradutor, após o que chamamos de "a projeção do fichário" sobre exemplares extraídos dos livros de textos traduzidos, tal como descrita adiante. É o caso dos campos "Ocorrência-LP" e "Contexto", para os quais foram transpostos os trechos extraídos dos textos presentes nos li-

vros, e também o campo "Área", cujas informações sobre classificação terminológica foram preenchidas com informações relativas ao contexto encontrado.

Este último procedimento – a classificação terminológica – foi realizado após a preparação e a constituição de uma amostragem do banco de textos digitalizados do tradutor. A partir da projeção do fichário sobre um determinado universo dos textos, foram encontrados e selecionados os termos que faziam parte da esfera terminológica correspondente e, só então, foi atribuída a eles uma classificação terminológica, dentro de um contexto estabelecido.

A classificação temática dos registros tomou por base domínios terminológicos mais abrangentes. A grande maioria – cerca de 53% – dos termos recontextualizados foi classificada como "Geral", pois não pertencia a nenhum domínio terminológico mais específico. O segundo maior índice de ocorrências foi de termos do domínio "Jurídico", que constituiu cerca de 18% do total recontextualizado. Os outros registros encontrados pertenciam a diversas áreas das Ciências Humanas, Médicas, Biológicas, Exatas etc., mas todos apresentaram ocorrências inferiores a 7,5%, sendo que a menor porcentagem encontrada foi 0,16% (casos que apresentaram apenas uma ocorrência).

3.2 Segunda etapa: Recuperação, digitalização e preparação dos textos contidos nos livros depositados junto à JUCESP

À conclusão do processo de digitalização dos registros contidos no fichário, seguiu-se a realização de um inventário do acervo de livros de Gustavo Lohnefink.

Atualmente, os livros do tradutor, bem como os livros de outros tradutores públicos falecidos ou aposentados, encontram-se na biblioteca da JUCESP. Há algum tempo, todos os livros que faziam parte do acervo foram provisoriamente acondicionados em uma sala que infelizmente não é o ambiente adequado para o armazenamento desse tipo de material, deixando-o exposto a agentes de vários tipos e que contribuíram para a deterioração do acervo.

Embora sujeitos ao mesmo tipo de exposição que outros livros menos afortunados, os volumes produzidos por Gustavo Lohnefink apresen-

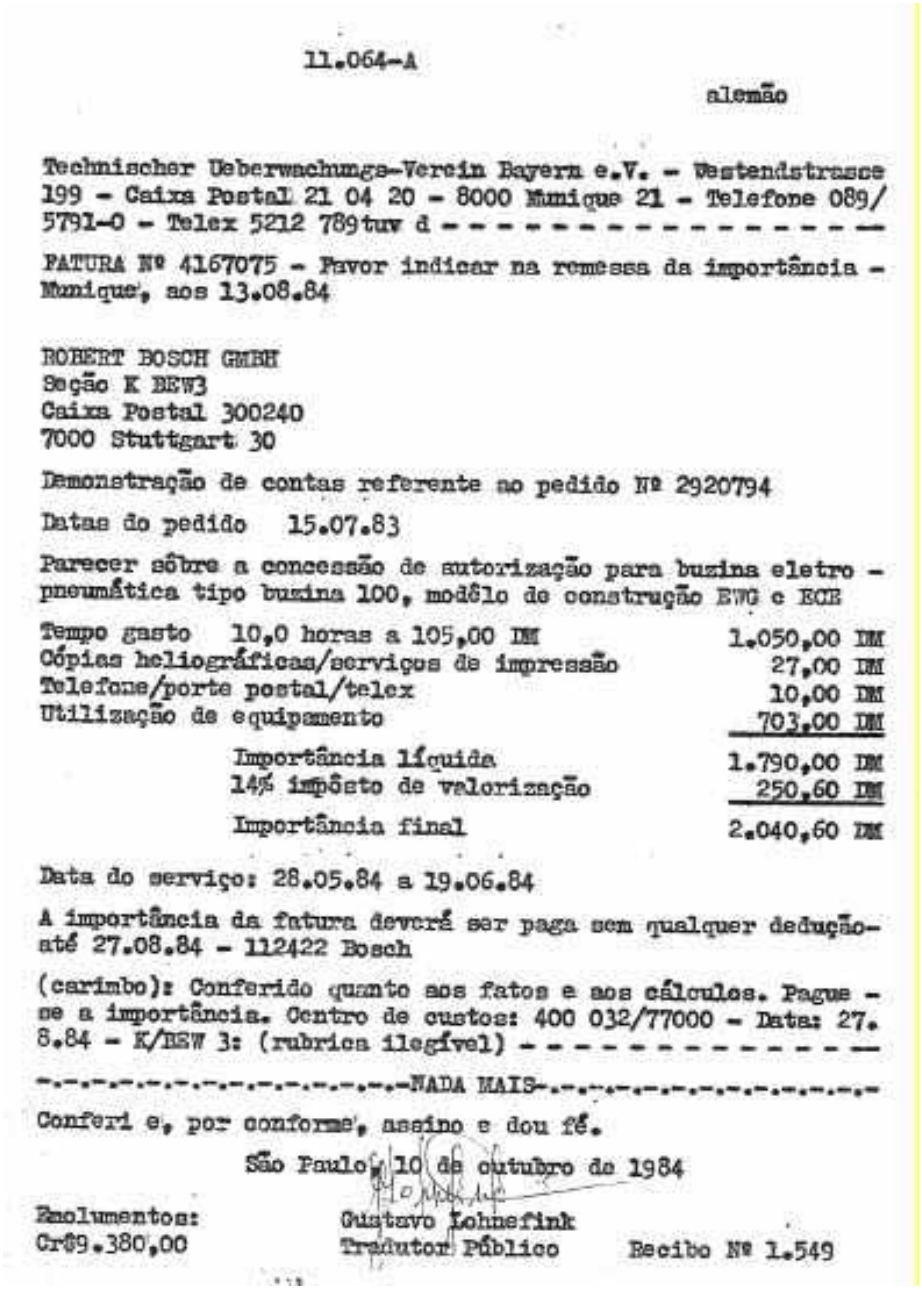
tam-se razoavelmente bem conservados. A umidade do ambiente promoveu um leve emboloramento das capas e beiradas das páginas dos livros, mas não houve grande comprometimento da integridade do papel nem da impressão dos textos.

Foi elaborada uma descrição detalhada do estado de conservação dos 247 livros do tradutor encontrados na biblioteca da JUCESP, do meio de registro das traduções – impressão, fotocópia, cópia de carbono ou mimeografada –, e das proporções entre os idiomas envolvidos (português, alemão e inglês).

Desse total de livros, 49 volumes estavam encadernados com 500 páginas, 196 volumes com 400 páginas, um volume com 270 páginas e um outro com apenas 184 páginas, somando 103.354 páginas de tradução registradas. No que diz respeito à distribuição por idiomas, os livros são sempre encadernados com textos em português e em uma única outra língua estrangeira, de forma que há 85 livros para o par de idiomas alemão – português e 162 livros para o par de idiomas inglês – português. O livro mais antigo do tradutor data de 1962 e o mais recente de 1993. São, portanto, mais de 30 anos de trabalho, dos quais a fase mais produtiva foi a da década de 80, com 171 livros encadernados. Em seguida vem a década de 70, com 53 livros, e as décadas de 90 e 60, com 12 e 11 livros, respectivamente.

As formas de registro dos textos são relativamente condizentes com a época em que os livros foram encadernados. Entre os volumes do tradutor encontram-se exemplares com textos copiados pelo processo de mimeografia, amplamente utilizado para fazer cópias dos textos do tradutor, particularmente dos mais antigos, encadernados até o final da década de 70.

Figura 4: Texto reproduzido por carbono



11.064-A

alemão

Technischer Ueberwachungs-Verein Bayern e.V. - Westendstrasse 199 - Caixa Postal 21 04 20 - 8000 Munique 21 - Telefone 089/ 5791-0 - Telex 5212 789tuv d - - - - - - - - - - - - - - - - - -

FATURA Nº 4167075 - Favor indicar na remessa da importância - Munique, aos 13.08.84

ROBERT BOSCH GMBH
Seção K BEW3
Caixa Postal 300240
7000 Stuttgart 30

Demonstração de contas referente ao pedido Nº 2920794

Datas do pedido 15.07.83

Parecer sôbre a concessão de autorização para buzina eletro - pneumática tipo buzina 100, modêlo de construção EWG e ECE

Tempo gasto 10,0 horas a 105,00 DM	1.050,00 DM
Cópias heliográficas/serviços de impressão	27,00 DM
Telefone/porte postal/telex	10,00 DM
Utilização de equipamento	703,00 DM
Importância líquida	1.790,00 DM
14% impôsto de valorização	250,60 DM
Importância final	2.040,60 DM

Data do serviço: 28.05.84 a 19.06.84

A importância da fatura deverá ser paga sem qualquer dedução- até 27.08.84 - 112422 Bosch

(carimbo): Conferido quanto aos fatos e aos cálculos. Pague - se a importância. Centro de custos: 400 032/77000 - Data: 27. 8.84 - K/BEW 3: (rubrica ilegível) - - - - - - - - - - - - - -

-.-.-.-.-.-.-.-.-.-.-.-.-.-.-NADA MAIS-.-.-.-.-.-.-.-.-.-.-.-.-.-

Conferi e, por conforme, assino e dou fé.

São Paulo, 10 de outubro de 1984

Emolumentos:
Cr$9.380,00

Gustavo Lohnefink
Tradutor Público

Recibo Nº 1.549

Figura 5: Texto fotocopiado

Foram escolhidos para a constituição do *corpus* livros com textos mais legíveis e bem conservados. Além de serem mais recentes, as cópias desses textos foram feitas utilizando carbono, material mais resistente e que se conserva melhor com o passar do tempo. A Tabela 4 abaixo fornece um breve panorama dos meios de registro dos textos e do número de livros nos quais são encontrados.

Tabela 4: Meios de registro dos textos nos livros

Meio de registro	Quantidade de livros
Mimeografia	55
Carbono	146
Carbono e fotocópia	29
Fotocópia	8
Impressão	8

O levantamento inicial sobre o estado de conservação dos livros, feito durante o inventário dos exemplares da biblioteca, forneceu os seguintes dados sobre a qualidade gráfica dos textos encadernados:[3]

Tabela 5: Estado de conservação dos textos

Estado de conservação/legibilidade do texto	Quantidade de livros
Ruim (apagado)	31
Médio (borrado, mas legível)	138
Bom (levemente borrado)	62
Muito bom (suficientemente legível para escaneamento)	7
Sem classificação[1]	9

Infelizmente essa etapa dos trabalhos teve seu cronograma reformulado por conta de alguns imprevistos e dificuldades de ordem burocrática na negociação com a entidade para a obtenção dos textos. O atraso causado pelos trâmites desse processo ocasionou também uma redução emergencial no *corpus* de textos a serem digitalizados, a fim de que fosse possível realizar, dentro do período de vigência da bolsa de iniciação científica, todas as etapas do projeto com suas respectivas fases e a conclusão do trabalho de projeção do *corpus* de fichas sobre uma parte, ainda que pequena, dos livros – condição essencial para testar a metodologia desenvolvida durante o projeto de pesquisa.

O acesso aos textos dos tradutores deve, portanto, constituir uma preocupação fundamental para futuras pesquisas. Os livros com os textos não são de propriedade dos tradutores; são instrumentos públicos e, como tais, ficam sob a guarda da JUCESP. Sua liberação será, talvez, o ponto mais delicado a ser

---

[3] Volumes aos quais não foi possível o acesso durante o processo de inventariação dos livros na biblioteca da JUCESP.

negociado diretamente com essa Instituição em pesquisas futuras. Primeiramente, porque não existe na rotina da JUCESP a prática de negociação com pesquisadores. Em segundo lugar, para a preservação do sigilo acerca do conteúdo desses textos, torna-se necessária a negociação de um termo de confiabilidade para que esse material não venha a ser divulgado indevidamente, fora do meio acadêmico. A experiência adquirida na execução deste projeto de pesquisa demonstrou que a negociação e obtenção do material devem constituir etapas prévias ao início dos trabalhos de pesquisa, pois são condição essencial para a realização dos mesmos. Em suma: o vínculo com instituições externas à Universidade, com uma rotina distinta e toda ela regulada por estatutos e regimentos, é uma variável que permeia todo o trabalho de uma pesquisa dessa natureza. A dependência da anuência da JUCESP estende-se, inclusive, para o momento em que se vão tornar públicos os resultados, depois de concluída a pesquisa.

Foi obtida autorização da JUCESP para que fossem feitas fotocópias de textos extraídos de cinco dos 247 livros do tradutor Gustavo Lohnefink que se encontram armazenados junto àquela Instituição. Por uma questão de tempo, a escolha dos textos foi pautada por dois critérios principais:

1. a legibilidade do texto, diretamente relacionada à qualidade da impressão e da cópia, questão fundamental para que o *software* pudesse reconhecer os caracteres adequadamente;
2. a necessidade de elaboração de dois *corpora* de textos, um em alemão e um em português.

A escolha do material a ser fotocopiado para posterior digitalização obedeceu, portanto, basicamente ao critério da qualidade gráfica do texto. O primeiro esforço no sentido de tentar fazer uma amostragem que se pautasse em uma regularidade para a constituição de um *corpus* estatisticamente válido mostrou-se inviável, assim como a realização de uma amostragem dos textos por tema ou domínio específico, uma vez que os livros do tradutor contêm textos produzidos num determinado espaço de tempo, de forma que textos de diferentes assuntos são encadernados num mesmo volume, sem nenhum encadeamento temático entre si. Um trabalho que tentasse proceder a uma seleção temática dos textos teria que recorrer a todos os exemplares dos livros, muitos dos quais eram inadequados – ilegíveis demais – para passar por um processo de digitalização e reconhecimento de caracteres.

Do universo de 247 livros foram isolados para essa etapa da pesquisa apenas aqueles cujos textos pertenciam ao par de línguas alemão-português, reduzindo o *corpus* em potencial a 85 livros. Desses exemplares foram selecionados cinco volumes, que melhor atenderam às exigências de qualidade para a digitalização.

Foram fotocopiadas 195 páginas, 136 com textos em português e 59 com textos em alemão. A amostragem de textos foi aleatória no sentido de que não foram predeterminados os tipos de texto que seriam fotocopiados ou o idioma (alemão ou português) dos mesmos. Novamente utilizou-se o critério de qualidade da impressão, razão pela qual em alguns dos livros apenas algumas páginas foram fotocopiadas. Cumpre ressaltar que em um mesmo livro, inclusive nos volumes mais recentes, foram encadernados textos com diferentes meios de registro, dos quais os impressos e os fotocopiados de maior legibilidade eram, muitas vezes, raros.

Para a etapa subseqüente, o reconhecimento ótico de caracteres (OCR – Optical Character Recognition), foram testados *softwares* especializados, com o intuito de detectar qual atenderia melhor às necessidades prementes deste projeto de pesquisa:

1. reconhecimento de caracteres em ambos os idiomas, alemão e português;
2. rapidez e acuidade na digitalização dos textos;
3. capacidade de conversão do texto para um arquivo de formato amplamente aceitável (.txt, .doc, .pdf, ou semelhantes).
4. relação custo – benefício do *software* viável para um projeto deste porte.

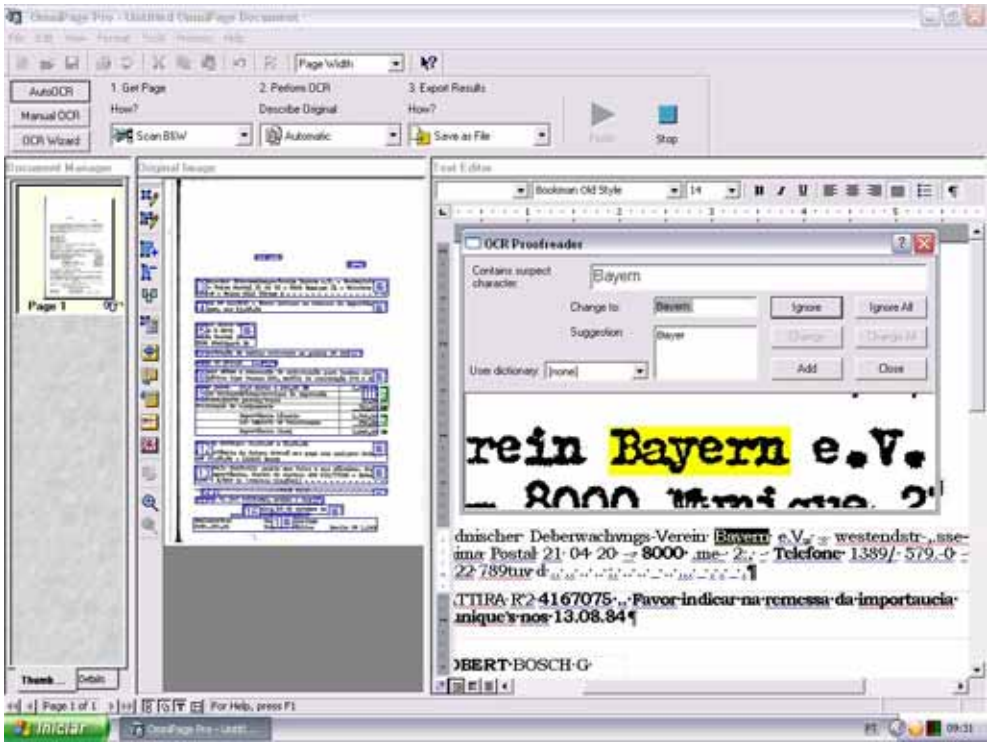
Todos os quatro *softwares*[4] testados apresentavam características gerais bastante semelhantes, executando basicamente as mesmas operações. O *software* que apresentou melhor desempenho e se mostrou mais adequado para os trabalhos de digitalização deste projeto de pesquisa foi o OmniPage Pro®, versão 11, produzido pela ScanSoft, Inc[5].

---

4 ABBYY Fine Reader®7 Professional Edition, Cuneiform ProÒ OCR 6, ReadIris Pro® 9 Demo e OmniPage Pro® 11

5 http://www.scansoft.com/omnipage/

Figura 6: Tela do OmniPage Pro durante o reconhecimento de caracteres



O processo de digitalização dos textos por reconhecimento ótico de caracteres desenvolveu-se nas seguintes etapas.

1. Escaneamento dos textos: operação realizada pelo próprio *software* de OCR, que escaneava o texto, reconhecia os caracteres, solicitando a confirmação do usuário para os casos mais duvidosos, e permitia que o texto obtido fosse salvo com a extensão ".doc".

Os arquivos, devidamente identificados (codificados) de acordo com o idioma, número do livro e número da página em que estavam localizados, foram divididos em dois grupos distintos, um para o idioma alemão, outro para o português. Por exemplo, o arquivo ".doc" de um texto em português localizado no livro nº 92, à frente da folha nº 315, foi identificado como P-L92P315-A, enquanto que o texto contido no verso da mesma folha recebeu a identificação P-L92P315-B e, em seguida,

ambos os arquivos foram integrados ao banco de textos digitais em português.

2. Revisão dos textos escaneados e reconhecidos, já no formato ".doc": procedimento realizado imediatamente após o reconhecimento do texto pelo *software* OmniPage Pro®, e que consistiu basicamente em uma leitura rápida do texto para corrigir erros mais evidentes de reconhecimento e para verificar e rearranjar a formatação do arquivo ".doc".

Durante a verificação da acuidade do processo de reconhecimento, foi possível constatar que os textos que haviam sido datilografados com máquina de escrever elétrica apresentaram o menor índice de erros de reconhecimento, seguidos pelos textos impressos (impressora matricial). Ambos apresentavam caracteres com acabamento de boa definição, o que possibilitava ao *software* de OCR uma identificação muito mais acurada e rápida.

Os principais problemas encontrados neste processo foram erros de colocação de vírgulas, acentuação e crase, em sua grande maioria agravados pelo próprio processo de reconhecimento, pois o *software* tende a reconhecer muitos dos traços presentes na folha, impurezas do papel por exemplo, como caracteres (normalmente vírgulas e acentos). Além disso, foram encontrados erros de grafia, concordância (número e gênero, por exemplo), colocação pronominal ou partícula ("se"), e erros cometidos pelo próprio tradutor quando da redação e/ou datilografia dos textos. Parte desses erros foi corrigida, particularmente os erros de ortografia que poderiam prejudicar o processo de localização dos termos na etapa subseqüente, quando do cruzamento do fichário com o banco de textos. Contudo, tentou-se preservar ao máximo a grafia original a fim de conservar os textos inalterados para poderem ser disponibilizados como base de dados para outros tipos de pesquisa, que possam utilizá-los com outras abordagens.

A opção pelo formato ".doc" para o armazenamento dos arquivos reconhecidos também teve a intenção de simplificar o cruzamento dos dados. Embora os programas de reconhecimento de caracteres ofereçam a possibilidade de armazenamento em vários tipos de arquivo de texto – .txt, .doc, .rtf, .xls, .csv, .wpd etc –, optou-se pela utilização do formato padrão do MS Word®, simplificando-se assim não apenas o acesso aos textos, mas também a sua manipulação na etapa ulterior da pesquisa.

3.3 Terceira etapa: Cruzamento dos dados obtidos

A recontextualização de parte dos termos do fichário foi feita através da projeção do acervo digitalizado de fichas sobre os textos, também digitalizados, através de um processo simples de busca e complementação (inserção manual de dados no formulário), realizado por meio dos *softwares* MS Word® e MS Access®.

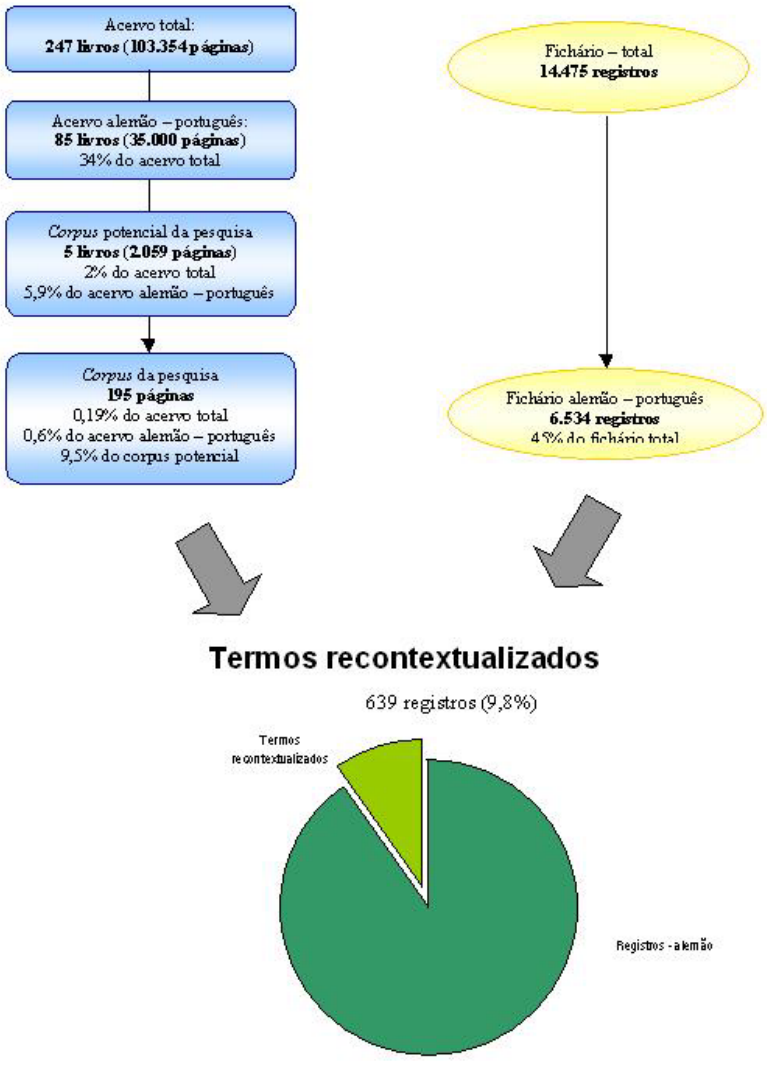
Partindo diretamente dos 6.534 registros no MS Access® para os quais o campo para os termos em alemão estava preenchido – o *corpus* de registros utilizado deste momento em diante na pesquisa – iniciou-se a busca pelos termos no arquivo unificado dos textos em português no MS Word®, utilizando-se o recurso "Localizar" desse aplicativo.

O cruzamento dos dados foi feito apenas com buscas nos textos em português, pois o fichário de Gustavo Lohnefink está organizado essencialmente por esses termos e, portanto, cada um dos registros contém uma (e apenas uma) entrada em português com o seu correspondente ou correspondentes em alemão (e/ou inglês). O caminho inverso – com busca pelo termo em alemão – teria sido inviável, pois na maioria das vezes há vários termos em alemão para um único termo em português, o que demandaria uma quantidade muito maior e praticamente imprevisível (não há como fazer um levantamento do número total desses termos) de tempo para que fossem encontrados os contextos de cada um. Além disso, o armazenamento de tal quantidade de informações dificultaria sobremaneira o seu registro no formulário do MS Access®.

Foram recontextualizados 639 dos 6.534 termos constantes do fichário português-alemão de Gustavo Lohnefink. Esse montante equivale a 9,8% – praticamente 10% – desse total de termos. Trata-se de um número extremamente significativo, pois resulta da projeção de 45% dos registros do fichário (6.534 registros nos idiomas português-alemão contra um total geral de 14.475 registros) sobre apenas 0,6% (195 páginas) do total de textos para o par de línguas português-alemão (35.000 páginas, 34% do acervo total dos 247 livros). Esse resultado, além de evidenciar que a metodologia é pertinente e permite a recuperação de dados preciosos que poderiam ser perdidos, leva a crer que a projeção do fichário sobre um universo maior dessas 35.000 páginas de textos possibilitaria a recontextualização de praticamente todos os registros, caso essa proporção se mantenha. Os números

resultantes confirmam, portanto, a validade deste trabalho de pesquisa e apontam auspiciosamente para o futuro, mostrando que é válido efetuar o cruzamento das fichas e textos digitalizados e evidenciando, conseqüentemente, que o trabalho desenvolvido neste projeto é uma iniciativa econômica, teórica e tecnologicamente viável e que pode vir a ter desdobramentos extremamente proveitosos.

Figura 7: Diagrama gráfico ilustrando os resultados obtidos pela pesquisa



4. Conclusões parciais e desdobramentos da pesquisa

A realização de uma análise objetiva dos trabalhos desenvolvidos durante o projeto "O trabalho de tradutor como fonte para constituição de base de dados" destaca essencialmente duas características marcantes desta pesquisa: o caráter documental e o seu pioneirismo.

Na verdade, o caráter pioneiro deste trabalho na esfera das relações entre os Estudos da Tradução na USP e o mundo profissional da tradução, através da recuperação de dados registrados assistematicamente e – com o apoio de uma metodologia – a disponibilização dos mesmos, pressupõe uma preocupação com a documentação das etapas, sucessos, insucessos e resultados, que garanta sua aplicação *mutatis mutandis* para outras pesquisas semelhantes.

Ao explorar pela primeira vez a aproximação entre instituições, especificamente entre a Universidade de São Paulo e a Junta Comercial do Estado de São Paulo (JUCESP), esta pesquisa incorporou como elementos condicionantes todos os passos e percalços que marcaram sua execução e concentrou seus esforços no desenvolvimento e no teste de uma metodologia a ser aprimorada em pesquisas futuras. Nesse sentido, alguns procedimentos tiveram de ser desenvolvidos "a partir do zero", freqüentemente por tentativa e erro, incorrendo na alternância de procedimentos na busca pela metodologia mais adequada.

Em todas as etapas de trabalho com o acervo, os resultados obtidos com as iniciativas tomadas eram o norte que estabelecia a direção a ser seguida para o desenvolvimento da metodologia utilizada e, conseqüentemente, para o recorte teórico de apoio e sua relativização. Durante todo o tempo, as características do objeto de trabalho – por exemplo, as técnicas de reprodução das traduções nos livros e as técnicas de registro (do tradutor) nas fichas – condicionaram as decisões sobre o emprego de recursos tecnológicos e sobre os procedimentos científicos a serem adotados. Tais características foram fatores determinantes para direcionar os trabalhos e permitir o desenvolvimento da metodologia estabelecida, e certamente resultarão em economia de tempo na replicação desta pesquisa em *corpora* de características similares.

No caso específico deste projeto, o acervo utilizado foi adquirido através de doação, mediada pela ATPIESP (Associação dos Tradutores Públicos e Intérpretes Comerciais do Estado de São Paulo), feita pela família do tradutor falecido ao CITRAT (Centro Interdepartamental de Tradução e Terminologia) da USP. O acesso a outros fichários preparados por tradutores públicos terá, como desta vez, de ser mediado pela Presidência da ATPIESP. Assim como nos entendimentos com a JUCESP, também na base desta pesquisa – a obtenção do *corpus* – está a mediação de uma Instituição que, neste caso específico, tem colaborado com os pesquisadores, talvez por ser esta parceria entre a Universidade e o mundo da tradução uma reivindicação antiga dos tradutores profissionais.

Além disso, a experiência deste projeto de pesquisa mostrou que o trabalho com este acervo pode ser continuado em etapas futuras de formação acadêmica. As possibilidades de trabalho são muitas, pois o material que se tem em mãos é muito rico e abre um leque de variadas opções de abordagem, que podem ser imediatamente retomadas e desenvolvidas após a conclusão deste projeto inicial.

Um desdobramento promissor é o aprimoramento da metodologia de recuperação de acervos de tradutores públicos aqui desenvolvida. O aproveitamento dos dados apresentados, associado à tendência concreta de uma simplificação do acesso ao trabalho dos tradutores, atualmente na sua quase totalidade realizados em ambiente eletrônico, permitirá a constituição de bases de dados cada vez mais ricas, a serem utilizadas para os mais diversos fins. A atualização de fontes dicionarizadas em setores do saber ainda em expansão e sem uma terminologia sedimentada é um exemplo. Criadas a partir de informações geradas por tradutores no dia-a-dia de sua profissão, essas bases seriam o reflexo da movimentação do léxico de uma especialidade em busca de uma padronização. Para tanto, a ampla divulgação, entre os tradutores profissionais, dos resultados aqui obtidos, no sentido de instruí-los quanto a um procedimento, cientificamente embasado, de registro de dados que garanta o seu reaproveitamento posterior, é de grande valia.

No caso particular da base de dados de textos traduzidos, uma outra possibilidade de aproveitamento seria a definição, ao longo do tempo, de uma tipologia de erros e inadequações em traduções feitas por profissio-

nais. O Projeto COMET[6], por exemplo, poderia acrescentar ao seu acervo os *corpora* gerados a partir de pesquisas semelhantes a esta.

Pesquisas realizadas em outras áreas que não a da Lingüística – História e Antropologia, por exemplo – também poderão usufruir desta metodologia para o reaproveitamento do material contido nos livros de traduções: os acervos dos textos gerados por tradutores documentam a história da imigração (como as certidões de nascimento, de casamento, os processos de naturalização de estrangeiros, entre outros), o intercâmbio tecnológico (contratos celebrados entre empresas estrangeiras que abriram filiais no Brasil e vice-versa, descrições de processos industriais, processos de patentes etc.), a atividade jurídica e financeira, a movimentação política flagrada na alternância entre estados de direito e regimes de exceção etc.

No caso particular dos Estudos da Tradução e da Terminologia, que constituem a preocupação mais imediata deste estudo, destacam-se duas importantes aplicações dessas bases de dados e da metodologia aqui desenvolvida para a constituição das mesmas: (1) a elaboração de glossários técnicos bilíngües, que complementem e atualizem os dicionários e obras de referência já existentes; e (2) a ampla oferta de material para estudos lingüísticos contrastivos entre pares de línguas.

Referências bibliográficas

Primária

LOHNEFINK, Gustavo. *Livro de registro de traduções*. Volumes 36521 (1978), 72, 73 (1984), 92 e 93 (1992). São Paulo: JUCESP.

Secundária

AUBERT, Francis Henrik. Introdução à metodologia da pesquisa terminológica bilíngüe. In: *Cadernos de Terminologia*, nº 02, São Paulo, Humanitas/FFLCH/USP 1996, 69-85.

---

6 Corpus Multilíngüe para Ensino e Tradução, da FFLCH/USP

AUBERT, Francis Henrik e TAGNIN, Stella E.O. *A Corpus of Sworn Translations – for linguistic and historical research*. São Paulo, Universidade de São Paulo, sem data.

BARBOSA, Maria Aparecida. Dicionário, vocabulário, glossário: concepções. *Cadernos de Terminologia*, nº 01, São Paulo, Humanitas/FFLCH/USP, 1996.

BOWKER, Lynne & PEARSON, Jennifer. *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York, Routledge 2002.

GONZALEZ, Marco e LIMA, Vera L. S. de. "Recuperação de informação e processamento da linguagem natural". Porto Alegre: PUCRS, Faculdade de Informática, sem data. http://ftp.inf.pucpcaldas.br/CDs/SBC2003/pdf/arq0026.pdf.

KRIEGER, Maria da Graça e FINATTO, Maria José Bocorny. *Introdução à Terminologia: teoria e prática*. São Paulo, Contexto, 2004.