Elaboração de um Glossário Automático Bilíngüe com Base no Modelo de Classes de Objetos

Renata Maria Odorissio
UFSCar / PPGL
Via Washington Luis, km 235
São Carlos – SP - Brazil
+55-16-3351.9316
renata.odorissio@gmail.com

Oto Araujo Vale
DL - UFSCar
Via Washington Luis, km 235
São Carlos – SP - Brazil
+55-16-3351.9316
otovale@ufscar.br

Abstract
The aim of this study is to create a bilingual electronic glossary, which is based on corpora of cooking recipes either in Portuguese and French. Gaston Gross's model of object classes is the one in which this research is based on. Using the analysis of three dictionaries, a gap was found in terms of equivalence criteria, which could make them inefficient for the dictionary user. This problem of equivalence was what motivated this study. As a result, this glossary is expected to efficiently cover a larger range of the proposed equivalence as an electronic language tool.

Resumo
Esse trabalho descreve a elaboração de um glossário automático bilíngüe, baseado análise das classes de objetos dos predicados designativos de procedimentos culinários extraídos de receitas culinárias dos corpora bilíngües em francês e português. O modelo de análise de classes de objetos de Gaston Gross tem como resultado a caracterização léxico-gramatical dos predicados associados aos argumentos, dando maior fiabilidade aos traços de equivalência dos termos bilíngües. Tal interesse surgiu a partir da análise de dicionários cujos critérios de equivalência terminológica não deixam claro ao usuário qual escolha seria a mais adequada ao termo traduzido.

Categories and Subject Descriptors
Language models
Machine translation
Text analysis

General Terms
Design

Keywords
Lexical-grammar, objects classes, bilingual glossaries, automatic dictionaries, terminological equivalence.

1. INTRODUÇÃO
Os dicionários bilíngües tradicionais ou automáticos com fins didáticos ou tradutológicos, em geral, não deixam claros os critérios de equivalência adotados, o que dificulta a escolha do melhor termo relativo à necessidade comunicativa do usuário estrangeiro.

Aqui usamos exemplos de dois dicionários bilíngües tradicionais e de um terceiro automático, instalado no site www.uol.com.br. O dicionário de papel The Pocket Oxford Hachette – French Dictionary é o único dos três citados que foi produzido com material de corpus e que traz a referência do domínio da qual advém o verbete, assumindo critérios de freqüência, inclusive, como pode ser lido em sua introdução. Entretanto, os critérios de equivalência aplicado nesse dicionário também não facilitam nem precisam a escolha a ser feita pelo usuário com relação ao objetivo comunicativo da tradução. O consulente recorre, assim, ao seu instinto lingüístico e conhecimento de mundo para fazer sua escolha do termo na língua de chegada, o que não garante adequação nem qualidade à tradução.

Tem-se, aqui, o exemplo do dicionário The Pocket Oxford Hachette que apresenta a estrutura sintática da língua de chegada, mas não incluir exemplos e se revela pouco elucidativo no que tange a questão lexical, mesmo sendo mais preciso em termos semânticos. A título de exemplificação foi escolhido o verbo *cuire*, cujos termos equivalentes são: (1) *to cook, to bake, to roast*, e complementando o verbete são propostos os seguintes equivalentes: *~ à la vapeur to steam*; *à ~ [apple] cooking*.

Diante desses dados e observações definimos, então, o objetivo do nosso trabalho, ou seja, o de construir as classes de objetos que serão a base das equivalências entre os termos da língua de partida, o francês, e a língua de chegada, o português.

Temos como modelo o estudo léxico-gramatical de Gaston Gross [2] que elabora uma caracterização criteriosa e detalhada

das classes de objetos a partir das estruturas sintáticas e semânticas da língua natural. Ele define classes de objetos como subclasses semânticas construídas a partir da subcategorização de traços e que sozinhas discriminam o sentido do operador com a precisão necessária à formação e reconhecimento de frases lingüisticamente aceitas. Todas essas classes e subclasses são arranjadas em forma de árvores semânticas que representam as relações existentes entre os elementos lexicais. Gaston Gross [3] cita como exemplo a classe dos humanos para mostrar a complexidade de codificação a partir de referências sintático-semânticas.

A priori pode parecer simples distinguir os elementos que fazem parte ou não da humanidade, porém, principalmente do ponto de vista lingüístico, Gross mostra que isso não se dá de forma tão evidente. No geral se define como humano o argumento de todo predicado estritamente humano como os verbos de apreciação ou de ordem. Outro critério é a oposição entre animados humanos e animados não-humanos, o que implica em dizer que um sujeito humano pode se associar a predicados do tipo crer, pensar, calcular. Da mesma forma, os adjetivos solteiro, ciente, desempregado, por exemplo, são qualitativos estritamente humanos. Não é preciso dizer o que há de tautológico nessa reflexão e nem tão pouco ressaltar a questão da aleatoriedade metafórica na formulação de frases que misturam predicados humanos a argumentos não-humanos.

O que interessa ao modelo de Gross é o comportamento dos argumentos e predicados de uma mesma classe. O autor mostra que a dificuldade de categorização existe, inclusive, com relação à diversidade interna desse conjunto. O autor demonstra que o traço humano não é um dado suficiente para descrever com a precisão necessária o comportamento sintático e os empregos dos predicados de uma língua natural. O que prova a necessidade de se recorrer a especificações semânticas mais precisas, fundadas sintaticamente em predicados apropriados.

Para exemplificar sua análise tem-se o verbo *nomear*, descrito como um predicado de três argumentos – sujeito e complementos – da classe de humanos. Essas informações não garantem nenhuma particularidade ou precisão ao termo. Para melhor esclarecimento tem-se a seguinte demonstração:

Luis nomeou Golveia chefe de gabinete. ⇔ Golveia nomeou Luis chefe de gabinete.

O diretor nomeou o professor chefe de departamento. ⇔ O professor nomeou o diretor chefe de departamento.

Nota-se que na primeira inversão não há problema de aceitação da frase quando os argumentos humanos são invertidos, pois se trata de nomes próprios sem traço implícito de hierarquia. Já na segunda frase ocorre uma alteração de ordem semântica: mesmo que gramaticalmente correta, a inversão dos argumentos produz uma sentença semanticamente inadequada. Ou seja, os argumentos e predicados carregam restrições específicas suplementares, de natureza léxico-gramatical, o que não pode ser desconsiderado quando o objetivo é a definição das equivalências em dois universos lingüísticos distintos. Nos exemplos, o predicado exige apenas complementos humanos, porém de um tipo exclusivo. Desse modo, tais restrições devem ser analisadas caso a caso e a partir dessa análise se construir a arborescência semântica das classes de objeto.

A subcategorização dos humanos em classes mais restritas se faz então sob uma base lingüística que repousa essencialmente sobre as compatibilidades presentes nos operadores e seus argumentos. Isso permite gerar para um dado operador uma lista de argumentos possíveis, o que a mera informação de se tratar de um traço humano não permite fazer.

Assim, o modelo de Gross cria codificações a partir dos traços sintático-semânticos dos argumentos uma vez entendidos como discriminantes aos vários empregos dos predicados a que se associam.

Com a mesma perspectiva analítica, Harris [4] já apontava para o estudo que visasse à construção de classes de objetos a partir das ocorrências em textos de domínio específico, no caso usava a Ciência como fonte de exemplificação e trabalhava com a idéia de uma unidade mínima de significação na estrutura da frase simples, entendida como aquela composta por operador e argumento ou, nos termos de Gaston Gross, predicado e argumento.

2. METODOLOGIA

As etapas metodológicas do nosso trabalho são:

a. Compilação do corpus em francês, aplicando a ferramenta BootCaT Tools [1]. O corpus de culinária produzido e disponibilizado pelo grupo de pesquisa lexicográfica COMET [8] está sendo usado para as análises dos predicados do português. Isso não exclui possíveis consultas feitas por meio de sites de busca como google e yahoo.

Todos os corpora estudados são manipulados por meio das ferramentas Unitex [5] e WordSmith Tools [6] – softwares que possibilitam o manuseio dos corpora e o estudo contrastivo dos dados lingüísticos levantados pela pesquisa.

b. Seleção dos predicados a serem investigados, levando-se tanto os valores freqüenciais como a representatividade do termo com relação ao domínio em questão. É importante ressaltar os verbos selecionados são todos aqueles designativos de procedimento exclusivamente culinário. Caracterização dos predicados e formulação da tipologia das classes de objetos (mapa conceitual) dos argumentos relacionados a esses predicados.

c. Estudo comparativo em busca dos traços sintático-semânticos que darão equivalência terminológica entre os predicados.

d. Construção do glossário automático de verbos da culinária baseado em critérios de equivalências de classes de objetos e predicados extraídos a partir de corpora eletrônicos.

3. PRIMEIROS RESULTADOS

A aplicação do modelo de classes de objeto pode ser entendida como uma via de mão dupla na observação dos eventos lingüísticos, pois pode revelar quais predicados e quais argumentos podem se associar, assim como o caminho contrário, ou seja, quais argumentos e predicados aparecem associados.

Outra possibilidade que essa análise nos permite fazer diz respeito às estruturas sintáticas especificas do domínio lexical em questão. Quando pesquisamos o predicado mais freqüente em se tratando de massas (pão, bolo, biscoito, lasanha, etc) em

francês, obtivemos o verbo *cuire* traduzido nos dicionários por cozinhar, cozer.

<cuire> - 4.040.000 ocorrências
<cuire><pâtes> - 998.000 ocorrências
<cuire><lasagnes> - 91.900 ocorrências
<cuire><tarte> – 130.000 ocorrências

No entanto, do ponto de vista morfossintático podemos observar uma particularidade presente nas ocorrências do francês com relação ao verbo *faire* como verbo-suporte, o que revela uma marca de separação sintática e semântica do agente e do receptor da ação, como se pode notar nos seguintes exemplos:

Faire fondre le chocolat et le beurre
Faire réduire une louche de court-bouillon avec les légumes
Faire cuire les biscuits 10 minutes au milieu du four.
<agente> faire + verbo principal + <receptor>

A mesma estrutura sintática não aparece no corpus do português e é fundamental que isso esteja descrito para o usuário do glossário visto que são eventos lingüísticos diversos de uma língua para a outra. Em termos semânticos, citamos algumas ocorrências extraídas da pesquisa feita com o verbo *assar* em corpus de receitas culinárias da Web. Obtivemos os seguintes predicados:

1. Para assar no espeto você precisa preparar a carne
2. Com 3 horas de geladeira já é possível assar os cookies
3. Antes de 3 horas em freezer ainda é complicado assar qualquer pão
4. Receita para assar peixe
5. Semi-assar: assar massa para tortas antes de colocar o recheio
6. Evite abrir a porta do forno quando estiver assando o bolo
7. Asse o pudim em banho-maria por 1 hora

Essa variedade de argumentos do verbo assar não se verifica no francês com os verbos dados como seus equivalentes pelos dicionários bilíngües analisados, no caso *rôtir* e *griller*. Por isso testamos a possível ocorrência desses verbos com tais complementos – *pizza*, *gâteaux*, *cookies*, *pudding* – porém nenhuma ocorrência com tais argumentos foi detectada. Isso demonstra que existe uma lacuna nas opções dicionarizadas por conta dos critérios de equivalência adotados pelos três dicionários examinados.

4. CONCLUSÃO

Os dados levantados pela pesquisa, mesmo que ainda embrionários, já demonstram que o modelo de classes de objetos pode em muito contribuir para a definição de equivalências terminológicas e, assim, constatamos a adequação necessária entre o método e os objetivos do nosso trabalho. Esse trabalho visa a contribuir tanto para os estudos lingüísticos, tradutológicos, como para o aprimoramento de ferramentas de programa de línguas naturais (PLN). O emprego prático dos resultados finais pode ser bastante amplo como, instalação do glossário em sites do domínio, ferramenta de tradução automática ou mesmo conteúdo lingüístico para ferramenta de desambigüização. Vale dizer que todo o material compilado para aplicação nesta pesquisa será disponibilizado para eventuais futuras pesquisas.

5. REFERÊNCIAS

[1] Baroni M.; Bernardini. 2004. S. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC. DOI= http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf

[2] Gross, Gaston. 1994. Classes d'objet et descriptions des verbes. Langages n. 115 p. 15-30. DOI= http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_1994_num_28_115_1684?_Prescripts_Search_isPortletOuvrage=false

[3] Gross, Gaston. 1995. À propos de la notion d'humain. In Lexiques-Grammaires comparés en français, Labelle, Jacques et Christian Leclère (dir.). Amsterdam : John Benjamins.

[4] Harris, Zellig S. 1952. Discourse Analysis. Language 28:1 (Reimpresso em Papers in Structural and Transformational Linguistics. Dordrecht: D. Reidel. 1981. 313-348)

[5] Paumier, Sébastien. 2004. Unitex - Manuel d'utilisation, Univ. Marne-la-Valée. DOI= http://www-igm.univ-mlv.fr/~unitex/UnitexManual.pdf

[6] Scott, Mike. 1998. WordSmith Tools Version 3. Oxford University Press, Oxford. DOI= http://www.lexically.net/wordsmith/

[7] Tagnin, Stella. E.O.; Teixeira, Elisa Duarte. 2004. Lingüística de Corpus e Tradução Técnica – relato da montagem de um corpus multivarietal de culinária. Tradterm, São Paulo, v. 10, n. 1, p. 313-358.

[8] Teixeira, Elisa Duarte. 2003. Em busca de um novo modelo tecno-formal para a construção de dicionários técnicos bilíngües – o exemplo da culinária. Revista Intercâmbio, São Paulo, SP, v. XII, p. 243-251.