Constituição de Um Dicionário Eletrônico Trilíngue Fundado em Frames a partir da Extração Automática de Candidatos a Termos do Domínio do Turismo

Simone Rodrigues Peron-Corrêa[1], Tiago Timponi Torrent[1]

[1]FrameNet Brasil – Programa de Pós-Graduação em Linguística – Universidade Federal de Juiz de Fora (UFJF)
Rua José Lourenço Kelmer, s/nº, Campus Universitário
36036-900 – Juiz de Fora – Minas Gerais – Brasil

speronjf@yahoo.com.br, tiago.torrent@ufjf.edu.br

Abstract. *This paper presents the methodology used for the creation of a frame-based trilingual terminological dictionary for the Tourism domain. This dictionary is one of the functions of a virtual assistant app designed for tourists. The work involved the compilation of a trilingual comparable corpus for Brazilian Portuguese, Spanish and English, composed of 1 million words per language. The extraction of candidate terms is carried using the TERMOSTAT tool and the electronic dictionary relies on FrameNet Brasil infrastructure and methodology.*

Resumo. *Este trabalho apresenta a metodologia utilizada para a constituição de um dicionário terminológico trilíngue fundado em frames para o domínio do Turismo. Tal dicionário constitui uma das funções de um aplicativo desenvolvido para funcionar como um assistente virtual para turistas. O desenvolvimento do dicionário envolveu a constituição de um corpus especializado trilíngue comparável para o português brasileiro, o espanhol e o inglês, composto por 1 milhão de palavras por língua. A extração dos candidatos a termos é realizada a través da ferramenta TERMOSTAT e a modelagem do dicionário eletrônico, através da FrameNet Brasil.*

1. Introdução

O trabalho aqui apresentado tem como escopo principal realizar a prospecção de termos para o domínio do Turismo, de modo a aperfeiçoar recurso lexical eletrônico já existente, denominado m.knob (*Multilingual Knowledge Base*), produzido para atender às demandas das Olimpíadas de 2016, sediadas no Rio de Janeiro. Tal recurso é uma aplicação computacional da Semântica de Frames [Fillmore 1982; 1985] cuja função primária é disponibilizar para o turista um guia turístico multilíngue (português, inglês e espanhol), combinado a um tradutor de sentenças e um dicionário de domínio específico, a Diciopédia.

Quando do lançamento da versão alfa para os Jogos Olímpicos, a Diciopédia do m.knob contava com 2.316 termos, sendo 1.153 em português, 777 em inglês e 386 em espanhol. Tais números apontam para a necessidade de melhoria desse recurso, de modo a equilibrar o quantitativo de termos. É visando a atender a esta necessidade e também a relatar o uso e as limitações de uma metodologia automática para a extração de termos, aquela possibilitada pelo *software* TERMOSTAT [Drouin 2003], que este trabalho se desenvolve.

2. Terminologia Baseada em Frames

O enfoque teórico-prático denominado Terminologia baseada em Frames, doravante TbF [Faber et al. 2005; 2006] apresenta como um dos principais objetos de estudo as unidades lexicais e terminológicas, que são determinadas de acordo com o contexto e estão intrinsicamente relacionadas a um domínio especializado. Tais unidades são consideradas como configurações de eventos complexos e estão interligadas através dos conceitos do domínio [L'Homme 2010]. Neste sentido, ao tratarmos de um domínio específico, considera-se primordial identificar as entidades e as ações básicas desse campo de especialidade, que se interconectam através de diversos tipos de relações conceituais. Partindo dessa premissa, podemos considerar o conhecimento especializado como uma subdivisão do conhecimento geral.

Segundo Faber e colaboradores (2005), as áreas especializadas devem ser organizadas em frames, ou seja, "qualquer sistema de conceitos relacionados de tal forma que, para entender um deles, é necessário compreender toda a estrutura na qual ele se encaixa" [Fillmore 1982:111], de modo que estes sirvam tanto de bases para a localização de conceitos hierárquicos dentro de um domínio especializado, quanto de um modelo de definição. Outro aspecto importante ressaltado por Faber e colaboradores (2005), no que diz respeito à noção de frame, é que através dele os conceitos definidos podem ser situados dentro de um contexto em que as categorias são relacionadas entre si. Dessa forma, ao vincularmos tais embasamentos teóricos na constituição de produtos lexicográficos e terminológicos, teremos, segundo Faber et al (2005:8),

> ...uma organização de áreas especializadas baseada em frames em que um frame dinâmico orientado para o processo fornece as bases conceituais para a localização das sub-hierarquias de conceitos dentro de um evento de domínio especializado, e a elaboração de um modelo de definição, abrindo assim a porta para uma representação mais adequada dos campos especializados, bem como fornecendo uma maneira melhor de ligar os termos aos conceitos[1].

Com base nesses princípios teóricos da Terminologia baseada em frames [Faber et al. 2005; 2006] e na Semântica de Frames [Fillmore 1982; 1985], alguns aspectos tornam-se relevantes na construção de um modelo de evento para um domínio de conhecimento especializado, como, por exemplo, auxiliar o usuário a processar o conteúdo conceitual com mais facilidade, já que fundamentadas na gestão da terminologia, as informações essenciais serão configuradas através de redes, que, por sua vez, são divididas em domínios e, estes, em frames, que poderão passar por vários níveis de especificidade usando herança hierárquica. Tais aspectos demonstram como os eventos podem ser considerados dinâmicos e modeláveis, uma vez que são considerados flexíveis, além de configurarem entidades do mundo real, que podem desempenhar diferentes papéis.

Segundo Krieger e Finatto (2004:17), os termos transmitem conteúdos específicos de cada área, com a finalidade de representar e transmitir o conhecimento especializado.

---

[1] (…) a frame-based organization of specialized fields in which a dynamic process-oriented frame provides the conceptual underpinnings for the location of sub-hierarchies of concepts within a specialized domain event, and the elaboration of a definition template, thus opening the door to a more adequate representation of specialized fields as well as supplying a better way of linking terms to concepts.

Sob o enfoque da Terminologia Sociocognitiva [Temmerman 1997; 2000], há um aspecto relevante, que é assim considerado também pela TbF: tomar os termos como unidades de compreensão e representação, configurando-os como modelos cognitivos e culturais, que designam categorias de uma estrutura prototípica delimitável.

Além das bases teóricas já mencionadas, a Terminologia também se ancora nos dados empíricos fornecidos pela Linguística de Corpus, uma vez que os *corpora* são considerados como fonte de seleção dos termos. De acordo com Cabré (1999:298), o *corpus* deve cumprir uma série de condições para garantir a confiabilidade dos resultados, tais como: ser pertinente, isto é, representativo do campo pesquisado; ser completo, incluindo, assim, todos os aspectos que devem estar relacionados com o tema investigado; ser atual, para que a lista de termos extraída reflita a realidade linguística do âmbito em questão, e, por último, ser original, ou seja, estar expresso na língua com que se pretende trabalhar.

Desde a década de 80, temos obtido avanços tecnológicos relacionados tanto aos trabalhos terminológicos quanto lexicográficos realizados com base em *corpora* automatizados. Podemos enumerar diversas vantagens dessa abordagem, tais como, manusear maior quantidade de textos e dados, conferindo maior flexibilidade e uma projeção múltipla destinada a diferentes usuários, além de proporcionar o armazenamento em grandes bancos de dados.

Para a extração automática de candidatos a termos para o domínio do Turismo relatada neste trabalho, utilizou-se a ferramenta TERMOSTAT [Drouin 2003], a qual é um *software* destinado à extração de termos simples e complexos. A ferramenta compara um corpus de especialista submetido pelo usuário a um corpus de domínio genérico de referência e extrai aqueles termos e colocações cujas frequência e colocações no *corpus* de especialista destoam daquelas encontradas no *corpus* genérico. A principal finalidade do TERMOSTAT é diminuir a quantidade de "ruídos", ou seja, de termos que não são correspondentes à área especializada em análise.

3. Aplicativo *Multilingual Knowledge Base (mknob)*

A extração de candidatos a termos do Turismo está sendo realizada com a finalidade de aperfeiçoar a modelagem já existente e auxiliar na prospecção de novos termos a serem acrescentados no aplicativo mknob[2], que se caracteriza por ser uma aplicação para usuários não especialistas nos domínios do Turismo e dos Esportes, que foi produzido para atender às demandas das Olimpíadas de 2016, sediada no Rio de Janeiro. É importante ressaltar que, mesmo depois do evento, o aplicativo pode ser usado como um guia turístico multilíngue. Tal recurso é uma aplicação computacional legível tanto por máquina quanto por homens, cuja função primária é disponibilizar para o turista um guia turístico multilíngue (português, inglês e espanhol), que recomenda locais e sugere atividades com base na interpretação semântica de *inputs* do usuário, além de traduzir sentenças (En, Es, Pt) nos domínios do Turismo e dos Jogos Olímpicos. O tradutor automático está em desenvolvimento, e o seu diferencial está em propor traduções com base na Semântica de Frames [Fillmore 1982; 1985], que elege os padrões de valência das Unidades Lexicais no ranqueamento das escolhas, além da implementação de novas

[2] http://www.ufjf.br/framenetbr/m-knob

relações ancoradas nas estruturas *qualia* [Pustejovsky 1995], em interface com Ontologias, Web Semântica e Dados Ligados.

De maneira mais elucidativa, explicamos a seguir as 3 funções principais do aplicativo:

a) Sistema de Recomendação baseado em busca semântica (Guia Local): esta função sugere atividades turísticas e de entretenimento, com base nos *inputs* que o usuário insere, em língua natural, na interface conversacional do aplicativo. Desse modo, a partir da definição dos interesses do turista, ele receberá informações pertinentes a eventos e atrações turísticas correspondentes ao que foi assinalado. A partir das recomendações, o usuário poderá acessar mais detalhes sobre os locais sugeridos, como mapas, horários, site oficial, descrição, assim como avaliar o nível da recomendação. Essas informações são extraídas automaticamente de bases de dados abertos disponíveis online, como Google Places e WikiData; já o processamento semântico, o armazenamento e o tratamento dessas informações são realizados pela FrameNet Brasil.

b) Tradutor híbrido (Intérprete Pessoal): esta função ainda está em desenvolvimento. A proposta inicial parte da premissa de que tanto a FrameNet quanto o Constructicon podem contribuir significativamente para a Tradução Automática, pois, em suas bases de dados, encontra-se um modelo do conhecimento acerca dos *frames* e da gramática de uma língua. Portanto, nossa proposta é a de que tais bases de conhecimento (frames e construções) possam alimentar os sistemas de tradução automática, aperfeiçoando a construção de equivalentes de tradução, a partir da anotação de valências sintático-semânticas, da organização ontológica do léxico e da incorporação das Estruturas *Qualia* [Pustejovsky 1995].

c) Diciopédia: esta função é considerada um repositório multilíngue, uma vez que o usuário pode acessar os dados em português, inglês ou espanhol. Nele encontrará conceitos e palavras relacionados aos domínios específicos do Turismo e dos Esportes, abarcados pelo m.knob. Cada entrada da diciopédia está vinculada a uma unidade lexical que evoca um frame um frame. Nela encontramos a definição e os possíveis equivalentes de tradução nas línguas mencionadas. Também encontraremos outras palavras relacionadas ao frame em análise ou interligados a um outro frame próximo. Será possível, durante a interação, sugerir novas palavras, conceitos, traduções, assim como propor correções às traduções inadequadas. As traduções que são apresentadas na Diciopédia são geradas automaticamente, seguindo duas metodologias distintas:

i. para nomes de entidades, ou seja, que se referem às pessoas, objetos e lugares, os dados são extraídos automaticamente de uma base de dados ligados, denominada BabelNet e à posteriori são validados pelos linguistas do projeto FrameNet Brasil. A BabelNet é um recurso computacional que liga informações extraídas de bases de dados como a Wikipédia a bases lexicais como a WordNet. Conforme descreve Navigli (2012:2), a BabelNet tem como resultado "um 'dicionário enciclopédico' que fornece babel synsets, ou seja, conceitos e entidades nomeadas lexicalizados em muitas línguas e conectados através de uma grande quantidade de relações semânticas." Esse recurso visa a fornecer uma cobertura lexicográfica e enciclopédica completa, incluindo 14 milhões de entradas em 271 línguas.
ii. para os verbos e os nomes que indicam evento, os dados são calculados tomando como base as anotações lexicográficas do projeto FrameNet Brasil, que

consideram as valências sintáticas e semânticas, de acordo com o uso destas em textos reais [Peron-Corrêa et al. 2016].

O trabalho de prospecção de termos relatado neste artigo tem por primeira aplicação enriquecer a Diciopédia.

4. Metodologia

O *corpus* de especialista utilizado foi constituído pela FrameNet Brasil e para o Português contém 536.918 palavras retiradas de guias turísticos e 551.932 palavras de blogs de viagem; para o Espanhol 605.782 palavras de guias e 500.183 palavras de blogs e para o Inglês são 589.203 palavras de guias e 502.510 palavras de blogs.

Após a análise automática do corpus de especialista pelo TERMOSTAT, cabe aos lexicógrafos da FrameNet Brasil validar os candidatos a termos. Tal validação consiste em um processo de três estágios:

1. Busca-se o candidato a termo na base corrente do mknob:
    a. Caso ele já tenha sido incluído, ele é considerado válido.
    b. Caso contrário, passa-se ao estágio 2.
2. Buscam-se no corpus sentenças que contenham o candidato a termo e realiza-se uma pré-anotação semântica das sentenças:
    a. Caso as anotações se enquadrem em um frame já incluído na base corrente do mknob, o termo é considerado válido.
    b. Caso contrário, passa-se ao estágio 3.
3. Discutem-se com a equipe de lexicógrafos as anotações não-conformes:
    a. Caso elas constituam um frame relevante, o termo é considerado válido.
    b. Caso contrário, ele é descartado.

5. Resultados Preliminares da Análise do TERMOSTAT

A comparação realizada pelo TERMOSTAT gera quatro arquivos de saída: (i) *Frequency*: que mostra a lista de candidatos a termos em ordem decrescente de probabilidade; (ii) *Percentage*: que mostra os candidatos a termos agrupados percentualmente de acordo com a suas propriedades morfossintáticas; (iii) *Collocations*: que mostra os candidatos a termos verbais seguidos dos seus argumentos prototípicos no corpus; (iv) *Bigrams*: que mostra os *bigrams* mais frequentes. Para este trabalho, realizamos uma análise apenas do arquivo de frequência tanto para o Português quanto para o Espanhol. No Português obtivemos 15217 termos selecionados, nos quais encontramos substantivos, adjetivos, verbos, advérbios e expressões. No Espanhol, foram 16697 termos, também distribuídos nas diversas classes mencionadas acima.

Esse primeiro resultado da tabela de frequência mostra em ordem decrescente os termos, com suas respectivas frequências, ao lado de sua especificação, ou seja, o resultado que contrasta essa frequência e as propriedades colocacionais encontradas com os mesmos parâmetros nos *corpora* de referência. Quanto maior a especificação, maior a chance de a palavra encontrada ser um termo do domínio em análise.

Tabela 1: Candidatos a termos do Domínio do Turismo extraídos pelo TERMOSTAT

Candidato	Frequência	Especificação	Frame do termo no mknob	Frame candidato no mknob
praia	5207	165.02	Locais_naturais	
rio	3728	123.33	Locais_naturais	

hotel	2924	117.59	Acomodação	
passeio	2384	113.3	Deslocar-se	
pousada	1834	105.41	Hospedagem	
cidade	4499	103.49	Locais_políticos	
ônibus	1497	97.71	Meios_de_transporte	
restaurante	1694	93.62	Alimentação	
dica	1159	85.63		nulo
viagem	1962	84.57	Viagem	
prato	1169	78.54	Alimentos_e_bebidas	
ótimo	921	76.62	Ser_desejável	
bar	1271	74.77	Alimentação	
parque	1690	71.85		Locais_naturais
café	1186	71.7	Alimentação	
ficar	3949	71.15	Hospedar-se	
pegar	1000	68.76		Usar_veículo
trilha	700	66.6		Vias
opção	1340	65.94	Possibilidades	
endereço	721	64.92		nulo
atração	649	64.3	Turismo_de_atração	
restaurante	789	62.36	Alimentação	
aeroporto	1035	61.77	Transporte	
lindo	691	61.71		Estética
mapa	836	61.68		Texto
café da manhã	585	60.97	Serviço_turístico	
oferecer	1281	60.51	Atrair_turista	
localizar	822	60.37	Ser_localizado	
mar	1350	59.54		Locais_naturais
bom	2548	59.23		Ser_desejável
gratuito	706	58.35		Custo
quarto	1052	57.3		Subpartes_de_prédios_e_locais
lagoa	650	57.17		Locais_naturais
cachoeira	507	56.82	Locais_naturais	
centro	2198	56.69		Locais_por_uso
morro	530	56.4	Locais_naturais	
táxi	588	54.1	Meio_de_transporte	
praça	1046	54.02	Locais_por_uso	

A título de ilustração, fizemos um recorte dos primeiros quarenta termos extraídos do TERMOSTAT para a Língua Portuguesa, conforme constam na Tabela 1.

Na primeira coluna, temos os candidatos a termos; na segunda, temos a numeração que indica a frequência deles; na quarta temos os números que indicam a especificidade.[3] Acrescentamos duas colunas que apresentam os resultados encontrados na validação dos candidatos a termos. Para os termos que já se encontram no aplicativo mknob, indicamos o nome do frame evocado; caso o termo não esteja no aplicativo, realizamos o processo descrito nos passos 2 e 3 da metodologia.[4]

Conforme se observa na Tabela 1, dos 40 termos extraídos pelo TERMOSTAT, apenas dois (5%) não devem ser incluídos na base do mknob, sendo que 11 novos termos (27.5%) foram descobertos e adicionados à base. Os resultados obtidos têm nos auxiliado no aperfeiçoamento da modelagem computacional já realizada, na implementação de novos dados que se fizerem necessários, na proposição de um padrão para construção de dicionários eletrônicos fundados em frames, assim como na projeção de novos projetos lexicográficos que atendam outros domínios especializados.

6. Considerações Finais

Neste artigo, apresentamos as principais bases teóricas que foram fundamentais para o desenvolvimento do aplicativo mknob, que visa atender a usuários não especialistas, nos domínios do Turismo e do Esporte. Realizamos um recorte curto dos dados fornecidos pela ferramenta TERMOSTAT para a língua portuguesa, com o objetivo de apresentar seu potencial para a extração de termos em língua portuguesa. O mesmo procedimento realizado para a língua portuguesa está sendo realizado para as línguas espanhola e inglesa. Com a análise desses termos, nossa proposta é equiparar os dados nas três línguas do aplicativo de modo a torná-lo ainda mais eficiente para que foi desenhado, mostrando assim uma maior cobertura de pesquisa e dados sobre os domínios especializados tanto do Turismo quanto do Esporte. Além disso, a prospecção de termos para um determinado domínio colabora no sentido de compreendermos e identificarmos a diferença do uso de tais palavras em domínio mais genérico ou específico, uma vez que o processo e a constituição dos termos de domínio específico requerem maior cuidado, pois não se trata apenas de listar, e sim de compreender os empréstimos linguísticos dentro de um contexto.

Referências Bibliográficas

Cabré, M. T. (1999). Una nueva teoría de la terminología: de la denominación a la comunicación. In: CABRÉ, M. T. La terminología: representación y comunicación. Barcelona: Universitat Pompeu Fabra.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. In *Terminology*, 9(1), pages 99—115. John Benjamins.

Faber, P., Márquez-Linares, C. and Vega-Expósito, M. (2005). Framing Terminology: A process-oriented approach. In *META*, 50(4), http://www.erudit.org/livre/meta/2005/000255co.pdf, January.

[3] Para análise neste artigo excluímos os dados referentes às variantes ortográficas e classe gramatical das palavras, que também são dados fornecidos pelo TERMOSTAT.

[4] Os frames podem ser consultados em http://www.ufjf.br/framenetbr/dados.

Faber, P., Martínez, S. M., Prieto, M. R. C., Ruiz, J. S., Velasco, J. A. P., Arauz, P. L., ... and Expósito, M. V. (2006). Process-oriented terminology management in the domain of Coastal Engineering. In *Terminology*, 12(2), pages 189—213.

Faber, P., Araúz, P. L., Prieto Velasco, J. A., and Reimerink, A. (2006). Linking images and words: the description of specialized concepts. In *International Journal of Lexicography*, 20(1), pages 39—65.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111—137. Hanshin Publishing.

Fillmore, C. J. (1985). Frames and the semantics of understanding. In *Quaderni di Semantica*, 6(2), pages 222–254.

Krieger, M. G.; Finatto, M. J. B.(2004). Introdução à terminologia: teoria e prática. São Paulo: Contexto.

L'Homme, M. C. (2010). Designing terminological dictionaries for learners based on lexical semantics: The representation of actants. *Specialised Dictionaries for Learners, Berlin/New York: De Gruyter*, 141—153.

Peron-Corrêa, S., Diniz, A., Lara, M., Matos, E., & Torrent, T. (2016). FrameNet-Based Automatic Suggestion of Translation Equivalents. In *International Conference on Computational Processing of the Portuguese Language* (pp. 347-352). Springer International Publishing.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, USA: MIT Press.