Compilação e anotação de corpus – um estudo sobre os compostos nominais

Lílian Figueiró Teixeira[1], Lucas Lermen[2]

[1]Programa de Pós-Graduação em Lingüística Aplicada – Universidade do Vale do Rio dos Sinos (UNISINOS)

[2]Faculdade de Engenharia da Computação – Universidade do Vale do Rio dos Sinos (UNISINOS)

lilianjoy@gmail.com, lucaslermen@gmail.com

Resumo. *A utilização de recursos computacionais para o estudo da língua tem se tornado cada vez mais freqüente. Neste trabalho, nosso objetivo é refletir sobre a metodologia que adotamos para realizar a coleta e compilação de corpus paralelo. Como parte de pesquisa em nível de mestrado, cujo propósito é estudar a semântica dos compostos nominais através da sua tradução do inglês para o português, a compilação do corpus exigiu uma série de procedimentos computacionais. Inicialmente foi feita a coleta de dez edições de uma revista em inglês e a sua tradução em português. Após a compilação do corpus, foram identificados os compostos nominais formados por dois substantivos (NN). Para fazer a extração, tornou-se necessário a etiquetação dos textos em inglês. Com o corpus devidamente etiquetado buscaram-se os compostos NN e a sua freqüência foi medida. Entre os resultados da extração, encontram-se predominantemente seqüências que ocorrem uma única vez no corpus (hapax). Posteriormente o corpus foi alinhado com o propósito de estudar os equivalentes de tradução dos compostos. Espera-se com este estudo contribuir para as pesquisas em PLN e em especial para as tarefas de tradução automática.*

Abstract. *Computational tools as resources for language studies has become very popular among linguistics. In this paper, our purpose is to consider a methodology adopted to collect and compile a parallel corpus. This experiment is part of a Master's research which aims the analysis of the compound nouns in English and their translation to Portuguese. First ten editions of a magazine were compiled in the two languages. Then the noun compounds formed by two nouns (NN) were identified. For the extraction, the texts in English were tagged and the number of occurrences for each compound noun candidate was counted. As the extraction results, most of the NN frequencies took place only once in the corpus (hapax). In order to analyse the translation equivalents for the expressions, the corpus was aligned. We expect to contribute to researches in NLP with this study, specially for the machine translation tasks.*

Palavras-chave: Lingüística de *Corpus*; *corpu*s paralelo; compostos nominais

1. Recursos computacionais e o estudo da língua

Uma das formas mais práticas e rápidas de se analisar a língua em uso é através dos recursos computacionais. Quando um professor, um aprendiz ou até mesmo um simples falante de uma língua qualquer tem dúvidas sobre o uso de alguma expressão ou palavra basta consultar o computador mais próximo. A sua fonte de informações pode ser um dicionário digital, enciclopédias on-line, *websites* especializados ou até mesmo os resultados de um *site* de busca.

No meio acadêmico, um dos principais recursos utilizados pelo lingüista tem sido o estudo de *corpus*, o qual constitui um conjunto de textos em formato digital compilado de forma que possa servir para pesquisas com fins lingüísticos. Um dos primeiros estudos que teve como base os dados de um *corpus* foi realizado em 1921 por Thorndike (*apud* Berber Sardinha, 2000). No seu trabalho, ele identificou as palavras mais freqüentes da língua inglesa através do levantamento manual de um *corpus* de 4,5 milhões de palavras.

O objetivo deste trabalho é refletir sobre a metodologia que adotamos para realizar a compilação e a análise de um *corpus* paralelo, que é constituído por textos na sua língua original e com a sua tradução alinhados frase por frase. Este estudo faz parte de uma pesquisa em nível de mestrado, cujo propósito é estudar a semântica dos compostos nominais através de sua tradução do inglês para o português.

Como nem sempre é possível encontrar um *corpus* disponível e apropriado para os objetivos de pesquisa, optamos pela compilação de um *corpus*, bem como, a criação de algumas ferramentas computacionais necessárias para obter as informações pertinentes ao estudo proposto. Apresentaremos nas próximas seções como se deram a compilação do corpus e a aplicação de ferramentas como itemizador, etiquetador morfológico, extrator e alinhador. Finalizaremos com uma pequena análise dos compostos nominais a partir dos resultados obtidos através destes recursos.

2. Compilação de corpus

Em um *corpus* paralelo, as traduções são feitas por seres humanos e servem para que se percebam as diferentes possibilidades de tradução de uma palavra ou expressão. Para essa primeira análise dos compostos, um *corpus* paralelo formado por dez edições, entre 2007 e 2008, da revista *National Geographic* foi compilado. A sua tradução é encontrada na mesma edição da revista, porém na sua versão brasileira. As duas reportagens foram obtidas nos *sites* da revista[1].

O *corpus* paralelo consiste basicamente em dois textos, um original e sua tradução, organizados de forma que cada linha esteja alinhada com o seu correspondente na segunda língua. As principais aplicações desse tipo de recurso estão relacionadas aos estudos de tradução, comparar semelhanças e diferenças entre original e traduções, ou até mesmo comparar as diferentes traduções de uma mesma obra.

[1] *National Geographic Magazine*, disponível em: <http://ngm.nationalgeographic.com/> e *National Geographic* Brasil, disponível em: <http://viajeaqui.abril.uol.com.br/ng/>

Em língua portuguesa ainda há poucos *corpora* desse tipo, entre eles temos o COMPARA (Frankenberg-Garcia, Santos, 2002), que apresenta romances em português europeu e brasileiro e suas traduções para o inglês. Há também traduções no sentido inverso, do inglês para o português.

Com o objetivo de disponibilizar um *corpus* paralelo em português e inglês que trate de outros gêneros textuais, resolvemos compilar um *corpus* com reportagens da revista *National Geographic*. Essa revista tem o seu original em inglês e também possui uma versão no Brasil.

Antes de apresentarmos o experimento realizado, é importante esclarecemos o conceito de *corpus* paralelo. Neste trabalho, adotamos a definição de Frankenberg-Garcia e Santos (2002), para quem o *corpus* paralelo é uma coleção bilíngüe de textos alinhados com as suas traduções. Esse tipo de *corpus* também é chamado *corpus de traduções* na tradição da lingüística contrastiva.

Entre os usos de *corpora* paralelos, McEnery e Wilson (1993) citam a tradução automática e a criação de léxicos. Tendo um *corpus* paralelo como base de dados, servindo como *corpus* de treinamento, podem-se criar métodos probabilísticos que auxiliem a tarefa de tradução automática. É possível extrair de um *corpus* paralelo as palavras correspondentes em mais de uma língua ou até mesmo expressões multivocabulares que podem ser incluídas em um léxico, ou dicionário multilingüe. Através de estudos de freqüência, é possível construir uma base de dados terminológica de textos especializados.

McEnery e Wilson (1993) apresentam outra definição para *corpus* paralelo, mas que em sua essência não difere da apresentada por Frankenberg-Garcia e Santos (2002). Os autores chamam de paralelos os *corpora* que contêm o mesmo texto em mais de uma língua, sendo que eles são tipicamente bilíngües, podendo ser também multilingües.

Para a compilação, copiamos e colamos no bloco de notas os arquivos correspondentes às edições entre agosto de 2007 a maio de 2008. Optou-se por arquivos em formato *txt*, pois este formato é pré-requisito para o seu processamento em diversos programas utilizados para o estudo de *corpora*. No tabela 1, temos as principais informações dos arquivos.

Corpora compilados da revista *National Geographic*				
Arquivo	Idioma	Tamanho	*Tokens*[2]	*Types*
national_geographic_ingles.txt	inglês	1.227 KB	212.552	20.263
national_geographic_portugues.txt	português	1.236 KB	208.201	24.327

Tabela 1. Dados dos *corpora*

No *site* não há informações sobre os tradutores do artigo, mas, considerando que a edição brasileira é publicada pela editora Abril, que está há anos no mercado,

[2] As informações quanto ao número de *tokens* (palavras) e *types* (formas diferentes) foram obtidas através da ferramenta *WordSmith Tools*, versão 5 (Scott, 1996).

imaginamos que os tradutores sejam profissionais qualificados. Procuramos manter a formatação apresentada no *site*, ou seja, a separação dos parágrafos, mas desconsideramos as imagens, *links* e quaisquer outros dados que não sejam texto.

3. Extração dos dados

O estudo da dissertação de mestrado é sobre os compostos nominais formados por dois substantivos na língua inglesa e dos seus correspondentes de tradução para o português. Desta forma, precisávamos extrair do *corpus* uma seqüência de dois substantivos sem que houvesse outro substantivo antes ou depois. Também parecia interessante obter uma lista de todas estas expressões seguidas pela quantidade de vezes em que elas ocorrem no *corpus*. Para obter estes dados, precisávamos inicialmente de um *corpus* com anotação morfológica, ou seja, com a informação de que a palavra é substantivo, adjetivo, verbo, etc. Há programas que criam etiquetas para cada uma destas informações automaticamente. Como o nosso *corpus* não era etiquetado, ele precisou passar por este processamento, pois só assim, outro programa poderia identificar a informação de que necessitávamos.

Optamos pelo etiquetador *TreeTagger* para a língua inglesa, por ser uma ferramenta gratuita e com bons resultados. No entanto, antes de passar os dados neste programa, o *corpus* requer um pré-processamento, ele precisa estar itemizado, ou seja, uma palavra por linha. Para fazer isto, foi criado um itemizador que utiliza a arquitetura Java J2SE. Após ter um *corpus* com as anotações morfológicas, precisamos extrair as seqüências de dois substantivos. Como não foi encontrado nenhum extrator apropriado e que fosse gratuito, também criamos esta ferramenta que tem como base as etiquetas do *TreeTagger* e utiliza a mesma arquitetura do itemizador. Essas três ferramentas citadas foram utilizadas inicialmente para o processamento do *corpus* em língua inglesa. O *corpus* em português também precisou ser itemizado para que pudesse ser alinhado com o *corpus* em inglês através do alinhador *Vanilla Alligner*.

Os resultados obtidos com cada uma dessas ferramentas serão expostos em maiores detalhes nas próximas seções.

3.1. Itemizador de texto

Os *corpora* em inglês e português precisaram ser itemizados, já que este formato é pré-requisito, tanto para o etiquetador *Tree Tagger* quanto para o alinhador *Vanilla*. Este programa que formata o texto em uma palavra por linha, chamado por nós de itemizador, utiliza a arquitetura Java J2SE. A principal vantagem desta arquitetura é o fato de ser multiplataforma, ou seja, é independente de sistema operacional, podendo funcionar em *linux, windows, mac*, entre outros.

Para separar cada palavra, o programa identifica os espaços em branco e os substitui por uma entrada. Desta forma, os sinais de pontuação não são separados das palavras. Expressões compostas separadas por hífen e siglas também são mantidas na mesma linha, conforme a figura 1. O itemizador salva o texto em itens no mesmo diretório em que o texto fonte estiver localizado, apenas acrescentando "Itemizado" ao nome do novo arquivo.

O único pré-requisito para o processamento dos textos é que eles devem estar no formato ANSI da extensão *txt*. Além disso, é necessário possuir o *Java Runtime*

Environment (JRE) instalado no computador. O arquivo de saída deve ser aberto através do *WordPad*.

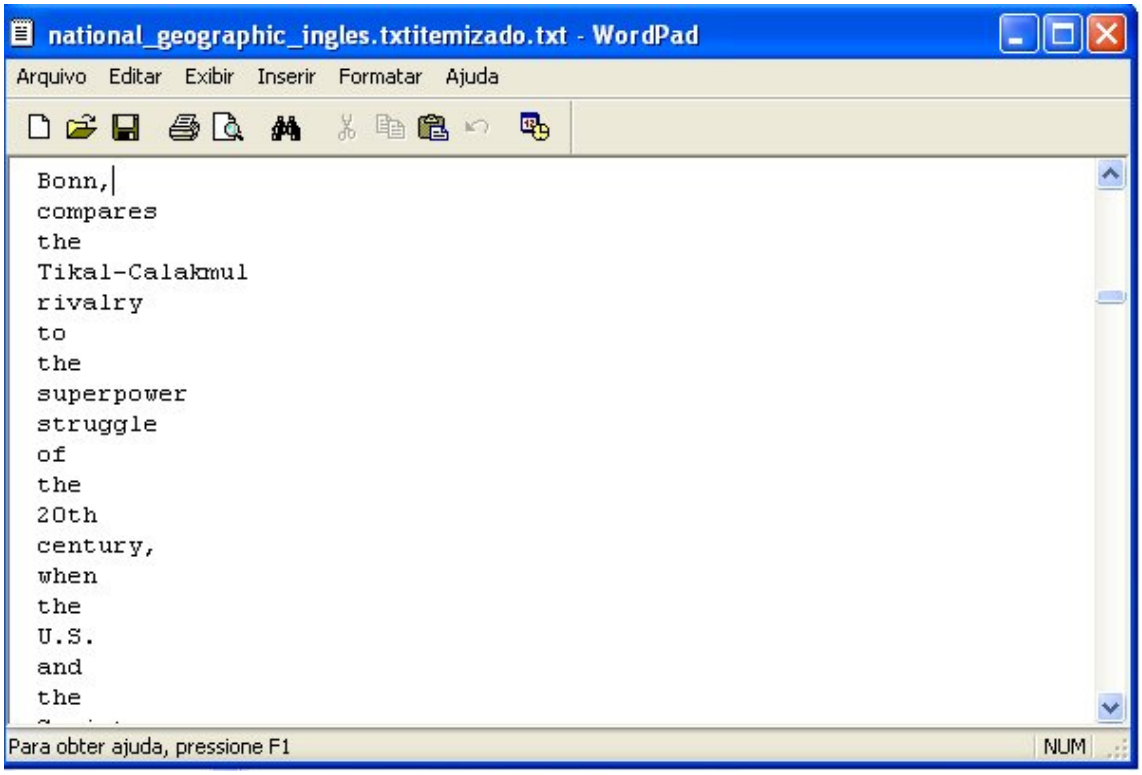

Figura 1. *Corpus* itemizado

3.2. Etiquetador *TreeTagger*

O *TreeTagger*[3] é um etiquetador *part-of-speech (POS)*, ou seja, é um sistema que faz automaticamente o reconhecimento das categorias gramaticais. Ele foi desenvolvido na Universidade de Stuttgart na Alemanha e é utilizado em mais de dez idiomas diferentes, dentre eles inglês, francês, alemão e italiano.

Como o nosso objetivo é conseguir extrair do *corpus* seqüências formadas por dois substantivos, consultamos o manual de etiquetas do *TreeTagger* (Santorini, 1990). Entre as etiquetas que nos interessam encontramos as seguintes: NNS (substantivo comum, plural) e NN (substantivo comum, singular). Para diferenciar o singular do plural, o programa identifica o verbo que acompanha o substantivo. É o verbo que irá determinar se o substantivo está no plural ou singular. Isso resolve o problema de substantivos que no singular terminam em "s", como *linguistics*. Há etiquetas referentes aos nomes próprios, mas estas não são interessantes para o nosso trabalho, já que analisaremos apenas os compostos nominais formados por substantivos comuns.

Uma informação importante quanto às escolhas do etiquetador, encontrada no mesmo manual, refere-se ao fato de que um substantivo modificador será etiquetado como substantivo e não como adjetivo. A importância desta escolha se dá, pois caso o

[3] Disponível em: <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

etiquetador anotasse um substantivo modificador como adjetivo, a nossa pesquisa seria muito prejudicada, já que buscamos por seqüências formadas por dois substantivos. Por outro lado, as cores também são etiquetadas como substantivos, embora nem sempre sejam considerados substantivos em determinadas frases.

O *TreeTagger*, segundo dados de Schmid (1994), utiliza um modelo probabilístico baseado em árvores binárias de decisões, isto é, a partir de trigramas, seqüências de três palavras encontradas em um *corpus*, criam-se relações entre as classes gramaticais. Para chegar à conclusão se determinada palavra é um substantivo ou um adjetivo é necessário responder afirmativamente ou negativamente a perguntas quanto às palavras que aparecem ao seu redor. A medida em que cada resposta afirmativa é dada, as informações na árvore são conectadas chegando-se a uma resposta, à folha da árvore. O etiquetador também possui um léxico que foi criado a partir de uma parte do *corpus Penn Treebank*. Dois milhões de palavras deste *corpus* foram etiquetadas e serviram de treinamento, ou seja, a partir dos dados obtidos neste *corpus*, criam-se regras probabilísticas que possam ser utilizadas na tarefa de etiquetação de quaisquer outros *corpora*.

Entre os resultados relatados por Schmid (1994), o *TreeTagger* atinge em torno de 96% de precisão, mostrando-se um etiquetador bastante eficiente. Na figura 2, há uma parte do *corpus* etiquetado, que é exibido em três colunas, a primeira com as palavras conforme são encontradas no texto, a do meio com as etiquetas morfológicas e a terceira com a forma canônica da palavra, como o infinitivo do verbo e o substantivo no singular e sem marca de gênero.

Figura 2. *Corpus* etiquetado

3.3. Extrator de seqüências NN

Após ter o *corpus* etiquetado, precisávamos extrair as seqüências formadas por dois substantivos com o objetivo de chegarmos aos compostos nominais. Para este fim, foi

desenvolvido um extrator com a mesma arquitetura utilizada pelo itemizador, Java J2SE.

Este extrator busca pelas seqüências de dois substantivos a partir das etiquetas do *TreeTagger*. Assim, ele busca por: NN NN, NN NNS, NNS NN e NNS NNS. Há também uma preocupação para que não haja um outro substantivo antes ou depois desta seqüência, pois trabalharemos apenas com compostos formados por dois substantivos. Se tiver três substantivos consecutivos, o programa verifica isso e descarta.

Como saída, o programa nos oferece uma lista com possíveis compostos nominais que inclui o seu número de ocorrências no *corpus*, que na figura 3 aparece como freqüência.

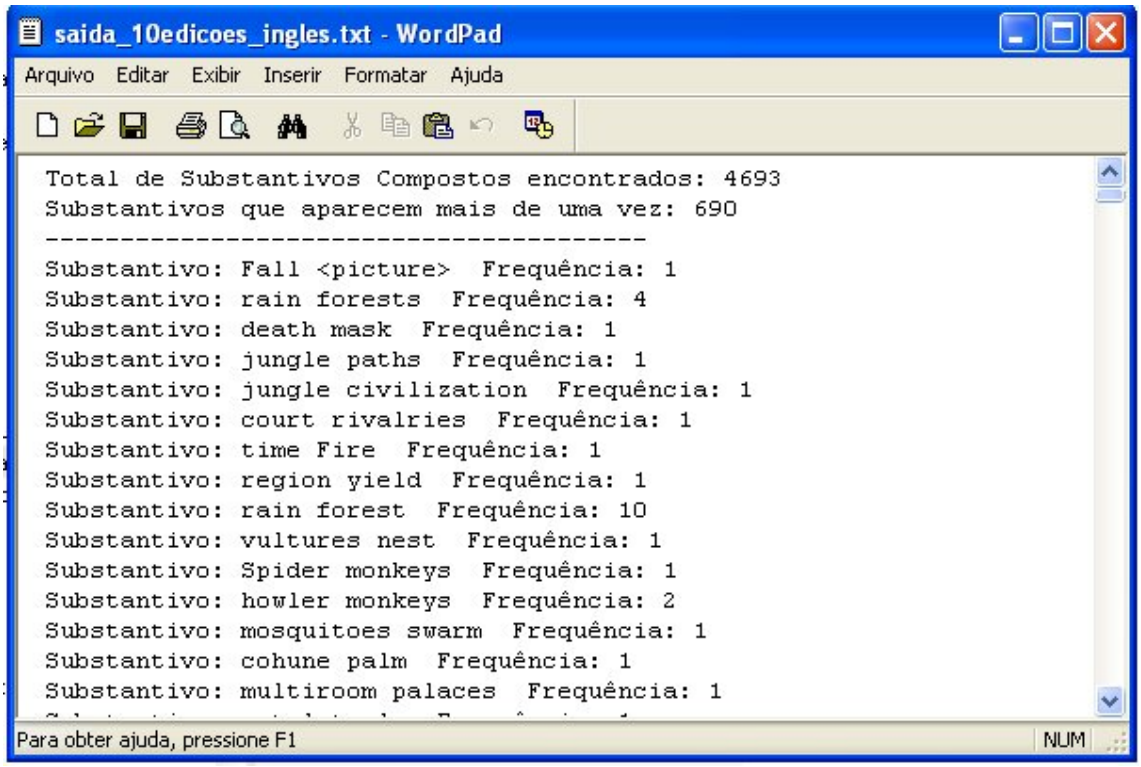

Figura 3. Resultados do extrator

3.4. Alinhador *Vanilla Aligner*

O alinhador *Vanilla Aligner* foi escolhido por ser gratuito e estar disponível on-line, além do mais, a sua precisão é alta. A função do *Vanilla* é alinhar frases de *corpora* bilíngües, ou seja, identificar as frases correspondentes em uma língua e na outra em conjunto de textos disponíveis em múltiplas línguas.

De acordo com os dados dos desenvolvedores do programa, Danielsson e Ridings (1997), para se trabalhar com o texto no alinhador *Vanilla*, é necessário um pré-processamento, que consiste em anotar as sentenças e os parágrafos. Entende-se por sentenças as unidades que um tradutor traduziria de uma só vez, incluindo, desta forma, títulos ou frases dentro de um parágrafo. Além disso, o texto deve ser itemizado. Para este alinhador, cada unidade de uma língua, uma frase, por exemplo, corresponde a uma

unidade de tamanho semelhante na outra língua. Entende-se por tamanho, a quantidade de caracteres.

Esse modelo probabilístico tem obtido bons resultados, atingindo uma média de 4% de erro em um *corpus* trilíngüe formado por 15 artigos de economia do *Union Bank of Switzerland*, nas línguas inglês, francês e alemão.

Conforme sugerido no *site* do LAEL[4], onde o programa é disponibilizado, optamos em preparar o *corpus* utilizando as etiquetas .EOS para final de sentença e .EOP para final de parágrafo.

A anotação das sentenças e dos parágrafos foi feita através do recurso localizar e substituir do bloco de notas. Toda vez que encontramos um ponto (. ! ou ?) substituímos por . .EOS, quando final de frase. Logo após a anotação foi conferida manualmente e a anotação dos parágrafos foi feita totalmente de forma manual. Poderíamos utilizar algum sistema automático para a identificação das sentenças e dos parágrafos. Não descartamos essa possibilidade no futuro, mas não há ainda uma opção de qual ferramenta seria a mais adequada para a tarefa. Como para esse alinhamento inicial utilizamos somente uma edição da revista, a conferência manual não tomou muito tempo.

Quanto ao alinhamento, este é feito em duas etapas. Primeiro os parágrafos são alinhados, e só após as frases de um parágrafo são alinhadas. Para o programa funcionar, os textos devem possuir o mesmo número de parágrafos. Isso dificulta um pouco a tarefa, já que a diferença de quantidade de parágrafos entre uma versão e outra era bem grande inicialmente. O texto em inglês possuía 286 parágrafos, enquanto que em português, havia 255. Foi necessária uma adaptação manual, de forma que as duas versões possuíssem o mesmo número de parágrafos, em torno de 260.

Utilizando o alinhador *Vanilla* do *site* do LAEL, obtém-se como resultado uma lista com as relações entre as unidades (denominadas *links*), que podem ser do tipo 1-1, em que uma frase foi alinhada com outra da tradução, 2-1, duas frases da primeira língua são alinhadas como apenas uma na segunda língua. Entre as outras possibilidades, temos: 0-1, 1-0, 1-2 e 2-2.

Em relação aos erros cometidos pelo programa, Gale e Church (1993) afirmam que há uma maior precisão nos casos de resultados 1-1, em que uma frase da primeira língua corresponde a apenas uma na segunda. Os casos mais problemáticos são os de 1-0, em que a frase em uma língua não possui correspondente na outra. Nos exemplos apresentados pelos autores, todas as relações classificadas como 1-0 estavam erradas de acordo com a avaliação de seres humanos. Entre as conclusões desse artigo, temos o fato de que quanto mais semelhantes forem as línguas, com maior freqüência ocorrerão casos de 1-1, logo o programa será mais eficiente.

A partir de uma análise inicial, vimos que de fato o caso mais freqüente é 1-1 e os alinhamentos são feitos corretamente (ver figura 4). O *Vanilla* traz alguns erros de alinhamento, mas devido ao seu modelo estatístico, mesmo alinhando frases

[4] Disponível em: <http://www2.lael.pucsp.br/corpora/alinhador/index.html>

inadequadamente em um ponto do texto, isso não compromete o alinhamento das seguintes.

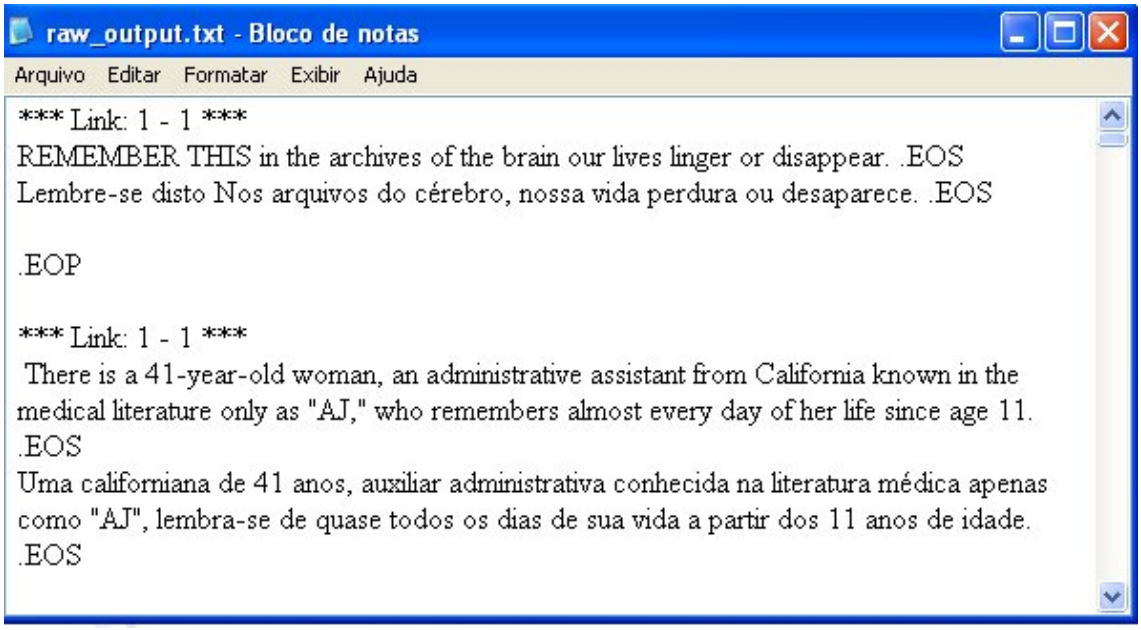

Figura 4. *Corpus* alinhado

4. Análise dos dados

A fim de avaliar de que forma a metodologia adotada pode ser útil no estudo da semântica dos compostos nominais e tendo em vista também os seus equivalentes de tradução em português, propomos um exercício de análise parcial dos dados.

Partimos dos resultados obtidos através do extrator de seqüências de dois substantivos que ocorrem juntos no *corpus* formado pelas edições da revista *National Geographic*. Avaliamos os vinte primeiros resultados obtidos de forma que possamos prever se essa metodologia é adequada para o estudo dos casos encontrados no *corpus*. Na tabela 2, há uma lista dos vinte primeiros candidatos a compostos seguidos pela quantidade de vezes em que os mesmos aparecem no *corpus.*

Nº	Seqüência	Ocorrências	Nº	Seqüência	Ocorrências
1	Fall <picture>	1	11	Spider monkeys	1
2	rain forests	4	12	howler monkeys	2
3	death mask	1	13	mosquitoes swarm	1
4	jungle paths	1	14	cohune palm	1
5	jungle civilization	1	15	multiroom palaces	1
6	court rivalries	1	16	metal tools	1
7	time Fire	1	17	kings-the kuhul	1
8	region yield	1	18	theater states	1
9	rain forest	10	19	temple mounds	2
10	vultures nest	1	20	limestone altars	1

Tabela 2. Vinte primeiros resultados do extrator

Quando havia uma figura no contexto original, uma indicação foi incluída no *corpus*, por isso aparece <picture> na versão dos dados em inglês. Desta forma, a primeira seqüência foi desconsiderada. A primeira questão que chama a atenção ao considerar os dados da tabela 2 é o fato de que a maioria dos supostos compostos (80%) ocorre apenas uma vez no *corpus* inteiro. Esses casos são chamados de *hapax* e não apresentam um desafio para o estudo da semântica dos compostos, pois não pretendemos estudar casos lexicais específicos, mas buscamos organizar os compostos em uma tipologia, e a partir dessa organização identificar alguns padrões.

Após termos as seqüências NN, precisamos conferir se esses itens são realmente compostos. A única forma de fazer isto é analisando o contexto em que cada item ocorre. Para obter estes dados, utilizamos o programa *WordSmith Tools*, versão 5, que, entre outros recursos, oferece concordâncias de uma palavra de busca. Por concordâncias, entendemos uma lista com todas as frases em que a palavra de busca ocorre no *corpus* (ver figura 5).

N	Concordance
1	is critical. Savanna-woodland chimps, unlike their rain forest brethren, spend most of their waking time
2	the border in western Mali. Unlike their better-known rain forest kin, savanna-woodland chimps spend
3	fish, and huge flightless birds could be found in the rain forest, virtually tame since they had never seen a
4	2008 Philippine Eagles: How to Help In the rain forests of Mindanao island the high-pitch cries of
5	people are waking up to its plight. Avian king of the rain forest canopy, the Philippine eagle is
6	only home, so it became, by default, the king of the rain forest. Expanding into an empty ecological
7	world's wonder. It glides through its sole habitat, the rain forests of the Philippines, powerful wings spread
8	their homeland is one of the last chunks of pristine rain forest left in the Congo Basin. Even so, nearby
9	earth, often by surface-mining that imperils pristine rain forests. Currently, less than 20 percent of
10	times or for trade—could not be sustained in the rain forest. Instead, each city-state produced small
11	shed their own metal armor in the sweltering rain forest in favor of these Maya "flak jackets." The
12	the centuries, as the Maya learned to prosper in the rain forest, the settlements grew into city-states, and
13	first Maya arrived, in perhaps 1000 B.C.—a dense rain forest where scarlet macaws, toucans, and
14	of the Maya unfolded against the backdrop of the rain forests of southern Mexico and Central America.

Figura 5. Concordância de *rain forest/rain forests*

Como o extrator busca por substantivos no singular ou no plural, *rain forest* e *rain forests* aparecem separadamente. Podemos juntar os resultados e afirmar que esta é a seqüência mais freqüente neste primeiro estudo, totalizando 14 ocorrências. O foco do estudo é em compostos formados por apenas dois substantivos, mas o extrator acabou incluindo duas ocorrências em que *rain forest* faz parte de um grupo maior, como em *rain forest brethren* (irmãos das florestas tropicais), *rain forest kin* (parentes das matas tropicais) e *rain forest canopy* (copa da floresta pluvial). Mesmo com estes problemas, não há como negar que *rain forest* constitui um caso de composto nominal, pois o seu referente é um só. Analisando o aspecto sintático adotado por Adams e Bloomfield (*apud* Ryder, 1994), ele também é aprovado como composto, pois não há como incluir outra palavra entre os dois elementos, e se utilizamos um adjetivo, este modifica o composto como um todo. Outra característica interessante sobre este composto é que ele é freqüente na revista de forma geral e não aparece em apenas uma reportagem. As ocorrências foram encontradas em cinco edições diferentes. Na tabela 3, vemos quais as seqüências obtidas pelo extrator foram confirmadas como compostos nominais a partir

da análise das concordâncias. Através da análise do *corpus* paralelo, também é possível identificar os equivalentes de tradução para o português.

Nº	Compostos NN	Equivalentes de tradução
1	rain forest	floresta tropical, floresta pluvial, mata tropical
2	death mask	máscara mortuária
3	jungle paths	caminhos através da floresta
4	jungle civilization	civilização se estendia pelas florestas
5	court rivalries	rivalidades violentas
6	spider monkeys[5]	SEM TRADUÇÃO
7	howler monkeys	macacos, SEM TRADUÇÃO
8	mosquitoes swarm[6]	SEM TRADUÇÃO
9	multiroom palaces	palácios
10	metal tools	ferramentas de metal
11	theater states[7]	SEM TRADUÇÃO
12	temple mounds	montes de templos, templos
13	limestone altars	altares de calcário

Tabela 3. Compostos e seus equivalentes de tradução do *corpus* paralelo

As seqüências da tabela 2 que foram descartadas após uma maior análise são as que fazem parte de um grupo com mais de dois substantivos (*cohune palm nuts*) ou que o anotador cometeu algum erro. Em *time Fire*, o segundo elemento é um substantivo próprio, que se refere ao maia Fogo Novo, mas a sua etiqueta é de substantivo simples. Em outros casos, o segundo elemento é um verbo, mas o etiquetador reconheceu como um substantivo (*region yield, vultures nest*[8]).

Com a análise dos vinte primeiros resultados, percebemos que algumas palavras se repetem em diferentes seqüências, tais como *monkey* e *jungle*. Ryder (1994) chama essas palavras de *core words*, que podem ocupar tanto a posição do primeiro quanto do segundo substantivo. Como esta palavra pode ser encontrada em outros compostos, temos *família de compostos*. Há outros *core words* nestes compostos que são encontrados em outras seqüências ao longo do *corpus*: *court case*, *distric court*, *basketball court*, *tennis court*.

5. Considerações finais

A metodologia adotada para a identificação e extração dos compostos do *corpus* formado pelas edições em língua inglesa da revista *National Geographic* trouxe bons resultados, já que dos vinte candidatos a compostos, treze foram confirmados como tal.

[5] Tanto *spider monkeys* como *howler mokeys* são tipos de macacos, macaco-aranha e bugio ou macoco-uivador. As traduções apresentadas em nota de rodapé foram feitas pelos autores deste trabalho.

[6] Tradução: enxame de mosquitos.

[7] Tradução: Estados de teatro.

[8] Tradução dos verbos *yield* e *nest* respectivamente: proporcionar e aninhar-se.

Houve alguns problemas em relação ao anotador automático que cometeu alguns erros de etiquetação e o extrator ainda precisa ser aprimorado de forma que seqüências com três substantivos não sejam identificadas.

A próxima etapa deste estudo se ocupará do alinhamento do *corpus* inteiro. Precisamos ainda desenvolver algum recurso para anotar os finais de frases e de parágrafos automaticamente e uma forma mais prática de exibir o *corpus* alinhado, de maneira que facilite a identificação dos equivalentes de tradução.

6. Referências

BERBER SARDINHA, Antonio Paulo. Lingüística de *corpus*: Histórico e problemática. **DELTA,** São Paulo, v. 16, n. 2, p. 323-367, 2000.

DANIELSSON, Pernilla, RIDINGS, Daniel. **Practical Presentation of a 'Vanilla' Aligner**. Presentation held at the TELRI Workshop in alignment and exploitation of texts in Ljubljana, 1997.

FRANKENBERG-GARCIA, Ana, SANTOS, Diana. COMPARA, um corpus paralelo de português e inglês na Web. In: TAGNIN, Stella E. O. (Org.). **Cadernos de Tradução: Corpora e Tradução.** Florianópolis: NUT, 2002, v. 1, n. 9, p. 61-79. Disponível em: <http://www.cadernos.ufsc.br/online/9/ana.htm>

GALE, William A., CHURCH, Kenneth W. A program for aligning sentences in bilingual corpora. **Computational Linguistics**, n. 19, v. 1, p. 75-102, 1993.

MCENERY, Tony, WILSON, Andrew. **Corpora and Translation: Uses and Future Prospects.** Lancaster, Unit for Computer Research on the English Language (UCREL), 1993. Relatório Técnico. Disponível em: <http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf>

RYDER, Mary Ellen. **Ordered Chaos: The Interpretation of English Noun-Noun Compounds.** Berkeley: University of California Press, 1994.

SANTORINI, Beatrice. **Part-of-Speech Tagging Guidelines for the Penn Treebank Project.** Pennsylvania, Departament of Computer & Information Science, 1990. Relatório Técnico.

SCHMID, Helmut. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing (NeMLaP-1), 1994, London. **Proceedings.** London: USL Press, 1994, p. 44–49.