Identificação de Expressões Multipalavra em Domínios Específicos

Aline Villavicencio[1,2], Carlos Ramisch[1,3], André Machado[1],
Helena de Medeiros Caseli[4], Maria José Finatto[5]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (Brasil)
[2]Department of Computer Sciences, Bath University (Inglaterra)
[3]GETALP – Laboratoire d'Informatique de Grenoble, Université de Grenoble (França)
[4]Departmento de Ciência da Computação, Universidade Federal de São Carlos (Brasil)
[5]Instituto de Letras, Universidade Federal do Rio Grande do Sul (Brasil)
{avillavicencio,ceramisch,ammachado}@inf.ufrgs.br,
helenacaseli@dc.ufscar.br, mfinatto@terra.com.br

Resumo

Expressões Multipalavra (EM) são um dos grandes obstáculos para a obtenção de sistemas mais precisos de Processamento de Linguagem Natural (PLN). A cobertura limitada de EM em recursos linguísticos pode impactar negativamente o desempenho de tarefas e aplicações de PLN e pode levar à perda de informação ou a problemas de comunicação, especialmente em domínios técnicos, em que EM são particularmente frequentes. Este trabalho investiga algumas abordagens para a identificação de EM em corpora técnicos com base em medidas de associação, informações morfossintáticas e de alinhamento lexical. Primeiramente, examina-se a influência de alguns fatores sobre o seu desempenho, tais como fontes de informação para a identificação e avaliação. Se, por um lado, as medidas de associação enfatizam revocação, por outro, o método de alinhamento centra-se em precisão. Neste trabalho, propõe-se uma abordagem combinada que une os pontos fortes das diferentes abordagens e fontes de informação utilizando um algoritmo de aprendizado de máquina para produzir resultados mais robustos e precisos. A avaliação automática dos resultados mostra que o desempenho do método combinado é superior aos resultados individuais das abordagens associativa e baseada em alinhamento para a extração de EM de português e inglês. Além disso, é discutida a efetividade de cada um desses métodos para a identificação de EM específicas em comparação com EM de domínio genérico. O método proposto pode ser usado para auxiliar o trabalho lexicográfico, fornecendo uma lista de candidatos a EM.

1 Introdução

A cobertura dos recursos lexicais tem um impacto significativo sobre o desempenho de muitas tarefas e aplicações de Processamento de Linguagem Natural (PLN), e nesse sentido, muitas pesquisas têm se dedicado à proposição de métodos para automatizar a aquisição lexical. Nos últimos anos, alguns desses trabalhos têm se centrado em um conjunto de fenômenos para os quais recursos lexicais são particularmente carentes de cobertura, entre os quais destacam-se as *Expressões Multipalavra* (EM) (Baldwin, 2005; Villavicencio et al., 2007).

Essas expressões podem ser definidas como combinações de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas, pragmáticas ou estatísticas (Sag et al., 2002), e incluem, entre outros fenômenos, verbos frasais (*carry up, consist of*), verbos de suporte (*tomar um banho, dar uma caminhada*), compostos (*carro de polícia, bode expiatório*) e expressões idiomáticas (*engolir o sapo, nadar contra a corrente*). EM são muito numerosas dentro de uma língua e, segundo Biber et al. (1999), podem corresponder de 30% a 45% do inglês falado e 21% da linguagem acadêmica. De acordo com Jackendoff (1997), as EM têm a mesma ordem de magnitude, no léxico de um falante nativo, do número de palavras simples. No entanto, essas proporções são provavelmente subestimadas se considerarmos a linguagem de um domínio específico na qual: (i) o vocabulário especializado e a terminologia especializada vão ser compostos, na sua maior parte, por EM (*aquecimento global, sequenciamento de proteínas, litíase renal crônica*) e (ii) que novas EM estão sendo constantemente introduzidas na linguagem (*melhoramento genético, gripe suína).*

Os problemas causados pela cobertura limitada dos recursos lexicais podem ser ilustrados, por exemplo, no contexto de um analisador sintático. Em uma amostra aleatória de 20.000 sentenças do *British National Corpus* (Burnard, 2007), a baixa cobertura de EM no léxico utilizado resultou em 8% dos erros cometidos pelo analisador sintático (Baldwin et al., 2004), mesmo com uma gramática de ampla cober-

tura como a *English Resource Grammar* (Copestake e Flickinger, 2000).

Portanto, EM devem ser identificadas e tratadas adequadamente, pois, do contrário, a qualidade dos sistemas pode ser seriamente deteriorada, especialmente para tarefas de PLN que envolvam algum tipo de processamento semântico (Sag et al., 2002). Para tanto, acredita-se que métodos (semi-)automáticos robustos para a aquisição de informações lexicais sobre EM possam aumentar a cobertura dos recursos lexicais. Por exemplo, o número de construções verbo-partícula listadas em um dicionário, como o *Alvey Natural Language Tools* (Carroll e Grover, 1989), pode ser significativamente aumentado através da adição de construções verbo-partícula automaticamente extraídas de um corpus, como o *British National Corpus* (Baldwin, 2005).

Neste trabalho, são investigadas algumas abordagens para a identificação de EM a partir de *corpora técnicos*. Uma avaliação detalhada do desempenho destas abordagens é realizada, examinando-se o impacto das fontes de informação utilizadas. A mesma inclui uma comparação dos resultados obtidos para um domínio específico usando um corpus paralelo inglês–português (en–pt) composto por artigos científicos de uma revista brasileira bilíngue de Pediatria. O propóstito é verificar de que forma uma segunda língua pode fornecer pistas relevantes para a identificação de EM em português. São também discutidos alguns aspectos que influenciam uma avaliação mais profunda dos resultados, tais como a proporção de termos específicos e genéricos nas listas de referência, a filtragem dos candidatos e o número de palavras de cada *n*-grama.

Após avaliar as abordagens associativa e baseada em alinhamento separadamente em trabalhos anteriores (Caseli et al., 2009a; Villavicencio, Caseli e Machado, 2009), neste trabalho investiga-se sua combinação ponderada a fim de propor-se um método mais robusto que resulte em um conjunto mais preciso de EM candidatas do que as dos métodos individuais. A abordagem proposta pode ser utilizada para: a) auxiliar o trabalho de produção de dicionários especializados, quer sejam repertórios de termos ou de fraseologismos, fornecendo uma lista de EM candidatas para manter os recursos lexicais atualizados; e, b) também para a melhoria da qualidade dos sistemas de PLN, que poderiam vir a integrar listas de EM verificadas manualmente ou listas de candidatas a EM extraídas de forma totalmente automática.

O restante deste artigo está estruturado da seguinte forma. A seção 2 apresenta uma visão geral sobre EM e sobre alguns trabalhos relacionados que tratam da sua extração automática. A seção 3 descreve os recursos utilizados nos experimentos, enquanto a seção 4 descreve os métodos propostos para extrair EM. A seção 5 apresenta a metodologia de avaliação e de análise dos resultados. A seção 6 encerra este artigo com as conclusões e com algumas perspectivas de trabalhos futuros.

2 *Expressões Multipalavra: Problemas e Soluções para Identificação*

O termo Expressão Multipalavra vem sendo utilizado para descrever um grande número de construções distintas, mas fortemente relacionadas, tais como verbos de suporte (*fazer uma demonstração, dar uma palestra*), compostos nominais (*quartel general*), frases institucionalizadas (*pão e manteiga*), e muitos outros. Sag et al. (2002) definem EM como *interpretações idiossincráticas que cruzam os limites (ou espaços) entre as palavras*. Esses autores tratam da diferença que existe entre a interpretação de uma EM (por exemplo, *bode expiatório*) como um todo e os significados isolados das palavras individuais que a compõem (*bode* e *expiatório*). Os mesmos consideram que a definição de EM engloba um grande número de construções, tais como expressões fixas, compostos nominais e construções verbo-partícula. Ainda nessa linha, Calzolari et al. (2002) definem EM como *uma sequência de palavras que atua como uma única unidade, em algum nível de análise linguística*, a qual exibe algumas das seguintes características:

- transparência sintática e/ou semântica reduzida;
- composicionalidade reduzida;
- flexibilidade sintática reduzida;
- violação de regras sintáticas gerais;
- elevado grau de lexicalização;
- elevado grau de convencionalidade.

Para Moon (1998) *não há um fenômeno unificado que se possa descrever como EM, mas sim um complexo de atributos que interagem de formas diversas, muitas vezes desordenadas, e que representam um amplo contínuo entre o não-composicional (ou idiomático) e grupos composicionais de palavras*. Outros autores utilizam a noção de frequência e definem EM como sequências ou grupos de palavras que co-ocorrem com mais frequência do que seria esperado por acaso, e podem ultrapassar fronteiras sintagmáticas (Evert e Krenn, 2005). Isso incluiria também fórmulas de saudação como *Tudo bem? Como você vai?* e sequências lexicais, como *eu não sei se*. Santos (2008) aborda a questão de EM em relação a uma aplicação em particular, a tradução automática, e os desafios causados por expressões multipalavra ou expressões complexas, que envolvem tanto casos de traduções de uma palavra em muitas

(exemplo *miss* como *sentir a falta*), de muitas palavras traduzidas em uma (exemplo *get up early* como *madrugar*) e de sequências de palavras traduzidas como sequências também (exemplo *kick the bucket* como *bater as botas*). Por conseguinte, autores diferem nas definições que usam para EM em função dos aspectos particulares que estão sendo enfatizados e dos grupos de palavras e construções que consideram como EM.

As EM são muito frequentes na linguagem corrente e isso se reflete em várias gramáticas e recursos lexicais existentes, em que quase metade das entradas são dedicadas a EM. No entanto, devido às suas características heterogêneas, EM apresentam grandes desafios tanto sob o ponto de vista linguístico quanto computacional (Sag et al., 2002). Primeiramente, algumas EM são fixas, e não apresentam variação interna, como *ad hoc*, enquanto outras permitem diferentes graus de variabilidade interna e modificação, como *levar chumbo/ferro/pau* e *ir/descer/perder-se (por) água abaixo*. Em termos de semântica, algumas EM são mais opacas em seu significado como *lavar roupa suja* significando *discutir assunto particular geralmente conflituoso*, enquanto outras são mais transparentes, e seus significados podem ser inferidos a partir de seus componentes, tal como *carro de polícia*, em que o sintagma preposicional *de polícia* adiciona informação de função para a palavra *carro*.

No contexto de textos de um domínio específico, tem-se uma definição importante relacionada a EM, a de termo. Segundo Krieger e Finatto (2004), para especialistas de um domínio, termos são uma representação do conhecimento da área específica, ou seja, as terminologias contêm unidades lexicais que expressam conceitos abstratos ou mesmo elementos concretos de um domínio. Existem várias diferenças entre EM genéricas e termos. Primeiramente, termos podem ser compostos por uma única palavra ou por múltiplas, como locuções nominais, enquanto EM são inerentemente compostas por duas ou mais palavras. Em segundo lugar, EM são um fenômeno que integra tanto à linguagem técnica e científica quanto à linguagem cotidiana de propósito geral, enquanto termos são tipicamente relacionados com a primeira. Além, disso, é preciso considerar que uma mesma terminologia, quando ocorre simultaneamente em textos de linguagem cotidiana e em textos científicos, tende a adquirir trações semânticos mais e menos específicos conforme o tipo de comunicação envolvida. Essas são diferenças importantes, pois será necessário determinar até que ponto os métodos computacionais disponíveis para lidar com EM em textos genéricos podem ser aplicados para lidar com corpora de domínio especializados e vice-versa. Por outro lado, EM e termos têm também aspectos comuns: ambos têm idiosincrasia semântica e ambos são um desafio para os sistemas de PLN (Ramisch, 2009).

Uma classificação de EM que permite agrupá-las em classes de dificuldade para métodos de identificação automáticos, é a proposta por Sag et al. (2002). Eles classificam as EM divididas em dois grandes grupos: expressões institucionalizadas e expressões lexicalizadas. Expressões institucionalizadas se caracterizam por serem sintaticamente e semanticamente composicionais, mas estatisticamente idiossincráticas se comparadas a qualquer outra alternativa do mesmo conceito (*café forte* x ?*café possante*).[1] Dentre essas EM convencionalizadas, ou seja, observadas com uma frequência muito maior do que qualquer outra formulação equivalente, as mais representativas são as colocações (*sal e pimenta*, *bagagem emocional*, etc.). De acordo com Smadja (1993), as colocações podem variar muito quanto ao seu comprimento, porém, elas geralmente contêm uma média de duas a cinco palavras. Além disso, algumas vezes a colocação envolve palavras não adjacentes em uma frase, e nesses casos a distância entre as partes que a compõem depende da sintaxe da língua, sendo potencialmente tão longa quanto se queira. Essa característica acarreta em dificuldades para a identificação de EM, à medida que não se pode saber *a priori* quais os limites das EM a serem extraídas automaticamente de textos. Algumas características relevantes das colocações são:

- não composicionalidade: o seu significado é constituído pela composição dos significados de suas partes, somando-se a isto um componente semântico adicional não previsível a partir das palavras isoladas que a compõem.
- não substituibilidade: não é possível substituir cada uma das suas componentes por palavras que possuam o mesmo significado que estas (*sal e pimenta* x ?*sal e malagueta*).
- não modificação: existem restrições a quanto às possibilidades de modificação sintática de uma colocação, que variam em grau de rigidez de expressão para expressão (?*bagagem vermelha emocional*).

As expressões lexicalizadas, por outro lado, compreendem EM que *possuem pelo menos sintaxe ou semântica parcialmente idiossincráticas, ou contêm palavras que não ocorrem isoladamente*, ou seja, são expressões que apresentam certa rigidez formal. Como tipos de expressões lexicalizadas têm-se as expressões fixas, expressões semi-fixas e expressões sintaticamente flexíveis.

[1] A notação usada neste artigo marca sentenças não-gramaticais com "*"e sentenças gramaticais possíveis mas não usuais para um falante nativo com "?".

- As *expressões fixas*, tais como *ad hoc*, *Porto Alegre*[2] são consideradas as mais rígidas de todas e se caracterizam por não apresentarem variações morfossintáticas e não permitirem modificações internas.

- *Expressões semi-fixas*, diferentemente das expressões fixas, permitem certo nível de variação lexical. Esta variação pode ser referente à flexão, à forma reflexiva e à escolha de determinantes. Dentre estas, tem-se as expressões idiomáticas não decomponíveis, que permitem variações apenas quanto à flexão e quanto à forma reflexiva mas que não apresentam variabilidade sintática, tais como modificações internas ou até mesmo a transformação para a voz passiva. Um exemplo é a expressão em inglês *kick the bucket*, que apesar de permitir a conjugação do verbo *kick* (*kicked the bucket*), não admite modificações feitas internamente (**kick the big bucket*). Outros casos são os compostos nominais que não permitem variabilidade sintática e são caracterizados por permitir flexões de número (*wine glass* (taça de vinho), *orange juice* (suco de laranja) e *guarda-chuva*) e os nomes próprios que são altamente idiossincráticas do ponto de vista sintático.

- As *expressões sintaticamente flexíveis*, ao contrário das expressões semi-fixas, apresentam uma variabilidade sintática mais ampla. EM desse tipo incluem construções verbo-partícula do inglês (*look up* e *break up*), expressões idiomáticas decomponíveis (*ser barra pesada* e *a barra pesou*), verbos de suporte (*tomar um banho, dar uma caminhada*, etc). Os componentes de EM desse tipo podem estar separados uns dos outros por aceitarem constituintes variáveis ou devido a variação na ordem sintática causada por fenômenos como passivização, topicalização, entre outros. Por exemplo, em alguns verbos frasais do inglês, o verbo pode estar separado da partícula por complementos de tamanhos não previsíveis, como *eat up* em *eat up the delicious and very expensive Belgian chocolate* x ?*eat the delicious and very expensive Belgian chocolate up*. De acordo com Riehemann (2001), este grau de flexibilidade varia de expressão para expressão e é geralmente imprevisível. Por exemplo, *spill the beans* and *kick the bucket* são duas expressões idiomáticas formadas por verbo transitivos e sintagma nominais mas que têm comportamentos bem diversos em termos de flexibilidade, com a primeira sendo sintaticamente flexível e a segunda semi-fixa.

[2]Cabe aqui salientar que, embora *Porto Alegre* seja um nome próprio e que esse tipo de EM possa receber tratamentos específicos, neste trabalho traz-se um enfoque propositalmente mais geral de diferentes tipos de EM. Além disso, a inclusão de nomes próprios como EM é aceita por alguns autores, e o critério adotado neste trabalho irá considerar também nomes próprios.

Em termos da identificação de EM, o grau de dificuldade da tarefa aumenta com o grau de flexibilidade da expressão. Consequentemente, muitos dos métodos tendem a se concentrar em capturar expressões fixas e semi-fixas, em particular, como discutido a seguir, pois essas quase não aceitam modificação na sintaxe e ocorrem sob a forma de palavras adjacentes. O desafio está em decidir os seus limites, e se há elementos variáveis, como determinantes alternativos (por exemplo, *engolir [um/o] sapo*). Desta forma, métodos baseados em *n*-gramas contíguos podem ser empregados para a sua identificação com bons resultados. Porém para as expressões flexíveis há ainda a dificuldade adicional de que a ordem dos seus componentes pode variar de diversas maneiras, e eles podem estar separados por um número imprevisível de palavras. Para este tipo de EM, a abordagem para identificação deve ser capaz de reconhecer combinações de palavras recorrentes mesmo se a ordem das mesmas muda, e se há elementos opcionais ou variáveis. Para lidar com esses casos, neste artigo é investigada a utilização do método baseado em alinhamento.

Neste trabalho, adota-se a definição de EM como combinações de palavras que apresentam idiossincrasias lexicais, sintáticas, semânticas ou estatísticas, que inclui entre outras construções expressões idiomáticas, verbos de suporte, compostos nominais, e nomes próprios, seguindo Sag et al. (2002). Esta definição abrangente é compatível com o uso de medidas estatísticas para a identificação de EM, pois elas são independentes de tipo. Desta forma, para se restringir a extração de EM a um tipo particular de expressões, filtros morfosintáticos podem ser aplicados. De fato, neste trabalho, tais filtros são empregados dando-se ênfase a expressões nominais, dada a natureza dos recursos disponíveis para a avaliação dos métodos, como dicionários e glossários terminológicos. Porém, os métodos aqui apresentados podem teoricamente ser aplicados para qualquer EM englobada por esta definição.

2.1 Identificação de EM

Uma grande variedade de abordagens têm sido propostas para a identificação automática de EM em função de seus diferentes tipos e propósitos de identificação. As abordagens diferem teórica e metodologicamente entre si em função dos tipos de EM abrangidos, da língua a que se aplicam e das fontes de informação que utilizam.

Alguns desses trabalhos utilizam informações sobre uma língua, como Baldwin (2005) e Villavicencio et al. (2007) aplicadas para o inglês, (Silva et

al., 1999) e (Dias e Nunes, 2001) aplicadas para o português. Outros trabalhos se beneficiam ainda de informações de uma segunda língua para ajudar a identificar e a lidar com EM (Villada Moirón e Tiedemann, 2006; Caseli et al., 2009b). Como base para ajudar a determinar se uma dada sequência de palavras é realmente uma EM (por exemplo, *ad hoc* é uma EM porém *o menino pequeno* não), algumas dessas propostas empregam conhecimentos linguísticos, enquanto outras empregam métodos ditos fracos ou estatísticos (por exemplo, Evert e Krenn (2005) e Villavicencio et al. (2007)) ou combinam vários tipos de informações tanto linguísticas, como propriedades sintáticas e semânticas (Van de Cruys e Villada Moirón, 2007), quanto de frequência e estatísticas, resultantes de processos como por exemplo o alinhamento lexical automático em um par de línguas (Villada Moirón e Tiedemann, 2006). A combinação de diversos tipos de informação pode ser realizada através de classificadores aprendidos automaticamente a partir de conjuntos de dados anotados (Pecina, 2008). Este trabalho investiga a influência de diferentes fontes de informação na tarefa de identificação de EM.

Medidas estatísticas de associação têm sido amplamente empregadas na identificação de EM, visto que elas podem ser democraticamente aplicadas a qualquer tipo de EM e de língua. A idéia por trás de seu uso é que elas são um meio de baixo custo para a detecção de padrões recorrentes, dado que se espera que as palavras componentes de uma EM coocorram frequentemente. Dessa forma, essas medidas podem indicar a probabilidade de que um candidato seja uma EM verdadeira, independentemente do tipo de EM e da língua. No entanto, algumas medidas parecem fornecer previsões mais precisas que outras sobre a chance de um determinado candidato ser de fato uma EM, e não há ainda consenso sobre qual medida é mais adequada para identificar EM em geral. Uma comparação de algumas destas medidas para a detecção de EM independentes de tipo indicaram que informação mútua diferencia melhor EM de não-EM do que χ_2 (Villavicencio et al., 2007). Várias medidas comuns de associação, como a informação mútua e χ_2, têm sido amplamente usadas para a detecção de EM, além de outras que têm sido especialmente propostas para esta tarefa, por exemplo, por Silva et al. (1999).

Outra questão importante é a generalização de algumas destas medidas para aplicação a *n*-gramas com tamanho arbitrário, principalmente no que diz respeito às MA baseadas em tabela de contingência e sua conhecida aplicação a bigramas. Silva et al. (1999), por exemplo, propõem uma abordagem em que, para um dado candidato, todas as possíveis divisões dele em duas partes são geradas, onde cada uma das duas partes pode ser maior que um unigrama. Assim um *n*-grama formado por 4 palavras $(w_1 \dots w_4)$ gera 3 bigramas: $(w_1) + (w_2, w_3, w_4)$, $(w_1, w_2) + (w_3, w_4)$ e $(w_1, w_2, w_3) + ((w_4)$, que são analisados em termos da força de associação entre as suas partes.

Além disso, para a identificação de EM, a eficácia de uma determinada medida parece depender de fatores como o tipo de EM sendo identificada, o domínio e o tamanho dos corpora utilizados, e a quantidade de dados de baixa frequência excluídos através da adoção de um limiar (Evert e Krenn, 2005). No que se refere aos tipos de corpora utilizados, tanto têm sido utilizados corpora paralelos (Maia, 2003), que envolvem originais e traduções, quanto corpora comparáveis (Maia e Matos, 2008), que implicam pares de textos escritos sobre um mesmo tema ou tópico originalmente produzidos em línguas diferentes por pessoas diferentes. Esses corpora geralmente são tratados à medida que recebem algum tipo de etiquetamento ou anotação. Há, todavia, também estudos que se dedicaram a corpora monolíngues não tratados, tal como o de Dias e Lopes (2005).

Quanto a trabalhos que envolveram a extração de terminologias em corpora, pode-se dizer que têm sido muitos e diferentes os estudos publicados. Todos, entretanto, enfrentaram a dificuldade de distinguir, com apoio computacional, os limites entre o léxico especializado e o léxico da linguagem cotidiana. Ranchhod e Mota (1998), por exemplo, fizeram um estudo que justamente procurou qualificar a identificação de itens especializados em um analisador de texto integrado por uma ferramenta que pesquisa, arrola e traz informações sobre os termos nele contidos. O tratamento da informação, entretanto, não partiu de um bloco geral de EM de corpora previamente reunidos, mas consistiu em agregar ao sistema de busca informações trazidas de dicionários gerais e de dicionários específicos pré-existentes. Nesse trabalho, ainda que os textos a analisar tenham sido do tipo técnico, também foi enfrentado o problema da presença simultânea de uma mesma dada expressão em diferentes dicionários, fato que já reforçava o problema da distinção controversa entre o léxico comum e o léxico especializado.

Assim, considerada a dificuldade implicada nessa diferenciação, são investigadas aqui algumas abordagens para a identificação de EM de um domínio especializado e alguns aspectos que podem ter influência sobre os resultados obtidos, para uma avaliação mais precisa destes métodos. Para português, a combinação de algumas medidas baseadas em frequências e heurísticas para a identificação de termos para a construção de uma ontologia a partir de textos de domínio específico resultou em uma medida F de até 11,51% para bigramas e 8,41% para trigramas (Vieira et al., 2009).

Entre os métodos que utilizam informações adicionais para extrair EM, o proposto por Villada Moirón e Tiedemann (2006) parece ser o mais semelhante à abordagem baseada em alinhamento empregada neste trabalho. A principal diferença entre eles é a maneira com que o alinhamento lexical é usado no processo de extração de EM. Neste trabalho, o alinhamento de palavras é a base do processo de extração de EM, enquanto o método de Villada Moirón e Tiedemann usa o alinhamento apenas para a classificação dos candidatos a EM que foram extraídos com base em medidas de associação e em heurísticas de dependência de núcleo (em dados sintaticamente analisados). Outro trabalho relacionado reporta a detecção automática de compostos não-composicionais (Melamed, 1997) que são identificados através da análise de modelos estatísticos de tradução treinados com um corpus enorme em um processo demorado. Santos e Simões realizaram experimentos envolvendo alinhamento lexical em corpora paralelos (Santos e Simões, 2008), buscando, entre outros objetivos, mensurar a importância da combinação de dicionários e corpora, do uso de informações sintáticas neste processo e da direção de tradução entre os idiomas. Nesse sentido, pode-se considerar que metodologias para a extração de sintagmas nominais bilíngues a partir de corpora paralelos, por exemplo, através do uso da *Pattern Description Language* (Simões e Almeida, 2008), constituem em si formas de identificação de EM.

3 O Corpus e as Listas de Referência

Nos experimentos descritos a seguir, utilizou-se um conjunto de 283 artigos de Pediatria, o corpus JPED-Coulthard. Trata-se de um corpus paralelo português–inglês que contém 785.488 palavras em português e 729.923 palavras em inglês. A língua-fonte, isto é, a língua na qual os artigos foram originalmente escritos, é o português, enquanto a língua-alvo é o inglês. Os textos foram publicados no *Jornal de Pediatria* entre 2003 e 2004. Vale destacar que esse corpus foi inicialmente organizado e estudado quanto à adequação das traduções por Coulthard (2005). Não foram considerados os resumos/*abstracts* e as referências bibliográficas no cômputo do número de palavras dos textos.

A partir desses 283 artigos bilíngues, foram criados alguns recursos lexicais: o *Dicionário de Pediatria* e o *Catálogo de Pediatria*[3]. Ambos recursos contêm um levantamento de expressões conceitual e linguisticamente importantes do domínio. Esses dois recursos, entretanto, são produtos diferenciados, pois são voltados para o uso do aprendiz de tradução brasileiro que começa a trabalhar na área médica. Sua finalidade é auxiliar esses iniciantes, graduandos em tradução da área de Letras/Linguística, que têm pouca experiência com construções, noções e terminologias desse domínio. Enquanto o *Dicionário* apresenta expressões recorrentes nesse corpus que geralmente contêm algum termo de Medicina cuja definição é apresentada, o *Catálogo* apresenta um levantamento de construções frequentemente empregadas na linguagem do domínio em foco, representada pelo corpus, mas que não estão associadas a uma terminologia ou nódulo conceitual. Em síntese, o foco de um é para expressões associadas a definições e o de outro é dirigido para expressões com exemplos de uso e padrões, o que inclui colocações e fraseologias.[4] Assim:

- o *Dicionário de Pediatria* contém 747 itens em português e 746 itens em inglês
- o *Catálogo de Pediatria* possui 702 itens em português e 698 itens em inglês

O processo de seleção das entradas a serem adicionadas aos dois recursos passou pelas seguintes etapas:

1. geração de *n*-gramas (bi, tri e quadrigramas[5]);
2. seleção de *n*-gramas com frequência maior ou igual a 5, sendo de palavras técnicas ou não;
3. exclusão de *n*-gramas puramente gramaticais (*o leite*);
4. exclusão de *n*-gramas que contivessem preposições, pronomes, conjunções: (por exemplo, *n*-grama iniciado ou terminado por *de* [PREP]). Essa lista foi definida a partir da análise manual da lista de palavras mais frequentes nos *n*-gramas;
5. exclusão de *n*-gramas do tipo ([DET]+N+X[+Y]), ou seja, uma sequência de determinante (DET) seguido de um nome (N) e até dois elementos (X e Y) (por exemplo, *o leite o leite materno/ o leite materno*

[3] Produzidos e disponibilizados gratuitamente por TEXTQUIM/TERMISUL `http://www.ufrgs.br/textquim`

[4] Aqui cabe esclarecer que a divisão dos itens entre *Dicionário* e *Catálogo*, em sendo um julgamento que reproduz a distinção entre: a) o que é específico do domínio, associado a uma definição; b) o que é expressão da linguagem cotidiana; c) o que é uma expressão de natureza híbrida, associado a um padrão sintagmático e que inclua linguagem geral e especializada, tornou-se algo extremamente complexo. Nesse sentido, a divisão dos dados nesses três blocos, acomodados os dois últimos no *Catálogo*, foi feita por um grupo de pesquisadores com formação em Letras e Tradução que se ocupam de produtos dicionarísticos. A divisão, além de espelhar critérios objetivos, também reflete uma percepção subjetiva do fenômeno envolvido e sempre comportará críticas. O trabalho relatado neste artigo, que reúne pesquisadores de PLN e terminógrafos, pode justamente qualificar esses tipos de repertórios de EM à medida que o retoma e confronta o procedimentos que os geraram.

[5] Um esclarecimento sobre essa nomenclatura pode ser encontrado em Manning e Schütze (1999, p. 193).

ordenhado). O padrão DET foi preenchido com várias possibilidades de determinantes: *os/o/as/a/um/uma/alguma/cuja(o)*;

6. remoção de *n*-gramas começados ou terminadas por verbo; e
7. retirada de *n*-gramas que fossem subpalavras de *n*-gramas maiores.

Em resumo, os recursos apresentam como entradas apenas *n*-gramas do corpus com frequência superior a 5, os quais foram filtrados mediante o uso de informações morfossintáticas e manualmente verificados. O processo de filtragem gerou um total de 2.407 entradas. Desses itens, 1.421 são bigramas e 730 são trigramas. Ao se comparar as listas em português e em inglês, tanto no dicionário quanto no catálogo, há diferença na quantidade de bi, tri e quadrigramas de língua para língua. Isso ocorre porque nem todas as construções em português possuem equivalentes com construções idênticas em inglês, com o mesmo número e a mesma ordem de palavras. Por exemplo, *recém-nascidos de baixo peso* é um quadrigrama; no entanto, seu correspondente, *low birth weight*, é um trigrama.

Neste trabalho, para avaliar o reconhecimento de expressões candidatas a EM, foram utilizados para integrar as listas de referência tanto os *n*-gramas do *Dicionário* quanto os do *Catálogo de Pediatria*. Além disso, as candidatas a EM em inglês são avaliadas usando-se um dicionário geral de EM em inglês (Cambridge, 1994), que contém 24.160 entradas (dos quais 9.174 são bigramas e 2.946 trigramas). As listas de referência contem as EM selecionadas por lexicógrafos e terminólogos para cada uma das línguas. Para o português as listas de referência incluem candidatos com frequência maior ou igual a 5. Portanto, qualquer candidato identificado pelos métodos usados (em particular pelo método de alinhamento), que não atinja a frequência mínima de 5 ocorrências, não será considerado como verdadeiro positivo por não se encontrar nas referências, mesmo quando se tratar de uma EM.

A lista resultante do processo de enriquecimento foi produzida conforme descrito em Lopes et al. (2009)[6], adicionando-se todos os bigramas válidos contidos em trigramas da lista que haviam sido removidos durante a construção dos recursos. Esse processo foi feito para as listas de ambas as línguas, e as versões finais das listas de referência têm 2.150 *n*-gramas em português e 1.424 *n*-gramas em inglês.

Para verificar a proporção de EM genéricas e específicas de domínio que ocorreram no corpus, utilizou-se metodologias distintas para cada uma das línguas. Em português, além das informações sobre EMs nas listas de referência, usou-se também julgamentos humanos, para anotar as EM no que se refere à pertinência ou não de cada item ao domínio de Pediatria/Medicina. Desta forma cada lista foi anotada com informação de domínio, de comum acordo, por três pesquisadores de Terminologia e Tradução que estiveram envolvidos na produção do dicionário e do catálogo. O anotador humano seguiu as heurísticas listadas abaixo para decidir sobre cada EM.

E se a EM corresponde a um termo de Medicina ou Pediatria ou área afim, recebeu a etiqueta E (*teste tuberculínico, fase de indução*);

G se a EM corresponde a uma expressão da linguagem cotidiana, de fácil compreensão para qualquer falante medianamente escolarizado do português do Brasil, recebeu a etiqueta G (*falta de apetite, grupo de risco*);

H se a EM corresponde a uma expressão da linguagem cotidiana, mas com sentido específico em Pediatria/Medicina, sendo um híbrido entre linguagem cotidiana e linguagem especializada, constituindo casos de julgamento ambíguo, recebeu a etiqueta H (*nível de sódio, saturação de oxigênio*).

Em inglês a natureza das listas de referência foi utilizada como indicador de domínio, dada a indisponibilidade de anotações dos dados por falantes nativos. Desta forma considerou-se todas as entradas do dicionário e do catálogo como construções específicas ao domínio enquanto as construções genéricas provêm de um dicionário geral, o *Cambridge International Dictionary of Idioms* (Cambridge, 1994), com 1.270 *n*-gramas para o inglês com ao menos uma ocorrência no corpus. Essas duas fontes estão marcados na tabela 1 como especializada (E) e genérica (G), respectivamente.

A lista de referência em português anotada por domínio contém uma grande maioria (76,83%) de EM específicas de domínio. Dentre os 1.421 bigramas, 977 foram considerados específicos do domínio de Pediatria, 226 genéricos e 218 híbridos. No grupo de trigramas, há 730 trigramas, dos quais 419 específicos e 195 genéricos e 116 híbridos.

Entre os candidatos, há expressões recorrentes como *prevalence of elevate blood pressure*, que não serão extraídas por nenhum dos métodos, dado o foco em bigramas e trigramas. Consequentemente, nas avaliações reportadas, a revocação apresentada é subestimada em relação ao seu valor real.

4 Métodos

Neste trabalho, investiga-se o uso de duas abordagens independentes para identificação de EM, e propõe-

[6]Disponível em `www.inf.pucrs.br/~ontolp/downloads-ontolplista.php`

Tipo	português	inglês
Específico	1.396	1.424
Genérico	421	1.270
Híbrido	334	–
Total	2.151	2.694

Tabela 1: Número de EM nas referências.

se uma abordagem combinada. A primeira abordagem, doravante denominada *abordagem associativa*, aplica Medidas de Associação (MA) para todos os bigramas e trigramas gerados a partir de cada corpus. As candidatas a EM são avaliadas em termos dos valores obtidos para elas por cada uma das medidas de associação utilizadas.

A segunda abordagem, doravante denominada *abordagem baseada em alinhamento*, tem por princípio a extração de EM a partir dos alinhamentos automáticos lexicais de versões em português e em inglês do Corpus de Pediatria gerados pelo alinhador estatístico lexical `GIZA++` (Och e Ney, 2000b). O método combinado proposto, por sua vez, combina as duas abordagens usando redes bayesianas.

Nesta seção, são descritos os experimentos realizados usando-se cada uma das abordagens para extrair EM do corpus. A avaliação é realizada de maneira automática, comparando-se as EM identificadas pela abordagem com as listas de referência descritas na seção 3 para cada uma dos línguas, em termos da precisão (P), revocação (R) e medida F, calculadas, respectivamente, como:

$$P = \frac{(\text{\#candidatas corretas})}{(\text{\#candidatas propostas})}$$
$$R = \frac{(\text{\#candidatas corretas})}{(\text{\#candidatas na referência})}$$
$$F = \frac{(2 \times P \times R)}{(P + R)}$$

Primeiramente, uma lista prévia de candidatas a EM é extraída do corpus JPED-Coulthard com cada abordagem. Em seguida, as candidatas são analisadas morfosintaticamente pelas ferramentas do `Apertium`[7] (analisador morfossintático e desambiguador lexical) com base nos dicionários morfológicos originais aumentados conforme descrito em Caseli, Nunes e Forcada (2006) e em Caseli (2007). Vale ressaltar, aqui, que o desempenho do analisador morfossintático está relacionado à cobertura do mesmo, ou seja, utilizando-se os dicionários morfológicos aumentados citados a cobertura é de 1.136.536 formas superficiais em português e de 61.601, em inglês. Com relação ao desempenho do desambiguador lexical, não foi encontrado nenhum relato a esse respeito. Por fim, sobre as listas de candidatas propostas por cada um dos métodos, são aplicados os seguintes filtros:

- **f0** lista original, nenhum filtro é aplicado às candidatas;
- **f1** candidatas após a remoção de n-gramas contendo pontuação e números;
- **f2** candidatas após (a) **f1** e (b) cujo número de ocorrências no corpus[8] é no mínimo 5;
- **f3** candidatas após (a) **f1, f2** excluindo ainda aquelas que (b) se iniciam por uma palavra funcional[9] e algumas formas superficiais, como flexões do verbo *ser* (*são*, *é*, *era*, *eram*), pronomes relativos (*qual*, *quando*, *quem*, *por que*) e preposições (*para*, *de*)[10].

Para f3, padrões alternativos de filtros morfossintáticos têm sido propostos na literatura, como o aplicado para a construção das listas de referência, e podem ser otimizados com informações específicas para cada língua. No entanto, como neste trabalho o objetivo é investigar o desempenho de um conjunto de métodos, f3 foi definido seguindo a proposta de Caseli et al. (2009b) para o inglês com padrões equivalentes para o português. Além disto, os filtros foram aplicados independentemente para cada uma das línguas, gerando uma lista filtrada de candidatos para cada uma.

Uma vez obtida a lista filtrada de candidatas a EM, o objetivo é remover apenas e qualquer n-grama que não seja uma EM. Este processo difere para cada uma das abordagens, conforme descrito nas próximas seções, e o sucesso é avaliado de acordo com as EM contidas nas listas de referência. Isso significa que os n-gramas candidatos que sejam encontrados nas listas de referência são considerados EM, mas os não contidos não necessariamente são não-EM. Há limitações de cobertura das referências a se considerar, dada a natureza dinâmica das línguas, bem como questões das características de uma EM qualquer, como transparência e frequência, entre outras.

[7] `Apertium` (Armentano-Oller et al., 2006) é um sistema de tradução automática de código-fonte aberto disponível em `http://www.apertium.org`.

[8] O método de alinhamento considera as frequências de alinhamento de uma candidata (número de vezes em que as palavras da candidata foram alinhadas juntas). No entanto, o filtro **f1** é aplicado sobre o número de ocorrências dessa candidata no corpus independentemente dos alinhamentos.

[9] Nesse trabalho, considera-se que uma palavra funcional é um artigo, verbo auxiliar, pronome, advérbio ou conjunção.

[10] E analogamente para o inglês, considerando-se a tradução literal dos termos de filtragem.

4.1 Abordagem Associativa

Na abordagem associativa, a filtragem é feita com base na força de associação de uma candidata medida de acordo com a probabilidade de co-ocorrência das palavras que a compõem. Evidências estatísticas de associação forte têm sido bastante empregadas em trabalhos recentes da área (Evert e Krenn, 2005; Ramisch et al., 2008; Pearce, 2002; Pecina, 2008; Ramisch, 2009). Uma visão geral destes trabalhos é apresentada em Ramisch (2009).

A validação de uma EM candidata é feita utilizando-se um conjunto de medidas de associação (MA): informação mútua pontual (PMI, do inglês *pointwise mutual information*), informação mútua (MI, do inglês *mutual information*), estatística *t* de student, estatística χ^2 de Pearson, coeficiente de Dice, teste exato de Fisher, medida de Poisson-Stirling (PS) e razão de chances (OR, do inglês *odds ratio*), implementadas no Ngram Statistics Package (Banerjee e Pedersen, 2003). Estas medidas típicas de associação são resumidas na tabela 2 (adaptada de Ramisch (2009)) e as fórmulas são calculadas com base nas frequências obtidas no corpus de cada língua. Globalmente, as medidas assumem que a frequência de co-ocorrência das palavras em uma EM é superior à frequência esperada para uma sequência randômica de *n* palavras. A segunda coluna da tabela mostra quais os valores de *n* (ou seja, o comprimento do *n*-grama) para os quais a MA pode ser aplicada, onde "*" representa a ausência de limitação de tamanho.

Formalmente, considera-se a candidata a EM como um *n*-grama composto de *n* palavras adjacentes w_1 a w_n. A contagem do número de ocorrências (frequência) de um *n*-grama em um corpus é denotada $c(w_1 \dots w_n)$. A medida da força de associação entre as palavras do *n*-grama $w_1 \dots w_n$ é feita através da comparação da frequência relativa *observada* $c(w_1 \dots w_n)$ com a frequência relativa *esperada* E. A última é calculada supondo-se como hipótese nula que palavras em um corpus são eventos independentes, ou seja, que a frequência de um *n*-grama é igual ao número de palavras N do corpus ponderado pelo produto das probabilidades de cada uma das palavras que o compõem:

$$E(w_1 \dots w_n) = \frac{c(w_1) \dots c(w_n)}{N^{n-1}}$$

Algumas das MA usadas na identificação associativa de EM são baseadas em tabelas de contingência. Isto significa que, além de considerar as frequências individuais das palavras, elas também levam em consideração a frequência de não-ocorrência dessas palavras, construindo uma tabela com as combinações possíveis. Nesses casos, usa-se a_i para representar ambas possibilidades, w_i e $\bar{w}_i$, em que a notação $\bar{w}_1$ corresponde à ocorrência de qualquer palavra exceto w_i. Em um dado *n*-grama $w_1 \dots w_n$, cada célula da tabela de contingência corresponde a uma combinação possível de $a_1 \dots a_n$. Essas medidas são muito robustas para eventos raros e são particularmente adequadas para os *n*-gramas onde $n = 2$ mas não são facilmente estendidas para candidatos com comprimento arbitrário, como mostra a coluna intermediária da tabela 2. As três últimas MA apresentadas na tabela são baseadas em tabelas de contingência e possuem limitação do valor de *n*. Portanto, para todos os trigramas candidatos, os valores dessas medidas não puderam ser calculados.

4.2 Abordagem Baseada em Alinhamento Lexical

Nos últimos anos, a utilização de textos paralelos e textos paralelos alinhados tem se tornado cada vez mais frequente em inúmeras aplicações de PLN. Os textos paralelos, segundo a terminologia estabelecida pela comunidade de linguística computacional, são textos acompanhados de sua tradução em uma ou várias línguas. Se esses textos possuírem marcas que identificam os pontos de correspondência entre o texto original (texto fonte) e sua tradução (texto alvo) eles são considerados alinhados.

Métodos automáticos de alinhamento de textos paralelos podem ser usados para encontrar os pontos de correspondências entre os textos fonte e alvo. O processo automático de alinhamento de textos paralelos, resumidamente, pode ser entendido como a "busca", no texto alvo, de uma ou mais sentenças (ou unidades lexicais) que correspondam à tradução de uma dada sentença (ou unidade lexical) no texto fonte. Quando a correspondência se dá entre sentenças dizemos que o alinhamento é sentencial, quando a mesma ocorre entre unidades lexicais, dizemos que o alinhamento é lexical.

O corpus paralelo utilizado nos experimentos apresentados neste artigo passou por ambos os processos de alinhamento. O alinhamento sentencial foi realizado automaticamente por uma versão do *Translation Corpus Aligner* (TCA) (Hofland, 1996), descrita em detalhes em Caseli (2003) e Caseli, Silva e Nunes (2004). Após o processamento automático, os casos potencialmente alinhados de maneira incorreta (alinhamentos diferentes de 1 : 1) foram verificados manualmente. O alinhamento lexical, por sua vez, foi desempenhado automaticamente por meio da ferramenta `GIZA++`[11] (Och e Ney, 2000b), porém sem a verificação manual uma vez que esta seria uma tarefa extremamente árdua já que os alinhamentos diferentes de 1 : 1 são muito mais frequentes no alinhamento

[11] `http://www.fjoch.com/GIZA++.html`

Medida	n	**Fórmula**
PMI	*	$\log_2 \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)}$
t	*	$\frac{c(w_1 \dots w_n) - E(w_1 \dots w_n)}{\sqrt{c(w_1 \dots w_n)}}$
Dice	*	$\frac{n \times c(w_1 \dots w_n)}{\sum_{i=1}^{n} c(w_i)}$
MI	2,3	$\sum_{a_1 \dots a_n} \frac{c(a_1 \dots a_n)}{N} \log_2 \left[\frac{c(a_1 \dots a_n)}{E(a_1 \dots a_n)} \right]$
PS	2,3	$c(w_1 \dots w_n) \times \left[\log \frac{c(w_1 \dots w_n)}{E(w_1 \dots w_n)} - 1 \right]$
χ^2	2	$\sum_{a_1 a_2} \frac{[c(a_1 a_2) - E(a_1 a_2)]^2}{E(a_1 a_2)}$
Fisher	2	$\sum_{k=c(w_1 w_2)}^{\min\{c(w_1),c(w_2)\}} \frac{c(\bar{w}_1)!c(w_1)!c(\bar{w}_2)!c(w_2)!}{N!k!(c(w_1)-k)!(c(w_2)-k)!(c(\bar{w}_2)-c(w_1)+k)!}$
OR	2	$\frac{c(w_1 w_2)c(\bar{w}_1 \bar{w}_2)}{c(w_1 \bar{w}_2)c(\bar{w}_1 w_2)}$

Tabela 2: Medidas de associação utilizadas pelo método associativo

lexical do que no sentencial.

GIZA++ utiliza os modelos estatísticos da IBM (Brown et al., 1993) e o modelo oculto de Markov (HMM) (Vogel, Ney e Tillmann, 1996; Och e Ney, 2000b; Och e Ney, 2000a) para determinar as melhores correspondências entre palavras fonte e palavras alvo. Para os experimentos apresentados neste artigo, utilizou-se a versão 2.0 em sua configuração padrão na qual estão incluídas iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM.

Os modelos utilizados por GIZA++ variam no modo como é calculada a probabilidade do alinhamento $\Pr({f_1}^S, {a_1}^S | {e_1}^T)$, na qual ${a_1}^S$ é um alinhamento que descreve o mapeamento da palavra fonte f_j na palavra alvo e_{a_j} considerando-se que ${f_1}^S$ é uma cadeia de caracteres fonte e ${e_1}^T$, uma cadeia de caracteres alvo. Por exemplo, no modelo IBM-1, todos os alinhamentos têm a mesma probabilidade. O modelo HMM, por sua vez, usa um modelo de primeira ordem $p(a_j|a_{j-1})$ no qual a posição do alinhamento a_j depende da posição do alinhamento anterior a_{j-1}. A partir do modelo IBM-3, um modelo de fertilidade $p(\phi|e)$ é adicionado ao cálculo da probabilidade. Esse modelo descreve o número de palavras ϕ alinhadas com a palavra alvo e. O modelo IBM-4, por sua vez, busca modelar o efeito de mudança de posição das palavras fonte na tradução e inclui, portanto, um modelo de distorção para simular o fato de que a tradução de uma palavra fonte é deslocada na frase alvo.

O alinhamento foi realizado por GIZA++ no sentido pt–en e no sentido en–pt e a combinação (união) dos alinhamentos foi gerada resultando no alinhamento final. A união foi selecionada como método de simetrização dos alinhamentos gerados nos dois sentidos de tradução por se tratar do método que apresentou melhor revocação em experimentos prévios (Caseli, 2007). O desempenho no alinhamento de lemas, configuração utilizada nos experimentos apresentados neste artigo, não foi avaliado especificamente para o corpus de Pediatria, porém em avaliação prévia realizada em outro corpus pt–en o desempenho relatado foi de 8,94% AER (*Alignment-Error Rate*), o que está de acordo com os valores relatados para outros pares de línguas. Detalhes sobre a avaliação do alinhamento de lemas produzido por GIZA++ podem ser obtidos em Caseli (2007).

Além dos alinhamentos lexical e sentencial, o corpus pt–en também foi etiquetado morfossintaticamente usando os dicionários morfológicos e as ferramentas do Apertium (Armentano-Oller et al., 2006). Em particular, o corpus foi analisado morfossintaticamente com base nos dicionários morfológicos originais aumentados conforme descrito em Caseli, Nunes e Forcada (2006) e em Caseli (2007). A partir desse processo de etiquetação morfossintática é que foi possível aplicar filtros de categorias gramaticais na lista inicial de candidatas a EM.

Um exemplo de um par de sentenças paralelas pt–en alinhado lexicalmente por GIZA++ é apresentado

na Figura 1. Nesse exemplo, cada palavra é apresentada em uma linha separada na ordem em que ocorrem na sentença, sua posição na sentença é indicada na primeira coluna e os alinhamentos lexicais podem ser recuperados pelo número que segue o ":" ao final de cada palavra. Alinhamentos de omissão estão representados pelo "0". Além disso, cada forma superficial da palavra nesta figura é seguida por seu lema, categoria gramatical e traços morfológicos retornados pelo etiquetador morfossintático que, quando não reconhece uma determinada palavra, indica que a mesma é desconhecida inserindo um "*" em seu início, como ocorre com as palavras em português *helicobacter* e *pylori*. Por fim, é possível notar um alinhamento envolvendo a candidata a EM "*precisa para*" com sua correspondente tradução em inglês "*needs to*".

	Sentença em português
1	o/o<det><def><m><sg>:1
2	único/único<adj><m><sg>:2
3	fato/fato<n><m><sg>:3
4	aceito/aceitar<vblex><pri><p1><sg>:3
5	é/ser<vbser><pri><p3><sg>:4
6	o/o<detnt>:0
7	de/de<pr>:0
8	que/que<cnjsub>:5
9	o/o<det><def><m><sg>:0
10	*helicobacter/helicobacter:6
11	*pylori/pylori:7
12	precisa/precisar<vblex><pri><p3><sg>:8_9
13	entrar/entrar<vblex><inf>:10
14	para/para<pr>:8_9
15	o/o<det><def><m><sg>:11
16	estômago/estômago<n><m><sg>:12
17	através/através<adv>:13
18	da/de<pr>+o<det><def><f><sg>:0
19	boca/boca<n><f><sg>:15

	Sentença em inglês
1	the/the<det><def><sp>:1
2	only/only<adj>:2
3	certainty/certainty<n><sg>:3_4
4	is/be<vbser><pri><p3><sg>:5
5	that/that<cnjsub>:8
6	*helicobacter/helicobacter:10
7	pylori/pylorus<n><pl>:11
8	needs/need<vblex><pri><p3><sg>:12_14
9	to/to<pr>:12_14
10	enter/enter<vblex><inf>:13
11	the/the<det><def><sp>:15
12	stomach/stomach<n><sg>:16
13	through/through<pr>:17
14	the/the<det><def><sp>:0
15	mouth/mouth<n><sg>:19

Figura 1: Exemplo de um par de sentenças paralelas alinhadas lexicalmente por GIZA++

Diferentemente da abordagem associativa, na abordagem baseada em alinhamento, as candidatas a EM são identificadas a partir das correspondências entre palavras e sequências de palavras da língua fonte e alvo definidas pelo alinhador. Mais especificamente, usando o alinhamento lexical entre uma sequência de palavras origem S ($S = s_1 \dots s_n$ com $n \geq 2$) e uma sequência de palavras destino T ($T = t_1 \dots t_m$ com $m \geq 1$), o método de extração baseado em alinhamento assume que a sequência S será uma candidata a EM. Por exemplo, a sequência de duas palavras em português *aleitamento materno* — que ocorre 202 vezes no corpus utilizado nos experimentos — é uma candidata a EM porque essas duas palavras foram alinhadas em conjunto 184 vezes com a palavra *breastfeeding* (um alinhamento 2 : 1), 8 vezes com a palavra *breastfed* (um alinhamento 2 : 1), 2 vezes com *breastfeeding practice* (um alinhamento 2 : 2) e assim por diante. É essa frequência de alinhamento, ou seja, o número de vezes em que a sequência de palavras da língua fonte ocorre em um alinhamento $n : m$ com $n \geq 2$, que será usada como atributo na combinação das abordagens. Por procurar sequências de palavras-origem que são frequentemente unidas durante o alinhamento, independentemente do número de palavras-alvo envolvidas, o método baseado em alinhamento prioriza precisão sobre revocação.

Algumas observações podem ser feitas a respeito de como o produto do alinhamento lexical influencia as candidatas de EM geradas. Por exemplo, na Figura 1, pode-se notar que duas palavras em português não consecutivas (*precisa* e *para*) foram alinhadas com duas palavras consecutivas do inglês (*needs to*). Essa característica traz um diferencial para o método de alinhamento quando comparado às medidas de associação uma vez que estas últimas recuperam apenas n-gramas e, sendo assim, a abordagem associativa nunca gera EM compostas por itens não consecutivos, diferentemente do método de alinhamento, que é capaz de gerá-las. Como consequência, a avaliação realizada com base nas listas de referência subestima a revocação do método baseado em alinhamento, uma vez que o processo de construção das listas levou em conta apenas sequências de palavras consecutivas.

4.3 Abordagem Combinada

Dado que as abordagens associativa e baseada em alinhamento têm características diferentes, que podem fazer com que se capture diferentes conjuntos de EM, a proposta deste trabalho é desenvolver um método combinado que maximize as vantagens de cada uma. Para isto, as diferentes MA e as frequências de alinhamento obtidas para as candidatas podem ser con-

n-grama (α)	n	$c(\alpha)$	**Abordagem alinhamento**	**Abordagem associativa**								**classe**
				Dice	OR	PMI	PS	t	MI	χ^2	Fisher	
abnormal findings	2	11	9	,03	114,1	6,74	25,70	2,62	0	734,73	0	não
renal insufficiency	2	26	0	,13	767,7	9,10	138	5,09	,0003	14249,6	0	sim
ato cirúrgico	2	7	3	,08	989,1	9,64	39,79	2,64	,0001	5584,2	0	sim
academia americana	2	24	0	,52	74302	13,3	197,4	4,9	,0004	244244	0	não

Tabela 3: Exemplos de entradas dos conjuntos de treinamento contendo todos os atributos usados em cada uma das estratégias de combinação.

sideradas como atributos para algoritmos de aprendizado de máquina, em uma abordagem semelhante à adotada por Pecina (2008) e Ramisch (2009). Para a abordagem combinada, foram utilizados os algoritmos implementados pelo pacote Weka (Witten e Frank, 2005).

O classificador para cada língua foi construído a partir do conjunto de *n*-gramas filtrados e anotados com os valores das medidas associativas e com o diagnóstico do alinhamento lexical sobre se o *n*-grama é uma possível EM, isto é, a frequência com que ele foi alinhado conjuntamente com uma palavra ou sequência de palavras na língua alvo. Na próxima seção, avalia-se duas possibilidades de combinação dos métodos: a primeira consiste em enriquecer os candidatos extraídos pelo método de alinhamento com as MA do método associativo; a segunda consiste em enriquecer os candidatos extraídos pelo método associativo (ou seja, todos os *n*-gramas do corpus que passaram pelos filtros) com a frequência de alinhamento. Em ambos os casos, os atributos usados para treinar o classificador são idênticos, e estão exaustivamente enumerados nas colunas da tabela 3.

Para adicionar a informação de classe para cada candidata foi feita uma avaliação das mesmas em relação as listas de referência: se o *n*-grama está contido nas listas, ele tem a classe *sim* (correspondendo a uma EM), caso contrário ele pertence à classe *não* (não-EM). A tabela 3 mostra alguns exemplos de entradas de inglês e português do conjunto de treinamento.

Como discutido na próxima seção, os conjuntos de dados disponíveis para treinar o classificador são desbalanceados, com uma proporção muito maior de não-EM do que de EM. Desta forma, optou-se por utilizar um algoritmo de construção de redes bayesianas com pesquisa de solução ótima através da árvore de cobertura. Este algoritmo, além de ser especialmente adequado para dados numéricos como os da tabela 3, tem se mostrado robusto e pouco sensível ao uso de classes com tamanhos muito diferentes.[12] Experimentos realizados em outros conjuntos de dados demonstraram que o algoritmo de máquina de vetor de suporte produz classificadores de boa qualidade. Neste trabalho, no entanto, optou-se por empregar um classificador do tipo rede bayesiana porque ele é menos oneroso em termos de recursos computacionais e de tempo de treinamento do que o algoritmo de máquina de vetor de suporte, além de produzir resultados comparáveis ao mesmo (Ramisch, 2009).

[12] Utilizando, por exemplo, árvores de decisão sobre os dados em inglês obteve-se um modelo com uma única classe com um mesmo diagnóstico para todos os candidatos (*não*).

5 Resultados

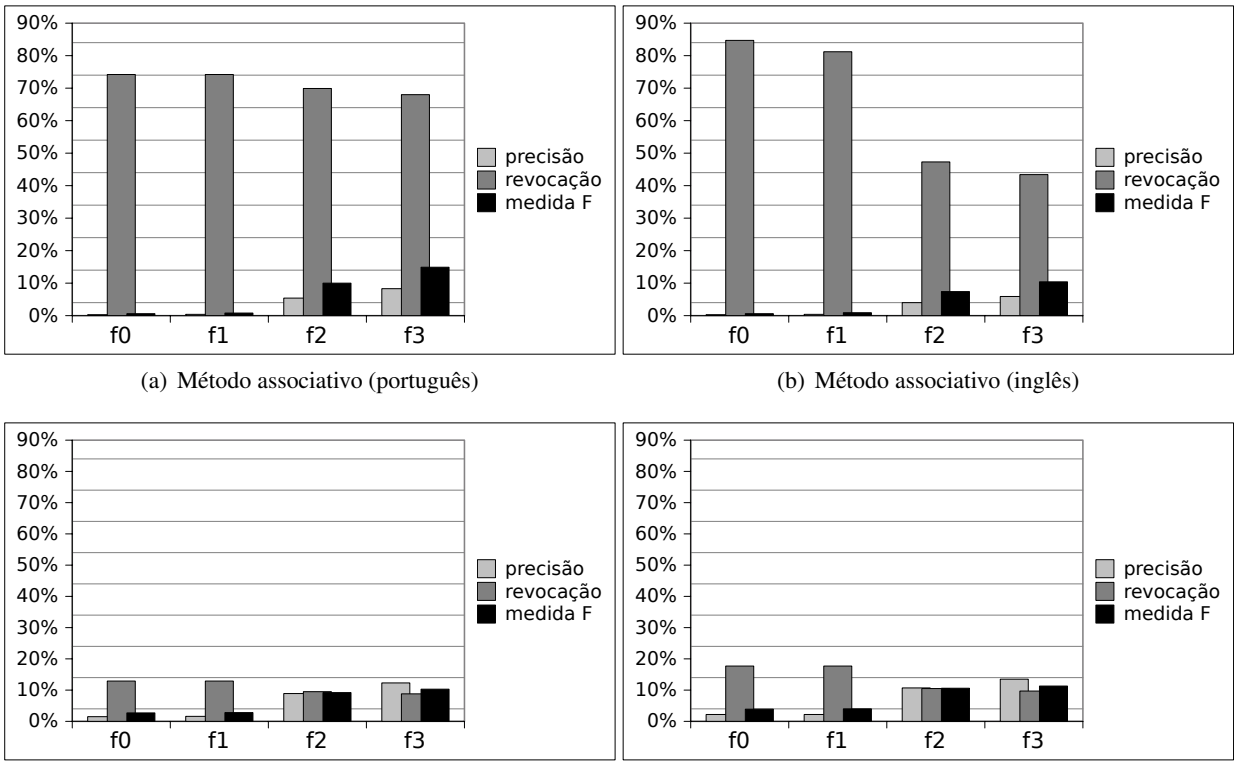
O desempenho obtido por cada um dos métodos na tarefa de identificação de EM será discutido a seguir. Após, discutir-se-á a taxa de acerto de cada método para EM de acordo com sua especificidade de domínio.

5.1 Identificação de EM

Primeiramente, descreve-se os resultados da avaliação da lista inicial de candidatas propostas por cada método e da aplicação dos vários filtros para remoção de ruído, como mostrado nas tabelas 4 e 5. Os resultados para português e inglês são descritos em termos de número de candidatos resultantes de cada processo e número de verdadeiros positivos (VP).

Para ambas as línguas e ambos os métodos a aplicação dos filtros melhorou os resultados em termos da precisão e da medida *F* (figura 2). Em particular o filtro f2 resultou em uma grande melhora da precisão, e mesmo nos casos onde houve uma redução na revocação, a medida *F* ainda refletiu a contribuição positiva do filtro. Por exemplo, para a abordagem associativa para o inglês, a revocação baixou em 33,9% mas ainda assim a medida *F* aumentou em 6,5%.

A diferença entre os métodos se refletiu em um número muito menor de candidatas a EM propostas pelo método baseado em alinhamento do que pelo método associativo: para o português 18.132 contra 572.893 respectivamente. Apesar desta grande diferença no número de candidatos propostos pelo alinhador (97% menos candidatos que a abordagem associativa), os resultados têm maior precisão para



(a) Método associativo (português)

(b) Método associativo (inglês)

(c) Método baseado em alinhamento (português)

(d) Método baseado em alinhamento (inglês)

Figura 2: Avaliação do efeito dos filtros aplicados a cada um dos métodos independentemente.

Abordagem associativa (pt)

	f0	**f1**	**f2**	**f3**
n-gramas	572.893	384.742	27.874	17.242
VPs	1.595	1.595	1.504	1.455

Abordagem baseada em alinhamento (pt)

	f0	**f1**	**f2**	**f3**
n-gramas	18.132	17.444	2.284	1.464
VP	277	277	204	189

Tabela 4: Desempenho dos filtros aplicados a ambos os métodos para os candidatos em *português*.

Abordagem associativa (en)

	f0	**f1**	**f2**	**f3**
n-gramas	586.431	391.850	25.478	15.399
VP	1.822	1.746	1.017	873

Abordagem baseada em alinhamento (en)

	f0	**f1**	**f2**	**f3**
n-gramas	17.516	16.972	2.108	1.527
VP	380	380	225	201

Tabela 5: Desempenho dos filtros aplicados a ambos os métodos para os candidatos em *inglês*.

ambas as línguas e maior medida F em todos os casos, exceto f2 e f3 para o português.

A tabela 6 mostra os resultados obtidos com f3 aplicado à intersecção entre os candidatos propostos por ambos os métodos, ou seja, considerando-se apenas candidatos extraídos simultaneamente pelo método baseado em alinhamento e pelo método associativo. Uma grande proporção dos candidatos propostos pelo alinhador estão contidos nos candidatos propostos pelas MA. Além disto, a precisão obtida melhora para ambas as línguas, com um número menor de candidatos que para o alinhador, mas o mesmo número de VP.

5.2 Combinação dos Métodos

A linha de base para a comparação do desempenho do método combinado é a obtida pelas abordagens associativa e baseada em alinhamento de forma independente.

São testadas duas alternativas diferentes para o método combinado, conforme mostrado na tabela 7. Em ambos os casos, foi realizada uma avaliação por validação cruzada fornecida pelo pacote Weka, usando-se 10 subconjuntos de dados. O tamanho do conjunto de dados fornecido é, respectivamente, de 1.464 e 17.242 candidatos para o português e de 1.527 e 15.399 para o inglês. A primeira estratégia,

	português	inglês
n-gramas	1.431	1.368
VP	190	208
Precisão	13,28%	15,20%
Revocação	8,83%	7,62%
Medida *F*	10,61%	10,16%

Tabela 6: Desempenho do filtro f3 aplicado à intersecção dos candidatos propostos pelas abordagens associativa e baseada em alinhamento.

alinhador → associativo, utiliza como base as candidatas propostas pelo método de alinhamento e usa os métodos associativos para subsequente validação. A segunda alternativa, **associativo → alinhador**, consiste em utilizar a lista de candidatas geradas pelas MA, adicionando às mesmas uma coluna que corresponde à informação fornecida pelo alinhador sobre a frequência de alinhamento da candidata. Vale relembrar que, em ambos os casos, os atributos usados pelo classificador são idênticos e foram descritos na tabela 3. Isso significa que as estratégias de combinação correspondem a duas maneiras diferentes de *escolher* quais candidatas serão consideradas pelo método combinado, mas *não têm influência no número ou tipo de atributos* usados pelo classificador. Por conseguinte, o conjunto de dados derivado da primeira estratégia contém algumas candidatas que não possuem nenhum valor de MA por se tratarem de *n*-gramas que não foram detectados pelo método associativo. Inversamente, o conjunto de dados derivado da segunda estratégia possui diversas candidatas contendo zero como informação de frequência de alinhamento, que correspondem a *n*-gramas identificados pelas MA mas que não fazem parte de nenhum alinhamento múltiplo.

A segunda alternativa parte de um número bem maior de candidatos do que a primeira, e para ambas as línguas tem-se um resultado muito superior em termos de medida *F* (mais de 30% superior para o português) do que o resultado obtido pela combinação dos métodos na direção contrária. Mesmo em relação ao desempenho máximo obtido por cada um dos métodos individuais, se pode ver uma melhora significativa nos resultados, em particular para o método associativo no português, onde a combinação resulta em uma aumento de quase 30% na medida *F*.

5.3 Especificidade dos Candidatos

Destes resultados gerais, o uso de uma lista de referência maior para o inglês, com EM genéricas, do que para o português, não parece ter contribuído para diferenças em resultado. Porém, a fim de determinar mais precisamente o desempenho de cada um dos métodos na identificação de termos específicos de domínio ou genéricos, os candidatos propostos por eles foram avaliados também em termos de tipo de EM, tabela 8. A maior proporção de candidatas específicas de domínio nas listas de referência foi também refletida nas candidatas VP retornadas por cada método. Para ambas as línguas, todas as abordagens têm melhor taxa de acerto para EM de domínio específico (E), com a identificação de um número maior destas candidatas. Estes resultados sugerem que a identificação de EM genéricas pode ser uma tarefa mais difícil do que a de EM específicas. O uso mais preciso e mais convencionalizado de EM dentro de um domínio pode contribuir para isto, com um menor grau de variabilidade léxica, sintática e semântica. Comparando-se as abordagens associativa e baseada em alinhamento, para EM específicas a abordagem que obteve uma melhor taxa de acerto foi a primeira, e para candidatas genéricas, foi a segunda. Isto pode ser devido à capacidade da abordagem baseada em alinhamento de identificar candidatos não-contíguos, sendo mais robusta à possível modificação ou variabilidade sintática. Porém, como as listas de referência possuem EM contíguas, não se pode calcular automaticamente qual o ganho trazido por esta capacidade. Investigações futuras serão feitas para avaliar este impacto. Além disso, a taxa de acerto para cada tipo obtida para o inglês é consideravelmente superior à obtida para o português.

6 Conclusões e Trabalhos Futuros

EM representam um conjunto complexo e heterogêneo de fenômenos que desafiam tentativas linguísticas e computacionais de capturá-los totalmente. Paradoxalmente, as EM têm um papel fundamental na comunicação oral e escrita e precisam ser levadas em conta quando da concepção de aplicações de processamento de linguagem que precisem de alguma interpretação semântica. Neste contexto, o tratamento de EM nos sistemas de PLN atuais é um grande desafio, dada a essência heterogênea e extremamente flexível dessas construções. Em decorrência do seu caráter simultaneamente complexo e essencial, as EM têm sido o foco de diversos trabalhos na comunidade científica, principalmente no que diz respeito à sua aquisição automática a partir de grandes bases textuais.

Neste trabalho, diferentemente de outros estudos com EM, lidou-se com um corpus especializado de originais e traduções e com listas de EM dele derivadas, as quais foram previamente identificadas por analisadores humanos como relevantes para a aprendizagem de tradução em Pediatria — tanto do ponto de vista conceitual quanto lingüístico — e incluídas em dois produtos de caráter dicionarístico diferentes. O conjunto geral dessas expressões, reunido em

português

	alinhador → associativo	**associativo → alinhador**
n-gramas	260	1.576
VP	137	787
Precisão	52,7%	49,9%
Revocação	6,4%	36,6%
Medida *F*	11,4%	42,2%

inglês

	alinhador → associativo	**associativo → alinhador**
n-gramas	97	1.130
VP	53	372
Precisão	54,6%	32,9%
Revocação	2,5%	17,3%
Medida *F*	4,7%	22,7%

Tabela 7: Desempenho do método combinado usando um classificador do tipo rede bayesiana.

português

	alinhamento	**associativa**	**alinhamento ∩ associativa**	**referências**
VP	190	1.463	190	2.151
E	55,79%	58,85%	55,79%	64,90%
G	24,21%	21,60%	24,21%	19,57%
H	20,00%	19,55%	20,00%	15,53%

inglês

	alinhamento	**associativa**	**alinhamento ∩ associativa**	**referências**
VP	208	934	208	2.694
E	63,46%	73,13%	63,46%	53,45%
G	36,54%	26,76%	36,54%	46,55%
H	0%	0%	0%	0%

Tabela 8: Proporção de EM por tipo em candidatos propostos por abordagens e na lista de referência

uma única lista de EM, com itens que integram tanto o léxico geral quanto o especializado, foi reavaliado pela mesma equipe e então dividido em três tipos: itens do léxico especializado, do léxico geral e itens de um "léxico híbrido", que constituiria, em tese, confluência entre linguagem cotidiana e linguagem especializada. O desafio aqui colocado foi o de encontrar metodologias de identificação para os itens associados ao léxico especializado combinando os diferentes fatores envolvidos nos materiais sob exame.

Nesse intuito, procurou-se investigar em que medida é possível utilizar e combinar recursos heterogêneos para automatizar a extração de EM, em específico no caso de textos técnicos em que grande parte das expressões possui simultaneamente um estatuto terminológico. Em primeiro lugar, analisou-se separadamente dois métodos de extração de EM: o método associativo, cuja lista de candidatos resultantes é gerada com base nas frequências de coocorrência das palavras que o formam; e o método baseado em alinhamentos, que por sua vez supõe que, em um corpus paralelo bilíngue, as expressões serão alinhadas de forma múltipla, extraindo-se assim a partir dos alinhamentos $n : m$ uma lista de candidatos a EM.

A fim de avaliar o desempenho individual e as possíveis estratégias de combinação de ambos os métodos, gerou-se uma lista de candidatos para cada

método e para cada língua a partir do corpus paralelo em português e em inglês do Jornal de Pediatria. Em um primeiro momento, foi investigado o impacto de diferentes fontes de informação na identificação de EM de domínios técnicos, através da aplicação de filtros sobre essas listas de candidatos. Os três filtros testados se mostraram bastante eficazes na remoção de ruídos e resultaram em melhoras significativas na medida F. Dentre esses, o que apresentou melhor compromisso entre um aumento na precisão e uma queda na revocação foi o filtro de frequência (**f2**); porém, o filtro morfossintático (**f3**) foi uma maneira simples e eficaz de eliminar o ruído com poucos efeitos colaterais.

Em termos das abordagens utilizadas (associativa e de alinhamento), uma avaliação dos desempenhos individuais indicou a natureza complementar de cada uma delas: a primeira identifica um maior número de candidatas, porém a segunda propõe um conjunto mais focado de candidatas com maior precisão. Ambos os métodos demonstraram maior sucesso na identificação de EM específicas de domínio, associadas ao léxico especializado, o que sugere que as EM genéricas apresentam um maior desafio para estes métodos. Está prevista uma investigação mais detalhada dos fatores que podem estar causando isso, como flexibilidade de uso e frequência, com comparação dos resultados obtidos em corpora genéricos. Pretende-se também verificar a portabilidade dos métodos para outros domínios.

Comparando as abordagens individuais, a associativa teve uma maior taxa de acerto nas EM específicas. Porém para as candidatas genéricas, a abordagem de alinhamento teve maior taxa de acerto. Dadas as diferenças dos candidatos propostos pelas duas abordagens, a combinação delas, proposta neste trabalho, trouxe um aumento significativo de desempenho na tarefa de identificação de EM. Foram avaliados dois modos para combinação dos resultados, e o que apresentou melhor desempenho foi que adotou o enriquecimento dos candidatos propostos pelos métodos associativos com informação de alinhamento. Neste caso a medida F aumentou de 14% para 42%.

Métodos como os apresentados neste artigo podem acelerar significativamente o trabalho de produção de repertórios de expressões recorrentes em corpora de textos científicos. Os resultados obtidos mostram que a adoção de abordagens simples, de baixo custo computacional e de conhecimento, pode trazer melhoras consideráveis de desempenho.

Para trabalhos futuros está prevista a investigação de maneiras alternativas para se obter a combinação ponderada das abordagens associativa e baseada em alinhamento, para produzir um conjunto de EM candidatas que é ainda mais precisa do que a fornecida pela primeira, mas que tem mais cobertura que a segunda. Além disso, seguindo a tendência de alguns trabalhos da área que exploram a extração de conhecimento de corpus comparável ao invés de corpus paralelo, como Fung (1998) e Haghighi et al. (2008), pretende-se, também, avaliar como as técnicas apresentadas neste artigo se comportam na extração de EM a partir de textos comparáveis. Por fim, a utilização dos resultados obtidos por este trabalho na construção semi-automática de ontologias também será investigada.

Agradecimentos

Este trabalho contou com a colaboração do grupos TERMISUL/TEXTECC da UFRGS, que disponibilizou o corpus de Pediatria JPED-Coutlhard e as listas de referência. Esses grupos têm apoio apoio financeiro do CNPq, FINEP e SEBRAE, e a pesquisa foi parcialmente realizada no projeto COMUNICA (FINEP/SEBRAE 1194/07).

Referências

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, e Miriam A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. Em R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, e M.C. Dias, editores, *Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, volume 3960 of *Lecture Notes in Computer Science*. Springer-Verlag, pp. 50–59, May, 2006.

Baldwin, T. 2005. The deep lexical acquisition of english verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, e Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. Em *of the Fourth (LREC 2004)*, Lisbon, Portugal, May, 2004.

Banerjee, S. e T. Pedersen. 2003. The Design, Implementation and Use of the Ngram Statistics Package. Em *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 370–381.

Biber, D., S. Johansson, G. Leech, S. Conrad, e E. Finegan. 1999. *Grammar of Spoken and Written English*. Longman, Harlow.

Brown, P., V. Della-Pietra, S. Della-Pietra, e R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312.

Burnard, Lou. 2007. User Reference Guide for the British National Corpus. Relatório técnico, Oxford University Computing Services, February, 2007.

Calzolari, Nicoletta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, e Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. Em *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1934–1940, Las Palmas, Canary Islands.

Cambridge. 1994. *Cambridge International Dictionary of English*. Cambridge University Press.

Carroll, J. e C. Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. Em B. Boguraev e E. Briscoe, editores, *Computational Lexicography for Natural Language Processing*. Longman.

Caseli, H. M. 2003. Alinhamento sentencial de textos paralelos português-inglês. Tese de Mestrado, Instituto de Ciencias Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). 101 p.

Caseli, H. M. 2007. *Indução de léxicos bilíngües e regras para a tradução automática*. Tese de doutoramento, Instituto de Ciencias Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). 158 p.

Caseli, H. M., M. G. V. Nunes, e M. L. Forcada. 2006. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation*, 20:227–245.

Caseli, H. M., A. M. P. Silva, e M. G. V. Nunes. 2004. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. Em *Proceedings of the SBIA 2004 (LNAI)*, number 3171, pp. 184–193, Berlin Heidelberg. Springer-Verlag.

Caseli, H. M., A. Villavicencio, A. Machado, e M. J. Finatto. 2009a. Statistically-driven alignment-based multiword expression identification for technical domains. Em *Proceedings of the 2009 Workshop on Multiword Expressions (ACL-IJCNLP 2009)*, pp. 1–8.

Caseli, Helena Medeiros, Carlos Ramisch, Maria das Graças Volpe Nunes, e Aline Villavicencio. 2009b. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 1:1–20.

Copestake, Ann e Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. Em *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Coulthard, R. J. 2005. The application of corpus methodology to translation: the jped parallel corpus and the pediatrics comparable corpus. Tese de Mestrado, Universidade Federal de Santa Catarina.

Dias, Gaël e Gabriel Pereira Lopes, 2005. *Extração Automática de Unidades Polilexicais para o Português*, pp. 155–184. Mercado de Letras / FAPESP, Campinas, SP, Brasil.

Dias, Gaël e Sergio Nunes. 2001. Combining evolutionary computing and similarity measures to extract collocations from unrestricted texts. Em *Proceedings of RANLP 2001 (Recent Advances in NLP)*, September, 2001.

Evert, S. e B. Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. Em David Farwell, Laurie Gerber, e Eduard Hovy, editores, *Machine Translation and the Information Soup: Proceedings of the Third Conference for Machine Translation in the Americas, AMTA'98*, volume 1529, pp. 1–17. Springer-Verlag, October, 1998.

Haghighi, Aria, Percy Liang, Taylor Berg-Kirkpatrick, e Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. Em *of the 46th : (ACL-08: HLT)*, Columbus, OH, USA, June, 2008.

Hofland, K. 1996. A program for aligning English and Norwegian sentences. Em S. Hockey, N. Ide, e G. Perissinotto, editores, *Research in Humanities Computing*, pp. 165–178, Oxford. Oxford University Press.

Jackendoff, R. 1997. Twistin' the night away. *Language*, 73:534–59.

Krieger, M. G. e M. J. B. Finatto. 2004. *Introdução à Terminologia: teoria & prática*. Editora Contexto.

Lopes, Lucelene, Renata Vieira, Maria José Finatto, Daniel Martins, Adriano Zanette, e Luiz Carlos

Ribeiro Jr. 2009. Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS. Electronic journal of communication information and innovation in health (English edition.Online)*, 3:72–84.

Maia, Belinda. 2003. Using corpora for terminology extraction: Pedagogical and computational approaches. Em B. Lewandowska-Tomasczczyk, editor, *PALC 2001 – Practical Applications of Language Corpora*, pp. 147–164.

Maia, Belinda e Sérgio Matos. 2008. Corpografo v4 - tools for researchers and teachers using comparable corpora. Em *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*, pp. 79–82, May, 2008.

Manning, Christopher D. e Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, USA.

Melamed, I. Dan. 1997. Automatic discovery of non-compositional compounds in parallel data. Em *of the 2nd (EMNLP-2)*, Brown University, RI, USA, August, 1997.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English. A Corpus-based Approach.* Oxford: Clarendon Press.

Och, F. J. e H. Ney. 2000a. A comparison of alignment models for statistical machine translation. Em *Proceedings of the 18th International Conference on Computational Linguistics (COLING'00)*, pp. 1086–1090, Saarbrücken, Germany, August, 2000.

Och, F. J. e H. Ney. 2000b. Improved statistical alignment models. Em *Proceedings of the 38th Annual Meeting of the ACL*, pp. 440–447, Hong Kong, China, October, 2000.

Pearce, Darren. 2002. A comparative evaluation of collocation extraction techniques. Em *of the Third (LREC 2002)*, Las Palmas, Canary Islands, Spain, May, 2002.

Pecina, Pavel. 2008. A machine learning approach to multiword expression extraction. Em *Proceedings of the LREC Workshop Towards a Shared Task for MWE 2008*, Marrakech, Morocco, June, 2008.

Ramisch, Carlos. 2009. Multiword terminology extraction for domainspecific documents. Tese de Mestrado, École Nationale Supérieure d'Informatiques et de Mathématiques Appliquées, Grenoble, França.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, e Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. Em *of the LREC Workshop Towards a Shared Task for (MWE 2008)*, pp. 50–53, Marrakech, Morocco, June, 2008.

Ranchhod, Elisabete Marques e Cristina Mota. 1998. Dicionários electrónicos de léxicos terminológicos. "Seguros". Em *Actas do Workshop sobre Linguística Computacional da APL*. APL.

Riehemann, Susanne. 2001. *A Constructional Approach to Idioms and Word Formation*. Tese de doutoramento, Stanford University.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, e D. Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. Em *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, volume 2276 of *(Lecture Notes in Computer Science)*, pp. 1–15, London, UK. Springer-Verlag.

Santos, Diana e Alberto Simões. 2008. Portuguese-english word alignment: some experiments. Em *of the Sixth (LREC 2008)*, Marrakech, Morocco, May, 2008.

Silva, Joaquim Ferreira da, Gaël Dias, Sylvie Guilloré, e José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. Em *Proceedings of the 9th Portuguese Conference on Artificial Intelligence (EPIA '99)*, volume 1695, pp. 113–132, London, UK. Springer-Verlag.

Simões, Alberto e José J. Almeida. 2008. Bilingual terminology extraction based on translation patterns. *Procesamiento del Lenguaje Natural*, 41:281–288, September, 2008.

Smadja, Frank A. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Van de Cruys, T. e B. Villada Moirón. 2007. Semantics-based Multiword Expression Extraction. Em *Proceedings of the Workshop on A Broader Prespective on Multiword Expressions*, pp. 25–32, Prague, June, 2007.

Vieira, Renata, Maria José Finatto, Daniel Martins, Adriano Zanette, e Luiz Carlos Ribeiro Jr. 2009. Extração automática de termos compostos para construção de ontologias: Um experimento na área da saúde. *Reciis - Revista Eletronica de Comunicação Informação e Inovação em Saúde*, 3:76–88.

Villada Moirón, B. e J. Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. Em *Proceedings of the*

Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), pp. 33–40, Trento, Italy.

Villavicencio, A., H. M. Caseli, e A. Machado. 2009. Identification of multiword expressions in technical domains: Investigating statistical and alignment-based approaches. Em *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology*, pp. 1–9.

Villavicencio, A., V. Kordoni, Y. Zhang, M. Idiart, e C. Ramisch. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. Em *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1034–1043, Prague, June, 2007.

Vogel, S., H. Ney, e C. Tillmann. 1996. HMM-based word alignment in statistical translation. Em *COLING'96: The 16th International Conference on Computational Linguistics*, pp. 836–841, Copenhagen, August, 1996.

Witten, Ian H. e Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.