Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

# Extração Automática de Multipalavras

Aline M. Paz da Silva
Maria das Graças V. Nunes

**NILC-TR-03-11**

Novembro 2003

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional

NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

# Extração Automática de Multipalavras

## 1. Introdução

Para que se possa traduzir e analisar textos em uma determinada língua é necessário levar em consideração o fato de que tais textos não são formados apenas por unidades simples (palavras), mas eles possuem também unidades mais complexas: unidades multipalavras – grupos de palavras que quando combinadas possuem um significado único. Diante disto, algumas ferramentas computacionais têm sido desenvolvidas para identificar e extrair unidades multipalavras a partir de um córpus eletrônico.

Este trabalho consiste no desenvolvimento de ferramentas para realização desta tarefa de extração automática de unidades multipalavras contidas em um córpus. Ele está sendo realizado no âmbito do projeto de mestrado sobre alinhamento lexical de textos paralelos Português-Inglês.

O alinhamento consiste em tentar encontrar as possíveis correspondências existentes entre as unidades que compõe o texto fonte e o texto alvo, qualquer que seja o nível da unidade utilizada. A classificação dos procedimentos de alinhamento tem como base a definição da unidade do texto a ser alinhada, podendo ser parágrafos, sentenças, palavras e unidades multipalavras. Essa classificação determina o nível em que se dá o alinhamento, denominado nível de resolução. No caso específico do alinhamento lexical as unidades consideradas para o processo de estabelecimento de correspondência são as palavras e as unidades multipalavras.

Como já mencionado, para o processo de alinhamento lexical o par de línguas selecionado foi o Português Brasileiro e o Inglês. Desta forma, a tarefa de extração automática de unidades multipalavras foi realizada tanto em córpus do Português quanto em córpus do Inglês.

Na seção 2, as técnicas selecionadas para a realização do processo de extração automática são mostradas de forma detalhada. Em seguida, na seção 3, são apresentados os resultados obtidos a partir da utilização das técnicas nos córpus construídos para a tarefa de extração de multipalavras. Na seção 4 é mostrada a avaliação realizada nas listas geradas pelas técnicas de extração de multipalavras, contrastando os resultados obtidos a partir de cada uma das técnicas selecionadas para este processo. Por fim, na seção 5, são apresentadas algumas conclusões deste trabalho. O Apêndice *Unidades Multipalavras Extraídas pelas Técnicas Implementadas* contém alguns exemplos de unidades multipalavras relevantes geradas por cada uma das técnicas de extração utilizadas neste trabalho.

## 2. Extração de Unidades Multipalavras

Uma das maiores dificuldades do processo de alinhamento lexical de textos paralelos é o tratamento das unidades multipalavras. Isto se deve ao fato de que os textos paralelos não podem ser somente alinhados palavra a palavra; grupos de palavras (unidade multipalavra), que têm significado como uma unidade única, devem ser tratados como unidades lexicais para efeito de alinhamento. Dessa forma, é necessário encontrar meios de se identificar no córpus as unidades que não podem ser alinhadas palavra a palavra, tais como "base de dados", as quais devem ser consideradas como um grupo de palavras. Uma das maneiras de se tratar tais unidades é utilizando-se listas de unidades multipalavras que são consultadas durante o alinhamento para auxiliar na identificação das unidades multipalavras contidas nos textos a serem alinhados. Essas listas podem ser construídas manualmente ou automaticamente a partir de um córpus, outra maneira é via detecção das multipalavras. Neste trabalho, optamos por utilizar listas pré-construídas de forma automática. Para isso, duas técnicas foram investigadas e aplicadas: a *Esperança Mútua* e o pacote de extração de *n-gramas* chamado *NSP*[1].

[1] Disponível em: http://www.d.umn.edu/~tdeperse/code.html

Os córpus e as técnicas que serão utilizados para a tarefa de extração automática das unidades multipalavras estão descritos nas seções a seguir.

### 1.1. Os Córpus utilizados para a Extração de Multipalavras

Para a tarefa de extração automática de unidades multipalavras é necessário utilizar córpus diferentes daqueles que serão usados como entrada para as técnicas de alinhamento lexical, para não influenciar o desempenho dos protótipos alinhadores. Sendo assim, foram construídos alguns córpus para cada uma das línguas envolvidas no processo de alinhamento, Português Brasileiro e Inglês.

Os córpus construídos para esta tarefa são formados por um córpus específico e um córpus *geral*. O córpus específico é constituído de textos técnicos da área de Ciência da Computação. A razão da utilização deste córpus é a necessidade de se obter unidades multipalavras específicas deste domínio de conhecimento, uma vez que os textos a serem alinhados pertencem a essa área específica. Já o córpus geral é composto por textos jornalísticos para que se possam obter as unidades multipalavras mais gerais possíveis (presentes em textos de qualquer domínio). Os córpus construídos foram codificados analogamente aos córpus que serão utilizados no processo de alinhamento, possuindo marcações de início e fim de texto, sentença e palavra. Além disso, tanto os córpus do Português Brasileiro quanto os córpus do Inglês foram etiquetados com um *tagger* do tipo MXPOST[2] para que pudessem ser analisados de acordo com suas classes gramaticais, ajudando a eliminar possíveis composições de multipalavras incoerentes.

Para o Português Brasileiro, os córpus construídos estão definidos da seguinte forma:

1. Córpus EP – Córpus Específico do Português: Composto por 52 arquivos contendo introduções de artigos, monografias, dissertações na área de Ciência da Computação desenvolvidos no ICMC-USP-SÃO CARLOS. A quantidade total de ocorrência de palavras neste córpus é de 1.940.708 (sendo considerados também números).

[2] Disponível em: http://nilc.icmc.usp.br/nilc/tools/nilctaggers.html

2. Córpus GP – Córpus Geral do Português: Composto por 53 arquivos contendo textos jornalísticos obtidos do Jornal do Brasil e da Folha de São Paulo. A quantidade total de ocorrência de palavras contidas neste córpus é de 3.367.219 (sendo considerados também números).

Analogamente, para o Inglês, os córpus construídos estão definidos da seguinte forma:

1. Córpus EI – Córpus Específico do Inglês: Composto por 100 arquivos contendo textos da área de ciência da computação obtidos do *ACM Journals*[3]. A quantidade total de palavras no córpus é 904.915 (sendo considerados também números).
2. Córpus EG – Córpus Geral do Inglês: Composto por 38 arquivos contendo textos jornalísticos obtidos do *New York Times*. A quantidade total de palavras no córpus é 942.212 (sendo considerados também números).

### 2.2. Técnicas de Extração de Unidade Multipalavras

Foram utilizados dois extratores distintos para a realização da tarefa de obtenção automática das unidades multipalavras. Um dos extratores foi implementado utilizando-se uma técnica baseada no cálculo da *Esperança Mútua* (Dias & Kaalep, 2002). O outro extrator utilizado foi o pacote *NSP* (*N-gram Statistic Package*). A seguir é apresentada uma descrição detalhada de cada um destes extratores de multipalavras.

#### 2.2.1. Esperança Mútua

Esta técnica baseia-se no fato de que unidades multipalavras são agrupamentos de palavras que tem probabilidade elevada de ocorrerem de forma combinada nos segmentos do texto. Dessa forma, a *Esperança Mútua* (*EM*) é usada para determinar o grau de coesão entre as palavras contidas em um *n-grama*. Para calcular esse grau de coesão entre as palavras, a medida *Esperança Mútua* utiliza uma outra medida denominada Esperança Normalizada (*EN*).

[3] Disponível em: http://www.informatik.uni-trier.de/~ley/db/journals/

A *EN* é definida como a esperança de ocorrência de uma palavra em uma dada posição, sendo conhecidas as ocorrências das demais *n-1* palavras e suas respectivas posições. Como exemplo, considere-se a unidade multipalavra "engenharia de software". O objetivo da *EN* é determinar o grau de esperança desse *trigrama*, levando em consideração a esperança da unidade "software" aparecer depois de "engenharia de", como também a esperança de ocorrência da unidade "de" ligando as unidades "engenharia" e "software" e finalmente a esperança da unidade "engenharia" ocorrer antes de "de software". Quanto mais coeso um grupo de palavras é, maior será sua esperança normalizada.

A fórmula utilizada para o cálculo da esperança normalizada é descrita pela razão entre a probabilidade de ocorrência de um *n-grama* e a média das probabilidades de ocorrência dos *n-1-gramas* que ele contém, como mostrado a seguir:

$$EN = \frac{prob(n-grama)}{\frac{1}{n}\sum prob(n-1-grama)}$$

A probabilidade, neste caso, é determinada pela razão entre a freqüência de ocorrência do *n-grama* em questão e o número total de possíveis *n-gramas* que ocorrem no córpus.

Uma vez calculada a *EN* para os *n-gramas*, é possível determinar qual deles tem maior chance de ser uma unidade multipalavra calculando-se a *EM* do *n-grama*. De acordo com (Daille, 1995 apud Dias & Kaalep, 2002) se forem considerados dois *n-gramas* com mesma esperança normalizada, o mais provável de ser uma unidade multipalavra será aquele que possuir maior freqüência relativa. Assim:

$$EM = prob(n-grama) \times EN(n-grama)$$

Dessa forma, a *Esperança Mútua* entre as palavras que compõem um *n-grama* é baseada na esperança normalizada e na freqüência relativa das palavras no córpus.

Depois que os valores de EM são calculados para todos os *n-gramas* candidatos e também para todos *n-1-gramas* que os compõem, é necessário determinar qual deles é realmente uma unidade multipalavra (o *n-grama* ou um dos *n-1-gramas* que o compõem). Para isso, utiliza-se um algoritmo conhecido como GenLocalMaxs (Dias & Kaalep, 2002).

Esse algoritmo elege as unidades multipalavras a partir do conjunto de *n-gramas* candidatas, baseado em duas suposições: 1) quanto mais coeso um grupo de palavras (*n-grama*) é, maior será o valor de EM associado a ele; 2) unidades multipalavras são grupos de palavras altamente associados. De acordo com essas suposições, é possível deduzir que: um *n-grama* é considerado uma unidade multipalavra se o grau de coesão entre as palavras que o compõem é maior ou igual ao grau de coesão de todos os subgrupos (*n-1-gramas*) que ele contém, e é maior que o grau de coesão de todos os supergrupos (*n+1-gramas*) que o contém. Como no caso do exemplo anterior, considerando o *trigrama* "engenharia de software", ele será considerado uma unidade multipalavra se o seu valor de ME for maior que o valor de ME para os *bigramas* "engenharia de" e "de software" e também seja maior que o valor de ME para qualquer *tetragrama* que o contém.

**Modificações no método da Esperança Mútua**

Os córpus descritos na seção 1 foram submetidos à técnica de extração de unidades multipalavras apresentada e, como resultado final, foram obtidas quatro listas de unidades multipalavras, uma para cada córpus construído.

Numa primeira fase, foi utilizada uma implementação seguindo exatamente a descrição do método apresentada nesta subseção. Os quatro córpus construídos foram fornecidos como entrada para o protótipo implementado. Dessa forma, foram obtidas quatro listas de unidades multipalavras, duas para o PB e duas para o Inglês.

As listas foram analisadas e foi possível observar que muitas unidades incoerentes eram geradas, por exemplo, as unidades "para de", "a de", "uma a", entre outras. Diante deste problema, decidiu-se utilizar, combinado a esta técnica de extração, um filtro capaz de minimizar a ocorrência destas unidades consideradas "sem sentido". O filtro é o mesmo utilizado na técnica de extração de unidades multipalavras baseada no grau de entropia entre as palavras descrito em (Merkel & Andersson, 2000). Tal filtro, que é aplicado durante a fase de construção da unidade multipalavra, consiste da utilização de três listas de palavras que não podem iniciar, ser parte de, ou finalizar uma unidade multipalavra. Tais listas são definidas como:

uma lista de palavras que não podem iniciar uma unidade multipalavra (por exemplo, artigos, preposições, números ou pontuação);

uma lista de palavras que não podem fazer parte de uma unidade multipalavra (por exemplo, números ou pontuação); e

uma lista de palavras que não podem terminar uma unidade multipalavra (por exemplo, artigos, números ou pontuação).

Um novo protótipo foi implementado, nesta nova fase já com o filtro para eliminar unidades multipalavras "sem sentido", e quatro novas listas foram obtidas. As novas listas ainda apresentavam unidades incoerentes, mas em um número bem mais reduzido. Os resultados são discutidos na seção 1.3.

### 2.2.2. N-gram Statistic Package (NSP)

O *NSP* é um pacote desenvolvido por Ted Pedersen que é composto por um conjunto de programas que buscam identificar e analisar os *n-gramas* contidos em um córpus. Para este processo são utilizados dois programas implementados em Perl, o ***count.pl*** e o ***statistics.pl***.

O programa ***count.pl*** toma um arquivo texto como entrada, formado pelo córpus do qual devem ser extraídos os *n-gramas*, e gera uma lista contendo todos os *n-gramas* contidos no arquivo e suas respectivas freqüências. O arquivo de saída possui na primeira linha a quantidade total de *n-gramas* encontrados e as linhas subseqüentes possuem os *n-gramas* acompanhados de suas respectivas freqüências de ocorrência no córpus, obedecendo ao seguinte padrão:

*token1<>token2<>...<>tokenN<>freq.n-grama freq.token1 freq.token2 …freq.tokenN*

Os tokens do *n-grama* são separados e delimitados pelo símbolo "<>". Assim, o *bigrama* "por exemplo" iria aparecer na saída como "por<>exemplo<>" seguido de sua freqüência de ocorrência no córpus e da freqüência dos *unigramas* ("por" e "exemplo") que o compõem. Da mesma forma, o *trigrama* "engenharia de software" aparecerá como "engenharia<>de<>software<>" também seguido de sua freqüências e das freqüências dos *n-1-gramas* que o compõem.

Como observado nos exemplos acima, este programa permite realizar a extração de *n-gramas* de tamanho *N*, sendo que *N* é o número de tokens que se deseja obter em cada *n-grama*. Por exemplo, para obtenção de *bigramas*, *N* = 2, para *trigramas*, *N* = 3, e assim por diante. Tal tamanho deve ser definido pelo usuário.

O outro programa, ***statistics.pl***, utiliza como entrada a lista gerada pelo programa anterior e computa um escore para cada um dos *n-gramas* de acordo com uma medida estatística. Entre as medidas possíveis estão o coeficiente de *Dice*, a *log-likelihood*, a informação mútua, entre outras. O ***statistic.pl*** toma as listas dos *n-gramas* com suas respectivas freqüências e os fornece aos pacotes da medida selecionada para que, com base nestas freqüências, sejam calculados os escores referentes aos *n-gramas* encontrados. Para o caso específico deste trabalho, a medida selecionada para o cálculo dos escores foi a *log-likelihood*, pois, das medidas disponíveis, ela é a única que pode ser utilizada tanto para *bigramas* quanto para *trigramas*.

A medida *log-likelihood* indica quanto uma determinada ocorrência de um *n-grama* é mais provável de ser uma unidade multipalavra do que uma outra ocorrência. Esta medida leva em consideração a ocorrência dos tokens que formam um *n-grama* e o seu complemento, ou seja, a não ocorrência destes tokens. Sua fórmula é definida como:

$$\log - likelihood = -2\log \frac{\left[P_X P_Y P_{\overline{X}} P_{\overline{Y}}\right]^{f_Y}}{\left[P_{XY} P_{\overline{XY}}\right]^{f_{XY}} \left[P_{X\overline{Y}} P_{\overline{X}Y}\right]^{f_{\overline{X}Y}}}$$

onde $f_X$ é a freqüência de ocorrência de uma palavra *X*; $f_Y$ é a freqüência de ocorrência de uma palavra Y; $P_X$ é a probabilidade de ocorrência de uma palavra *X*; $P_Y$ é a probabilidade de ocorrência de uma palavra *Y*. A variável XY representa um *bigrama* formado pelos tokens *X* e *Y*. A barra sobreposta indica o complemento de uma variável.

Uma vez que os escores foram calculados segundo a medida *log-likelihood*, o programa ***statistic.pl*** gera um arquivo de saída, no qual a primeira linha é a quantidade de *n-gramas* gerados e nas linhas subseqüentes estão os *n-gramas* ordenados segundo um *rank* que é determinado pelo valor do escore encontrado, seguindo o padrão:

*token1<>token2<> ... <>tokenN<>rank escore freq.n-grama freq.token1etoken2 freq.token1etokenN … freq.token1 freq.token2 ... freq.tokenN*

Para estabelecer o *rank* usa-se o seguinte critério: o maior escore recebe *rank* igual a 1, o segundo maior escore recebe igual a 2, o terceiro maior escore recebe *rank* igual a 3 e assim por diante.

Todo este processo é realizado de maneira semelhante para os *trigramas*, de forma que o escore é calculado pela medida *log-likelihood* estendida diretamente para os *trigramas*. A única modificação é que ao invés de serem considerados apenas dois tokens, passam a ser considerados três tokens.

## 3. Resultados obtidos

### Listas Originais

A partir dos métodos anteriormente descritos foram obtidas oito listas de unidades multipalavras. Quatro destas listas foram obtidas pela técnica baseada em *Esperança Mútua* aplicada ao córpus EP, ao córpus GP, ao córpus EI e ao córpus GI, respectivamente. De forma semelhante, as outras quatro listas foram extraídas pelo pacote *NSP* aplicado aos mesmos córpus.

A Tabela 1 mostra o contraste entre a quantidade de unidades contidas nas listas obtidas utilizando-se a técnica de *Esperança Mútua* e o pacote *NSP* para realizar o processo de extração de unidades multipalavras. É essencial lembrar que todas as listas geradas, independentemente de qual extrator está sendo utilizado, são dependentes dos córpus que estão sendo utilizados, sendo a abrangência das listas tão amplas quanto a abrangência dos córpus em questão.

Tabela 1 – Quantidade de unidades das listas geradas pelas técnicas de extração selecionadas.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua*** | 12.197 | 23.196 | 8.694 | 17.226 |
| ***NSP*** | 90.503 | 185.650 | 30.516 | 73.938 |

Analisando as listas geradas foi possível observar que muitas das unidades geradas não eram unidades relevantes (um exemplo de uma unidade considerada relevante é o bigrama "guerra fria", outros exemplos podem ser encontrados no Apêndice), ou seja, eram unidades que não deveriam necessariamente ser consideradas como unidades multipalavras, como por exemplo, as unidades "alto desempenho", "feminino e masculino", "O desempenho". Com base neste fato, percebe-se que era necessário realizar uma etapa de eliminação de tais unidades, permitindo que só fizessem parte da lista apenas as unidades que realmente satisfizessem as propriedades de uma multipalavra.

Todo o processo e os critérios estabelecidos para a tarefa de eliminação das unidades consideradas não relevantes estão descritos na subseção a seguir.

**Eliminando Unidades Não Relevantes**

Em (Schone & Jurafsky, 2001), unidade multipalavras são definidas como uma seqüência de palavras vizinhas conectadas cujo significado ou conotação não pode ser obtido a partir do significado ou conotação de seus componentes. Desta forma, é possível estabelecer alguns critérios que devem ser satisfeitos para que um determinado *n-grama* possa ser considerado uma unidade multipalavra. Tais critérios são: a não-composicionalidade, a não-substitutibilidade e a não-modificabilidade.

A não-composicionalidade é a propriedade que garante que o significado de um *n-grama* só pode ser obtido a partir do conjunto de palavras que o compõem como um todo, não podendo ser obtido a partir da decomposição de seus componentes. O *bigrama* "alto desempenho" é um exemplo de uma unidade que não satisfaz o critério da não-composicionalidade, uma vez que o seu significado pode ser obtido a partir da composição do significado das palavras que o compõem. Em contrapartida, o *bigrama* "guerra fria" é um exemplo de uma unidade que satisfaz o critério da não-composicionalidade, pois o seu significado só pode ser abstraído do grupo de palavras como um todo.

A não-substituibilidade garante que nenhum dos componentes do *n-grama* pode ser substituído por seus sinônimos, pois tal substituição não mais transmite o significado original. Por exemplo, o *n-grama* "compact disk" não necessariamente implica "densily-

packed disk"; pode-se tratar de uma mídia musical, nesse caso ele é considerado não-substituível.Por outro lado, um exemplo de unidade que não satisfaz a propriedade de não-substituibilidade é o *bigrama* "preço alto", pois se a palavra "alto" for trocada por "elevado" ("preço elevado") no *bigrama* em questão o significado ainda é mantido.

Por fim, a não-modificabilidade estabelece o critério de que a estrutura de uma unidade multipalavra não pode ser modificada sem que seu significado seja alterado. Considerando mais uma vez a unidade "compact disk", se mudarmos sua estrutura para "disk that is compact" seu significado não é preservado, satisfazendo o critério da não-modificabilidade. Em contraste, o *bigrama* "garota bonita" não perde seu significado se for modificado para "bonita garota" não obedecendo desta forma o critério da não-modificabilicade.

Além de satisfazer as propriedades acima descritas, foi estabelecida ainda a condição de que dos córpus em Português Brasileiro seriam consideradas apenas as unidades geradas em Português; analogamente, dos córpus em Inglês seriam consideradas apenas as unidades geradas em Inglês. Esta condição foi estabelecida pois algumas vezes eram geradas *n-gramas* em Inglês nas listas geradas a partir dos córpus em Português o que não era interessante para o objetivo deste trabalho.

As eliminações foram realizadas manualmente por agentes humanos, numa fase de pós-processamento (depois da geração das listas), levando em consideração todos os critérios pré-estabelecidos acima.

**Listas Finais**

Obedecendo aos critérios descritos, após o processo de eliminação das unidades consideradas não relevantes foram obtidas oito novas listas, equivalentes às listas originalmente obtidas pelos córpus de entrada (córpus EP, córpus GP, córpus EI e córpus GI) submetidos às técnicas de extração (*Esperança Mútua* e *NSP*), algumas das unidades contidas nestas listas são mostradas no Apêndice.

A Tabela 2 mostra como são compostas, em unidades, as listas finais obtidas após o processo de eliminação. Pode-se notar pela Tabela 2 que a quantidade de unidades obtidas

nas listas geradas pelo pacote *NSP* foi aproximadamente duas vezes maior que a quantidade obtida pela técnica baseada em *Esperança Mútua*.

Tabela 2 - Quantidade de unidades das listas obtidas após o processo de eliminação.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua*** | 153 | 1.115 | 140 | 348 |
| ***NSP*** | 260 | 1.556 | 262 | 645 |

## 4. Avaliação das Listas Obtidas

Utilizando-se os extratores automáticos de unidades multipalavras foram obtidas no total dezesseis listas de unidades multipalavras: oito listas originais, obtidas a partir da aplicação das técnicas de extração aplicadas aos córpus construídos, e oito listas processadas, obtidas a partir das listas originais após um processo de eliminação de unidades não relevantes. Com base nos resultados obtidos foi então feita uma comparação entre as listas geradas seguindo alguns critérios pré-determinados para que fosse possível avaliar as técnicas de extração utilizadas.

Os critérios estabelecidos para a comparação das listas foram: 1) a quantidade de unidades geradas; 2) a quantidade de unidades eliminadas (somente aplicado às listas originais); 3) a quantidade de unidades coincidentes e 4) quantidade de unidades contidas no conjunto diferença das listas geradas.

A avaliação consistiu de duas etapas, a primeira delas foi feita utilizando-se as listas originais e a outra utilizando-se as listas processadas. Todo o processo de avaliação é descrito de forma detalhada nas próximas subseções.

### 4.1. Etapa 1 - Antes da Eliminação (Listas Originais)

A primeira etapa foi realizada utilizando-se as listas originais exatamente da maneira como foram geradas pelas técnicas de extração, sem nenhum processamento ou eliminação de unidades não relevantes. Para possibilitar uma melhor visualização dos resultados obtidos, foram montados, durante o processo de avaliação, tabelas e gráficos comparativos

mostrando o desempenho de cada um dos extratores de multipalavras utilizados neste trabalho.

A Tabela 1, apresentada na seção 1.3, mostra a comparação baseada no primeiro critério de avaliação estabelecido, que contrasta a quantidade de unidades geradas pela técnica baseada em *Esperança Mútua* e pelo pacote *NSP*. O gráfico obtido com base nos valores contidos na Tabela 1 pode ser visualizado na Figura 1 mostrada abaixo.

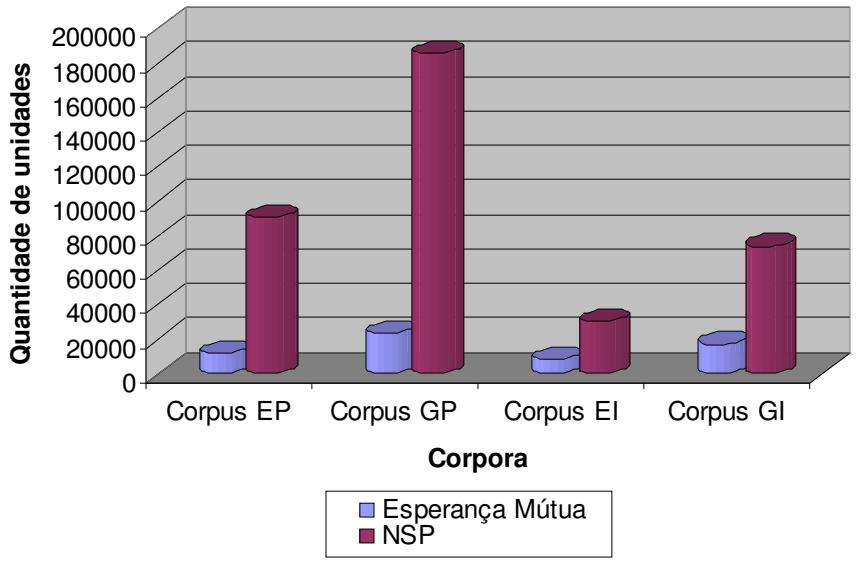


Figura 1 – Comparação entre a quantidade de unidades geradas pelas técnicas de extração.

Observando-se a tabela (Tabela 1) e a figura (Figura 1) obtidas nesta primeira comparação, pode-se perceber que o pacote *NSP* sempre gera mais unidades (aproximadamente cinco vezes mais) do que a técnica baseada em *Esperança Mútua*. Isto pode ser explicado pelo fato de que o pacote *NSP* gera todas as possíveis combinações de *n-gramas* existentes no córpus em questão, enquanto que a *Esperança Mútua* só gera os *n-gramas* que têm um elevado grau de associação entre as palavras que o compõem.

O segundo parâmetro consiste em estabelecer uma comparação entre a quantidade de unidades eliminadas em cada uma das listas geradas pelos extratores automáticos. A Tabela 3 foi montada de acordo com os resultados obtidos depois do processo de exclusão das unidades que não satisfaziam os critérios estabelecidos na seção 1.3. A Figura 2 apresenta o gráfico gerado a partir dos dados da Tabela 3.

Tabela 3 – Quantidade de unidades eliminadas das listas originais.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua*** | 12.044 | 22.068 | 8.554 | 16.878 |
| ***NSP*** | 90.243 | 184.094 | 30.254 | 73.293 |

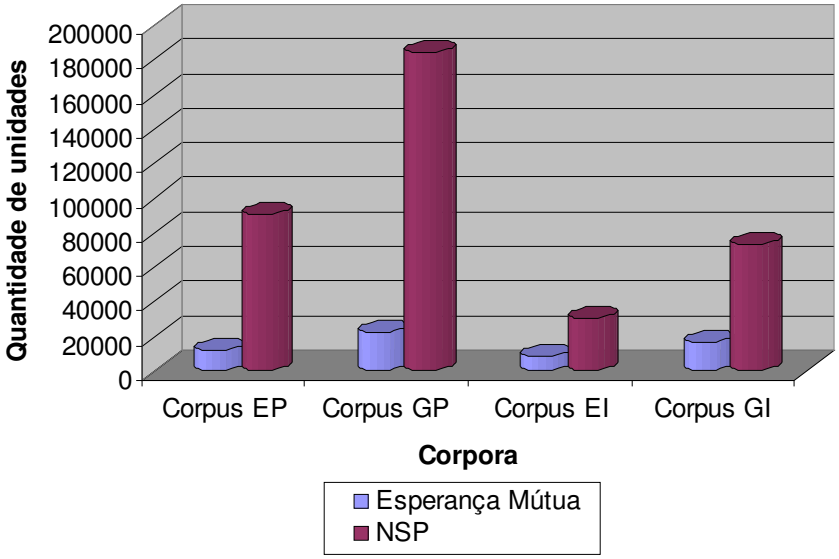


Figura 2 – Comparação entre a quantidade de unidades eliminadas das listas geradas pelas técnicas de extração.

Analisando-se a Tabela 3 e a Figura 2 é possível observar que o número de unidades que foram eliminadas foi bastante elevado tanto para a técnica de extração baseada em *Esperança Mútua* quanto para o pacote *NSP*, fazendo com que as Tabela 1 e Tabela 3 e Figura 1 e Figura 2 ficassem bastante semelhantes. Isto comprova o fato de que a maioria dos extratores de multipalavras existentes hoje necessita de melhoramentos à medida que não geram apenas unidades multipalavras relevantes (para as quais necessita-se considerar um grupo de palavras para obter sua definição).

O próximo critério focaliza as unidades coincidentes nas listas equivalentes geradas pelos dois extratores. A idéia é obter a quantidade de unidades contidas na intersecção entre as listas geradas pela técnica de *Esperança Mútua* e suas equivalentes geradas pelo pacote *NSP*. Os resultados desta operação são apresentados na Tabela 4.

Tabela 4 – Quantidade de unidades coincidentes nas listas originais.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua* ∩ *NSP*** | 5.095 | 9.288 | 2.933 | 5.831 |

Um aspecto interessante que foi observado, e que vale a pena ser mencionado, é que as listas provenientes da operação de intersecção continham algumas unidades não relevantes comuns às duas técnicas de extração, apesar de se esperar que a intersecção das listas nos fornecesse uma certa confiança para as unidades multipalavras geradas. Isto significa que, apesar de utilizarem estratégias diferentes para encontrar as unidades multipalavras, estas técnicas podem coincidir nas falhas, gerando falsas multipalavras comuns.

Por fim, o quarto e último critério procura encontrar as unidades que estão contidas em uma lista, mas não estão presentes em sua equivalente (lista gerada a partir do mesmo córpus, mas com outra técnica de extração automática). A Tabela 5 mostra as diferenças encontradas nesta comparação e a Figura 3 ilustra esta diferença.

Tabela 5 – Quantidade de unidades da diferença entre as listas geradas pelos métodos extratores.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua – NSP*** | 7.102 | 13.908 | 5.761 | 11.395 |
| ***NSP – Esperança Mútua*** | 85.408 | 176.362 | 27.583 | 68.111 |

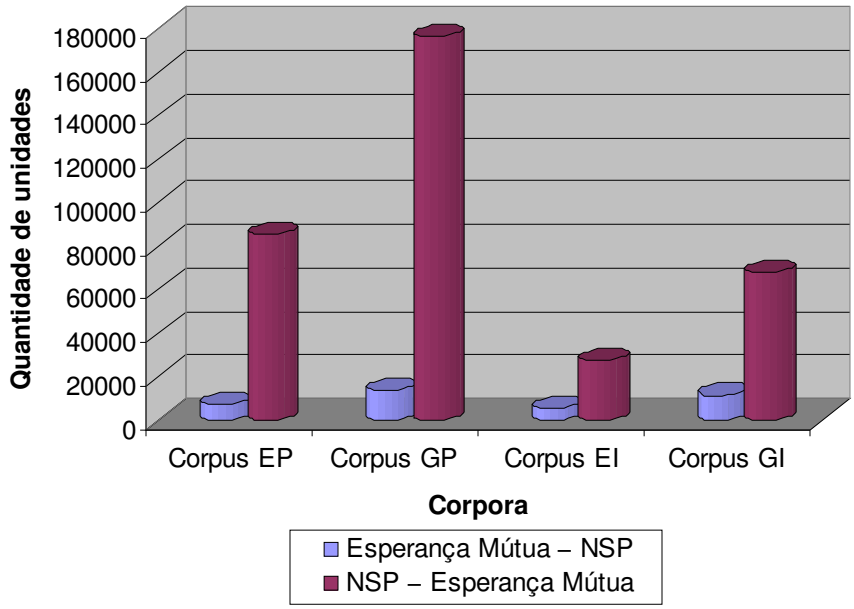


Figura 3 – Diferença entre as listas geradas pelos métodos extratores.

A primeira linha da Tabela 5 mostra quantas unidades estão contidas nas listas geradas pela técnica de extração baseada na *Esperança Mútua* e não estão incluídas nas listas geradas pelo pacote *NSP*. De forma análoga, a segunda linha fornece o número de unidade que são encontradas nas listas obtidas pelo pacote *NSP*, mas não são encontradas nas listas obtidas pela técnica de *Esperança Mútua*.

Depois de realizada esta etapa de avaliação nas listas originais é mais marcante ainda a necessidade de um processo de eliminação dos *n-gramas* que estão inseridos na lista, mas que não devem ser considerados como unidades multipalavras. Na subseção a seguir serão apresentados os resultados das comparações realizadas com as listas processadas.

### 4.2. Etapa 2 - Depois da Eliminação (Listas Processadas)

A etapa seguinte, etapa 2, foi realizada com as listas processadas, que são as listas obtidas depois da eliminação das unidades que não deveriam ser consideradas como multipalavras (unidades não relevantes). Os critérios utilizados para esta avaliação foram os mesmos utilizados na etapa realizada com as listas originais, com exceção do critério de número 2, pois nestas listas as unidades não relevantes já foram excluídas. Como na etapa de avaliação anterior, foram construídos tabelas e gráficos comparativos para exibir o desempenho obtido por cada um dos extratores de unidades multipalavras utilizados.

Com base no primeiro critério estabelecido, quantidade de unidades geradas, foi montada a Tabela 2 já mostrada na seção 1.3. A tabela mostra que as listas foram bastante "enxugadas" ficando com um número bem reduzido de unidades em relação às listas originais, sendo eliminados todos os excessos (unidades não relevantes). Como na comparação das listas originais, o pacote *NSP* obteve um desempenho superior (mais unidades geradas) ao da técnica baseada em *Esperança Mútua*. O gráfico da Figura 4 permite visualizar a diferença entre a quantidade de unidades geradas pelos dois extratores para cada um dos córpus utilizados como entrada.

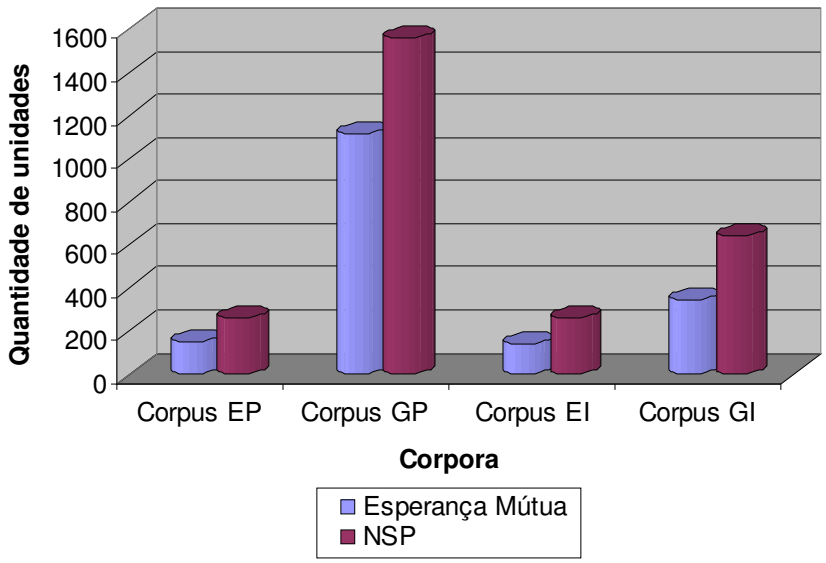


Figura 4 – Comparação entre a quantidade de unidades obtidas após o processo de eliminação.

O próximo critério de avaliação, terceiro critério da etapa 1, estabelece uma comparação entre as unidades coincidentes contidas nas listas de cada uma das técnicas de extração. A Tabela 6 mostra os valores obtidos.

Tabela 6 – Quantidade de unidades coincidentes nas listas processadas.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua ∩ NSP*** | 75 | 718 | 82 | 218 |

Com as listas obtidas nesta comparação foi possível certificar algumas ocorrências de unidades multipalavras, permitindo obter uma lista básica de unidades relevantes.

O último critério explora a propriedade da diferença entre as listas equivalentes gerada pelas diferentes técnicas de extração. A tabela 7 foi montada com os valores obtidos na comparação deste critério e o gráfico da Figura 5 mostra a relação entre tais valores.

Tabela 6 – Quantidade de unidades coincidentes nas listas processadas.

| | Córpus **EP** | Córpus **GP** | Córpus **EI** | Córpus **GI** |
|---|---|---|---|---|
| ***Esperança Mútua – NSP*** | 76 | 397 | 58 | 130 |
| ***NSP – Esperança Mútua*** | 185 | 840 | 181 | 429 |

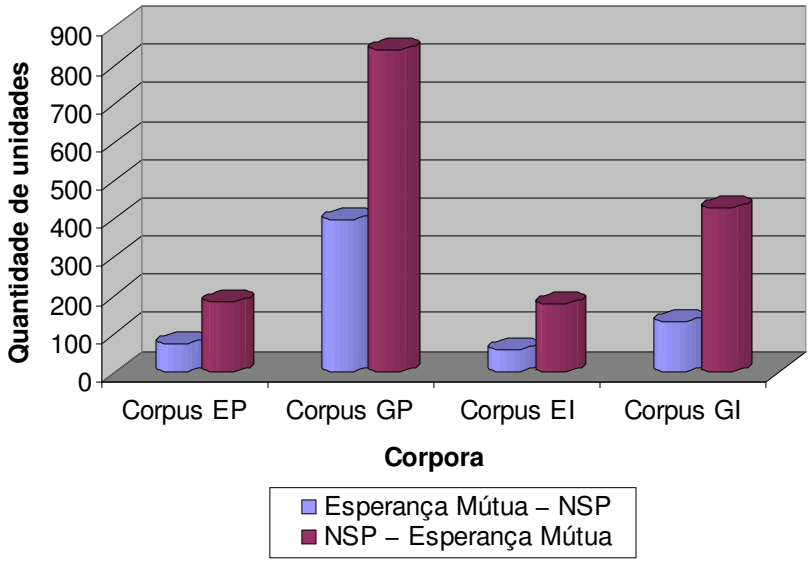


Figura 5 - Diferença entre as listas obtidas após o processo de eliminação.

A diferença na etapa 2 possui a mesma definição apresentada na avaliação das listas originais (etapa 1).

A partir dos resultados obtidos nestas comparações será possível gerar, para cada córpus utilizado como entrada, mais duas novas listas: uma composta pelos elementos da lista básica (formada pela intersecção das listas processadas) e outra formada pela lista básica acrescida dos elementos contidos nas listas compostas pela propriedade da diferença. Estas listas também serão utilizadas no processo de alinhamento para que se possa verificar qual a influência da redução (lista básica possui menos unidades que as listas processadas) e do acréscimo (lista completa – básica mais elementos da diferença – possui mais unidades que as listas processadas) de unidades multipalavras no desempenho do alinhador.

## 5. Conclusões

A extração automática de unidades multipalavras é uma tarefa difícil, pois na maioria das vezes envolve a utilização aspectos específicos da língua (como, por exemplo, os filtros que são utilizados para eliminação de unidades não relevantes) para melhorar o desempenho dos extratores.

Além disso, um outro ponto que deve ser considerado é a dependência dos córpus que estão sendo utilizados. Quantas e quais serão as unidades multipalavras que estarão presentes nas listas geradas depende dos córpus usados como entrada para os extratores automáticos. Desta forma, se os córpus não forem suficientemente abrangentes, as listas poderão não ser tão completas quanto deveriam.

Em relação às técnicas existentes para a tarefa de extração automática, foi possível confirmar o fato de que a maioria dos extratores de multipalavras existentes hoje necessita de melhoramentos à medida que não geram apenas unidades multipalavras relevantes.

Por fim, avaliando o desempenho dos protótipos utilizados na tarefa de extração, foi possível perceber que o pacote *NSP*, embora utilize uma heurística bem mais simples, obteve melhor desempenho do que o extrator baseado na técnica de *Esperança Mútua*, que produziu uma quantidade aproximadamente duas vezes menor de unidades multipalavras válidas.

## Referências

Dias, G.; Kaalep H. (2002). Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs. *In*: H. Metslang, M. Rannut (eds.) Languages in Development, Linguistic Edition 41, Lincom-Europa, München.

Merkel, M.; Andersson M. (2000). Knowledge-lite Extraction of Multi-word Units with Language Filters and Entropy Thresholds. *In*: Proceedings of RIAO'2000, Collége de France, Paris, France, April 12-14, 2000, Volume 1, pp. 737-746.

Shimohata, S.; Sugio, T.; Nagata, J. (1997). Retrieving Collocations by Co-occurrences and Word Order Constraints. *In*: Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL'97), Madrid: 476-481.

## Apêndice : Unidades Multipalavras Extraídas pelas Técnicas Implementadas

Neste apêndice são apresentadas algumas das unidades relevantes geradas pelas técnicas de extração utilizadas neste trabalho. As listas estão separadas de acordo com o córpus utilizado como entrada para cada técnica utilizada para a tarefa de extração.

### 1. Esperança Mútua

**Córpus EP**
Ad hoc
Além disso
Autômato finito determinístico
Ciclo de vida
Código fonte
De modo que
Donald Hung
Engenharia de software
Et al
Gill Harel
Guerra fria
Levam em conta
Ponto de vista
Por exemplo
Von Neumann

**Córpus GP**
Às vezes
Abrir mão de
Além disso
Apesar disso
Bob Marley
Cara ou coroa
Em nível de
Fernando Henrique Cardoso
Flauta doce
Ou seja
Por isso
Por que
Rio de Janeiro
São Paulo
Semana Santa

**Córpus EI**
Case study
Finite state automata
In fact
In this
Marry Monads
Of the
Of this
Pay atttention
Points of view
Points out
Taken into account
There are
Software engineering
Vice versa
World Wide Web

**Córpus GI**
Bar mitzvah
Be used to
Breaking apart
Carry out
Ed Thompson
Fell apart
Find out
For example
Garden of Eden
Make up
Of that
Of these
San Antonio
So far
Sooner or later

**2. Pacote NSP**

**Córpus EP**
Assim como
Bases de dados
Engenharia reversa
Estados Unidos
Et al

Klein Joshi
Ou seja
Ponto flutuante
Por alto
Por engano
Por exemplo
Por outro lado
Por que
São Paulo
Tom Jobim

**Córpus GP**
Ao caso
Assim como
Através de
César Maia
Daniela Mercury
Euclides da Cunha
Frente fria
No entanto
Pedro Paulo Diniz
Por favor
Por isto
Por meio de
Pouco a pouco
São Carlos
Via crucis

**Córpus EI**
As well as
At least
Based on
Break out
Et al
Finite state machine
In the
Instead of
Look at
Of this
On the
Operating systems
Points of view
Taken into account
World wide web

**Córpus GI**
Ariel Sharon
Based on
Be used to
C I A
How many
In the
New York
Of the
Point of view
Saddam Hussein
Set up
So much
Sooner or later
Take into account
United States