Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

TagAlign: Uma ferramenta de pré-processamento de textos

Helena de Medeiros Caseli
Valéria Delisandra Feltrim
Maria das Graças Volpe Nunes

Nº 169

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
Junho/2002

Resumo

Neste relatório é apresentada a descrição da ferramenta NILC TagAlign. Essa ferramenta possui duas funcionalidades básicas: a marcação de partes específicas do texto com etiquetas pré-definidas pelo usuário ou pelo sistema e o alinhamento de textos paralelos utilizando um módulo de suporte ao usuário. O alinhamento automático de textos paralelos também é uma funcionalidade prevista para essa ferramenta. A TagAlign processa textos no formato .txt e gera saídas também com o mesmo formato. As saídas podem ser o texto de entrada marcado ou desmarcado, dois textos paralelos alinhados ou apenas as sentenças alinhadas dos textos paralelos.

Índice

1 INTRODUÇÃO 1

2 A NILC TAGALIGN 2

2.1 MARCAÇÃO 3

2.1.1 Marcação Automática de Fronteiras de Parágrafos e Sentenças 4

2.1.2 Marcação Manual 10

2.1.2.1 Definição do Arquivo de Etiquetas 15

2.2 ALINHAMENTO 20

2.2.1 Alinhamento Sentencial Manual 21

2.3 FERRAMENTAS 27

2.3.1 Geração do Corpus Paralelo 28

2.3.2 Avaliação do Corpus Alinhado 31

2.4 AJUDA 35

3 CONCLUSÃO 36

4 REFERÊNCIAS BIBLIOGRÁFICAS 37

TagAlign: uma ferramenta de pré-processamento de textos. [1]

1 Introdução

O processamento de grandes quantidades de textos para os mais diversos fins e a troca de informação deles derivada são atividades cada vez mais freqüentes nos diversos campos de pesquisa e na sociedade, de uma forma geral. Essas atividades, que na maioria dos casos demandam um tempo relativamente grande para serem concluídas, podem ser agilizadas e/ou facilitadas com o uso de técnicas de pré-processamento dos textos.

Inúmeras ferramentas de pré-processamento de textos têm surgindo nos últimos tempos com o propósito de inserir significado (utilizando etiquetas), formatá-los de acordo com padrões pré-especificados para troca de informações ou apenas garantir mais agilidade no processo final.

Nesse contexto, apresentamos a NILC TagAlign: uma ferramenta de pré-processamento de textos que tem dois módulos (ou funções) básicos – o módulo de marcação e módulo de alinhamento – projetada, inicialmente, para auxiliar a geração de recursos lingüísticos necessários no projeto PESA[2] (*Portuguese-English Sentence Alignment*) e no projeto SciPo[3] (*Tool for Scientific Writing in Brazilian Portuguese*).

O módulo de marcação da TagAlign possui duas variantes: a marcação automática e a marcação manual. Na primeira, são inseridas etiquetas de início e fim de parágrafos e sentenças de modo totalmente automático; e na segunda, o módulo dá suporte ao usuário para a seleção de partes do texto e a inserção de etiquetas pré-determinadas e escolhidas por ele. O outro módulo da ferramenta é responsável pelo alinhamento sentencial manual de textos paralelos, ou seja, sua função é facilitar o processo de inserção de correspondências entre as sentenças de um texto em uma língua (texto fonte) e as sentenças de sua tradução (texto alvo). Como esse processo é manual, a seleção das sentenças que são a tradução umas das outras nos dois textos fica a cargo do usuário.

Os detalhes da TagAlign e de suas funções básicas são apresentadas na seção seguinte (Seção 2). Uma breve conclusão desse relatório é apresentada na última seção (Seção 3).

---

[1] Este trabalho foi apoiado por CNPq, CAPES e FAPESP.
[2] Mais detalhes em Caseli (2002) disponível em www.nilc.icmc.sc.usp.br “Projects/PESA”.
[3] Mais detalhes em Feltrim (2002) disponível em www.nilc.icmc.sc.usp.br “Projects/SciPo”.

2 A NILC TagAlign

A NILC TagAlign, ou simplesmente TagAlign, é uma ferramenta de pré-processamento de textos desenvolvida no Núcleo Interinstitucional de Lingüística Computacional (NILC), no Instituto de Ciências Matemáticas e Computação da Universidade de São Paulo (ICMC-USP) em São Carlos, cuja funcionalidade divide-se em dois grupos de tarefas distintos: Marcação e Alinhamento.

O módulo de Marcação possui duas variantes: a marcação automática de fronteiras de parágrafos e sentenças em todos os parágrafos e sentenças de um texto e a marcação manual de partes específicas do texto, determinadas pelo usuário, com etiquetas por ele definidas e selecionadas.

O módulo de Alinhamento possui, atualmente, apenas a opção de alinhamento sentencial manual, mas também possuirá, num futuro próximo, a opção de alinhamento sentencial automático de textos paralelos e, em futuro mais distante, as mesmas opções para o alinhamento de palavras. O alinhamento manual de sentenças auxilia o usuário na inserção de marcações de correspondência entre as sentenças dos textos fonte e alvo.

Ambos os módulos serão explicados em mais detalhes nas subseções a seguir: 2.1 e 2.2, respectivamente. A Figura 1 traz uma ilustração da tela inicial da TagAlign.

Figura 1. Tela inicial da TagAlign.

Todas as etiquetas inseridas pela ferramenta TagAlign, em qualquer um dos módulos (marcação ou alinhamento), estão de acordo com as especificações da linguagem de codificação

XML[4] (*Extensible Markup Language*). Essa padronização é muito importante, pois permite a reutilização desses textos codificados em projetos futuros ou mesmo no intercâmbio com outros grupos de pesquisa.

Além dos dois módulos citados, a TagAlign possui outras opções em seu menu principal (como pode ser observado na Figura 1): Ferramentas e Ajuda. As ferramentas disponíveis nesta versão da TagAlign (versão 1.1) são: geração do *corpus* paralelo e avaliação do *corpus* alinhado, as quais serão explicadas em detalhes na Seção 2.3. Há ainda uma Ajuda (ou *Help*) sobre as principais funcionalidades da TagAlign, apresentada resumidamente na Seção 2.4.

A entrada e a saída da TagAlign variam de acordo com o módulo utilizado. No módulo de marcação, a entrada e a saída são arquivos texto, sendo que o conteúdo do arquivo de saída será o mesmo do arquivo de entrada após o processo de marcação (ou desmarcação). No módulo de alinhamento, a entrada é, necessariamente, um par de textos paralelos e a saída pode ser um par de textos paralelos com marcações de correspondência entre as sentenças ou apenas um arquivo texto com as sentenças correspondentes dos dois textos.

A linguagem de programação escolhida para implementação da TagAlign foi o Borland Delphi 6 para Windows. Essa escolha se deu devido à adequação de tal linguagem aos propósitos da ferramenta.

A TagAlign é uma ferramenta fácil de ser utilizada e não exige do usuário nenhum conhecimento aprofundado em Marcação, Alinhamento de Textos Paralelos ou tradução Português-Inglês. Porém, no caso do alinhamento sentencial manual, um conhecimento mínimo dessas duas línguas é necessário para que as correspondências entre as sentenças dos textos fonte e alvo possam ser estabelecidas corretamente.

2.1 Marcação

A marcação, codificação ou etiquetagem de textos é um processo de inserção de marcações (códigos ou etiquetas) em partes específicas do texto com o intuito de acrescentar significado, diferenciar as partes marcadas, formatar o texto de acordo com uma padronização, entre outros.

---

[4] Em http://www.w3.org/XML/.

O módulo de marcação da ferramenta TagAlign se insere nesse contexto e é responsável pela marcação automática de fronteiras de parágrafos e sentenças e pelo suporte à marcação de partes específicas do texto com etiquetas pré-definidas pelo usuário. Esse módulo é acessado pela opção **Marcação** no menu principal, como mostrado na Figura 2.

Figura 2. Módulo de Marcação.

Nas subseções seguintes serão apresentadas em detalhes as duas variantes do módulo de marcação da TagAlign: a marcação automática de fronteiras de parágrafos e sentenças (Subseção 2.1.1) e a marcação manual de outras partes do texto (Subseção 2.1.2).

2.1.1 Marcação Automática de Fronteiras de Parágrafos e Sentenças

Ao selecionar a opção **Parágrafos/Sentenças (automática)** na tela da Figura 2, o usuário terá acesso a todas as funções referentes à inserção automática de marcações de fronteiras de parágrafos e sentenças.

No processo de marcação automática da TagAlign são inseridas etiquetas de início e fim de parágrafos (e) e/ou sentenças (<s> e </s>) e também uma etiqueta de identificação do texto inserida no seu início. Essa etiqueta, <text>, possui dois atributos para identificar o idioma no qual o texto foi escrito e o nome do arquivo que o contém: lang e id, respectivamente. Desta forma, por exemplo, um texto em português armazenado no arquivo art1.txt possuiria a seguinte etiqueta <text>:

```
<text lang=pt id=art1>
```

A etiqueta final </text>, correspondente à inicial <text>, também é inserida, no fim do texto.

As etiquetas text, p e s podem ser alteradas pelo usuário nas caixas de edição mostradas na Figura 3.

Figura 3. Marcação automática de fronteiras de parágrafos e sentenças.

Para determinar as fronteiras de parágrafos e sentenças o algoritmo responsável pela inserção das etiquetas leva em consideração os seguintes fatores:

- Parágrafos
  - Início – Considera-se início de parágrafo a existência de um caractere precedido por:

- início do arquivo. Ex:

  <text lang=pt id=art1>

  Hoje é dia 20 ...
- zero ou mais espaços e/ou caracteres de mudança de linha. Ex:

  Hoje é dia 20 ...

- Fim – Considera-se fim de parágrafo a existência de um caractere de mudança de linha (Enter) seguido por:
  - zero ou mais espaços e/ou caracteres de mudança de linha e um caractere diferente destes. Ex:

    Hoje é dia 20 ...
  - fim do arquivo. Ex:

    </text>

- Sentenças
  - Início – Considera-se início de sentença a existência de uma letra maiúscula precedida por:
    - um ou mais espaços e/ou zero ou mais caracteres de mudança de linha e uma pontuação indicativa de fim de sentença (.!?). Ex:

      . <s>Hoje é dia 20 ...
    - uma etiqueta de início de parágrafo. Ex:

      <s>Hoje é dia 20 ...
    - início do arquivo. Ex:

      <text lang=pt id=art1>

      <s>Hoje é dia 20 ...
  - Fim – Considera-se fim de sentença a existência de uma pontuação indicativa de fim de sentença (.!?) seguida por:
    - um ou mais espaços e/ou caracteres de mudança de linha uma letra maiúscula. Ex:

      .</s> A

      ou

?</s>

A ...

- uma etiqueta de fim de parágrafo. Ex:

  .</s>

- fim do arquivo. Ex:

  !</s>

  </text>

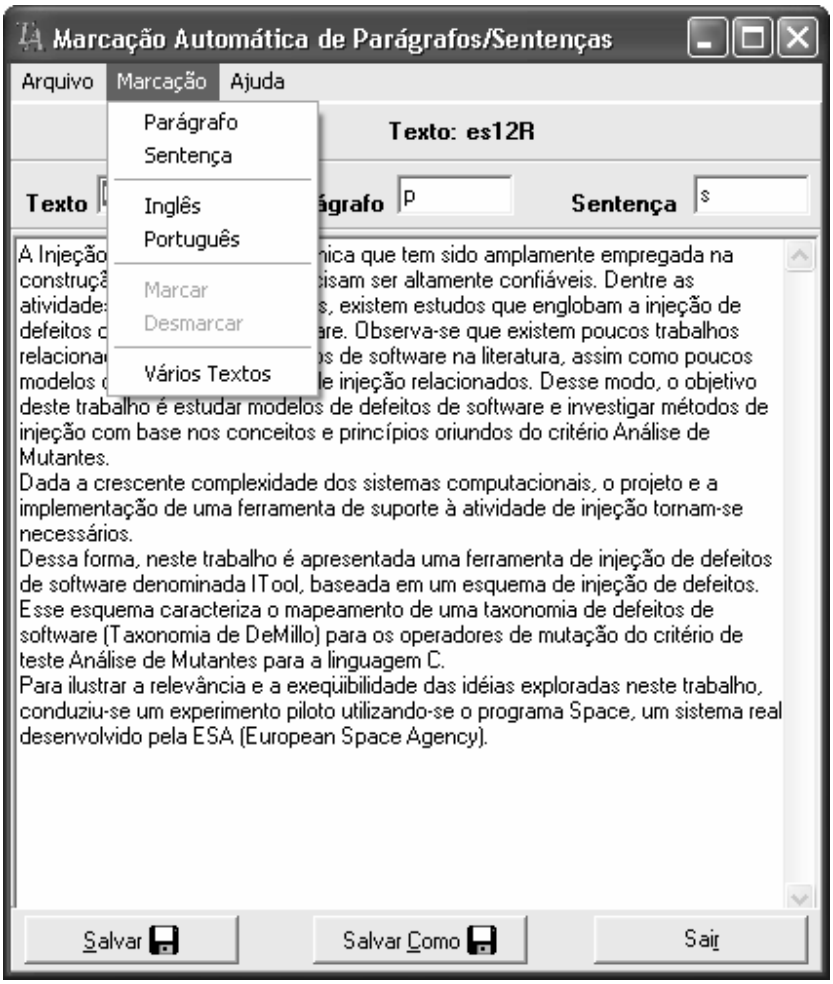


Figura 4. Opções do módulo de marcação.

O módulo de marcação possui sete opções, mostradas a seguir e visualizadas na Figura 4:

1. **Parágrafo** – habilita ou desabilita a inserção de etiquetas de início e fim de parágrafos (e , por exemplo).

2. **Sentença** – habilita ou desabilita a inserção de etiquetas de início e fim de sentenças (<s> e </s>, por exemplo).
3. **Inglês** – seleciona o idioma inglês como o idioma do texto a ser marcado.
4. **Português** – seleciona o idioma português como o idioma do texto a ser marcado.
5. **Marcar** – marca o texto aberto de acordo com as opções selecionadas. A opção **Marcar** só estará habilitada se estiverem selecionados: pelo menos uma das opções de marcação (Parágrafo e/ou Sentença) e apenas um dos idiomas (Inglês ou Português).
6. **Desmarcar** – retira todas as etiquetas existentes no texto, ou seja, todos os pares <x> e </x>, em que x é qualquer combinação de caracteres alfanuméricos.
7. **Vários Textos** – abre uma nova janela, mostrada na Figura 5, para a seleção do diretório no qual estão armazenados os arquivos com os textos a serem marcados (ou desmarcados). Apenas os arquivos .txt existentes no diretório escolhido são exibidos na tela; os arquivos com outro tipo de extensão são ocultados. O usuário pode selecionar quantos arquivos quiser. Porém, se nenhuma seleção explícita for feita, todos os textos armazenados em todos os arquivos .txt existentes no diretório serão marcados (ou desmarcados). A marcação é feita de acordo com as opções escolhidas pelo usuário.

Figura 5. Marcar/Desmarcar vários textos.

Assim, por exemplo, para marcar um texto em português com as etiquetas padrão de início e fim de parágrafos e sentenças, devem ser selecionadas as opções Parágrafo, Sentença e Português do menu Marcação, conforme mostrado na Figura 6 e o resultado será o apresentado na Figura 7.

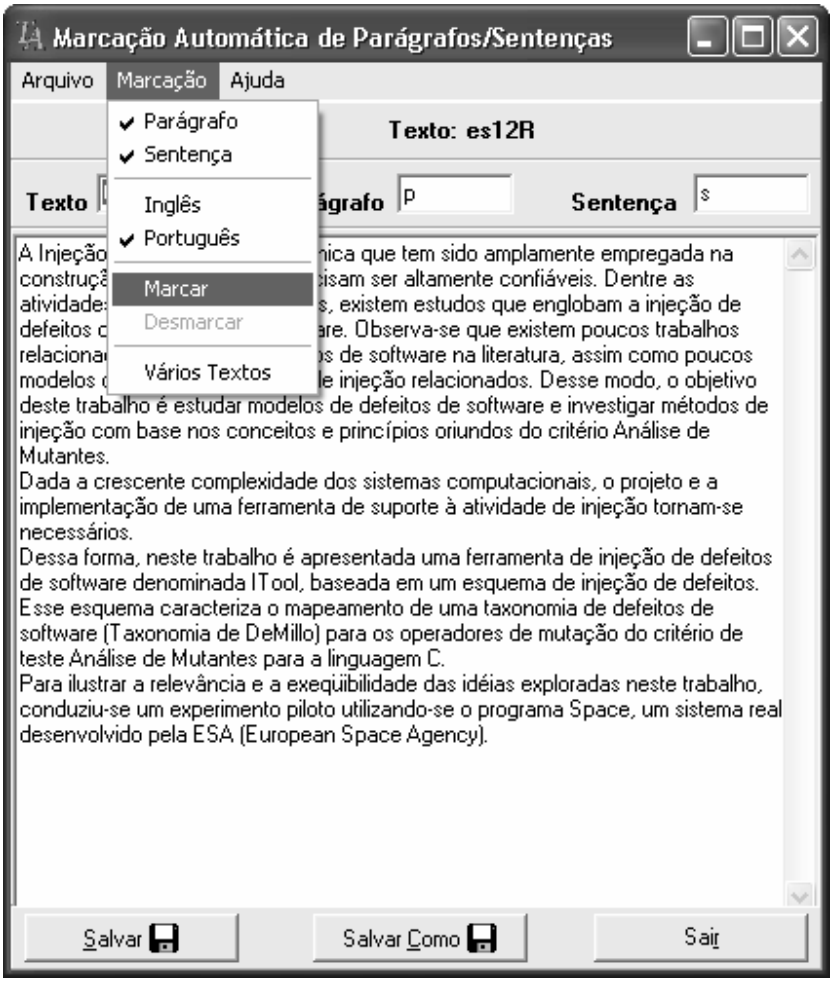


Figura 6. Texto em português antes da inserção de etiquetas de início e fim de parágrafos e sentenças.

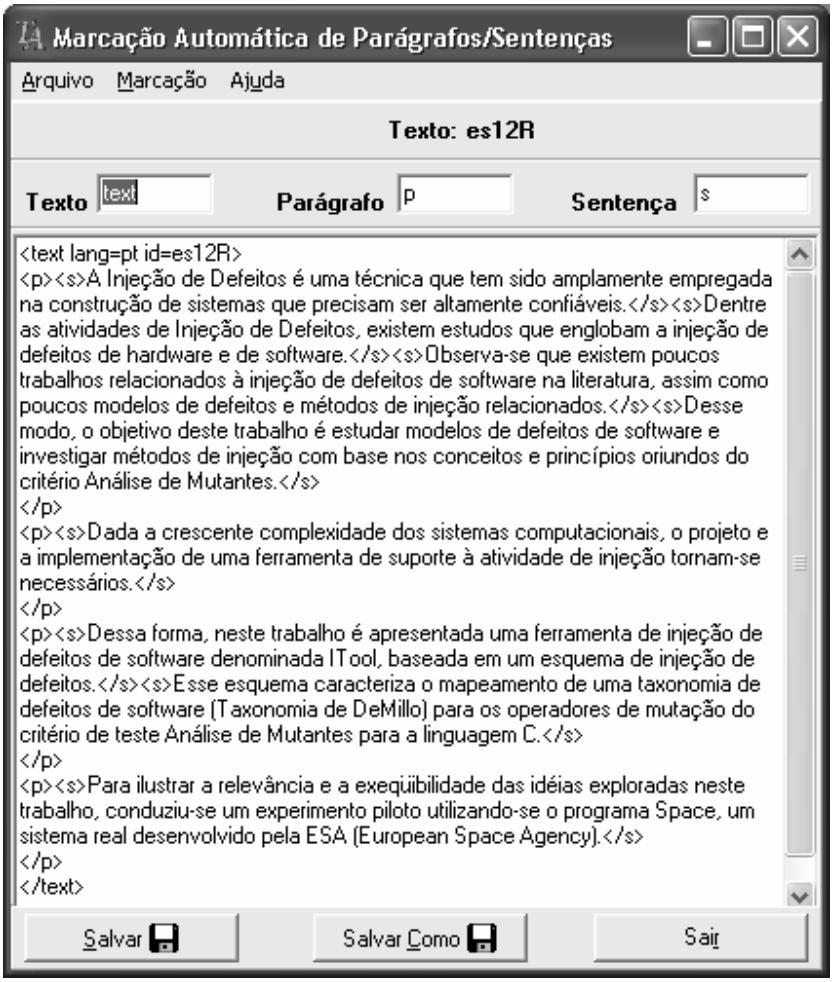


Figura 7. Texto em português após a inserção de etiquetas de início e fim de parágrafos e sentenças.

O processo inverso, ou seja, a remoção das etiquetas existentes no texto também é possível. Nesse caso, o resultado será o texto de entrada sem nenhuma etiqueta previamente existente, ou seja, sem qualquer par <x> e </x>.

2.1.2 Marcação Manual

Ao selecionar a opção **Outra (manual)** na tela da Figura 2 (menu **Marcação**), o usuário terá acesso a todas as funções referentes ao módulo de suporte à marcação manual. O módulo de marcação manual permite que o usuário escolha o conjunto de etiquetas a ser utilizado e o texto a ser marcado. Essa marcação é chamada de manual, pois é o usuário quem define o trecho de texto

a ser marcado e a etiqueta a ser inserida, sendo que a marcação só é realizada após o comando explícito do usuário. A Figura 8 apresenta a tela inicial do módulo de marcação manual.

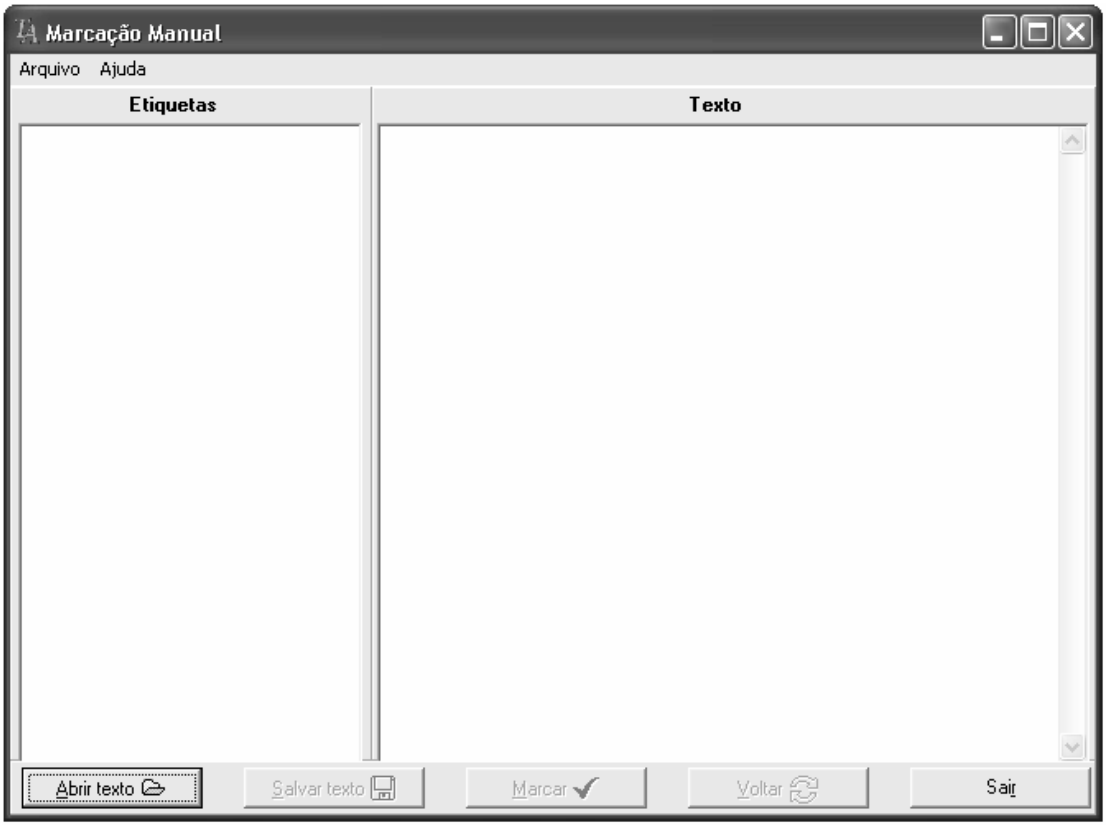


Figura 8. Tela inicial do módulo de marcação manual

Para iniciar a utilização desse módulo, o usuário deverá carregar o conjunto de etiquetas que deseja utilizar por meio do menu **Arquivo**, opção **Carregar etiquetas**, ou por meio do atalho **Ctrl + E**, como mostrado na Figura 9. O conteúdo do arquivo de etiquetas selecionado pelo usuário será mostrado na janela **Etiquetas**. O arquivo de etiquetas deve ser um arquivo do tipo Texto sem Formatação (.txt) editado de acordo com a sintaxe pré-definida para esse módulo do programa. Tal sintaxe, juntamente com instruções para a composição do arquivo de etiquetas, é descrita na subseção 2.1.2.1.

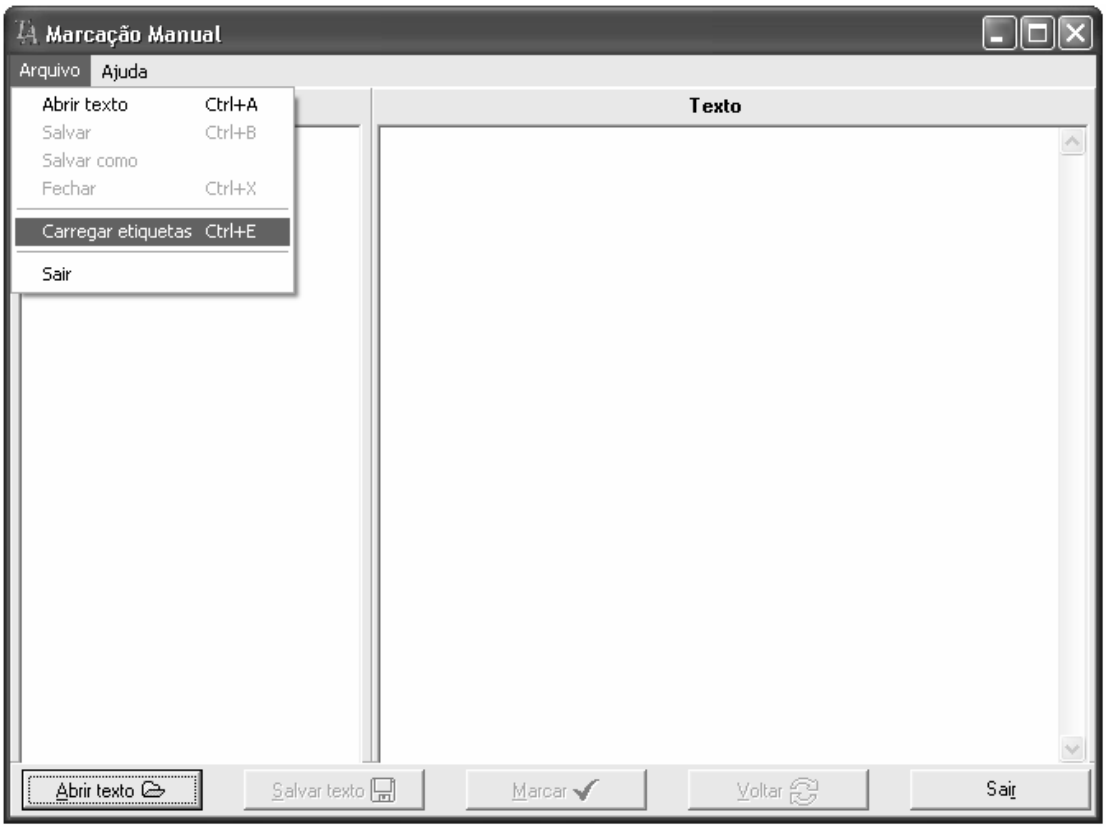


Figura 9. Tela do módulo de marcação manual mostrando as opções do menu **Arquivo**, com destaque para a opção **Carregar etiquetas.**

O texto a ser marcado também deve ser do tipo .txt e pode ser carregado para a janela **Texto** por meio do menu **Arquivo**, opção **Abrir texto**, por meio do botão **Abrir texto** ou por meio do atalho **Ctrl + A**, conforme também pode ser visualizado na Figura 9. Para fechar o arquivo e limpar a janela **Texto** deve ser utilizada a opção **Fechar** do menu **Arquivo** (atalho Ctrl + X). O texto também pode ser editado por meio dos comandos usuais de edição (digitação comum, Ctrl + C, Ctrl + V, Ctrl + Z). A Figura 10 mostra a tela do módulo de marcação manual com ambos os arquivos carregados, de etiquetas e texto.

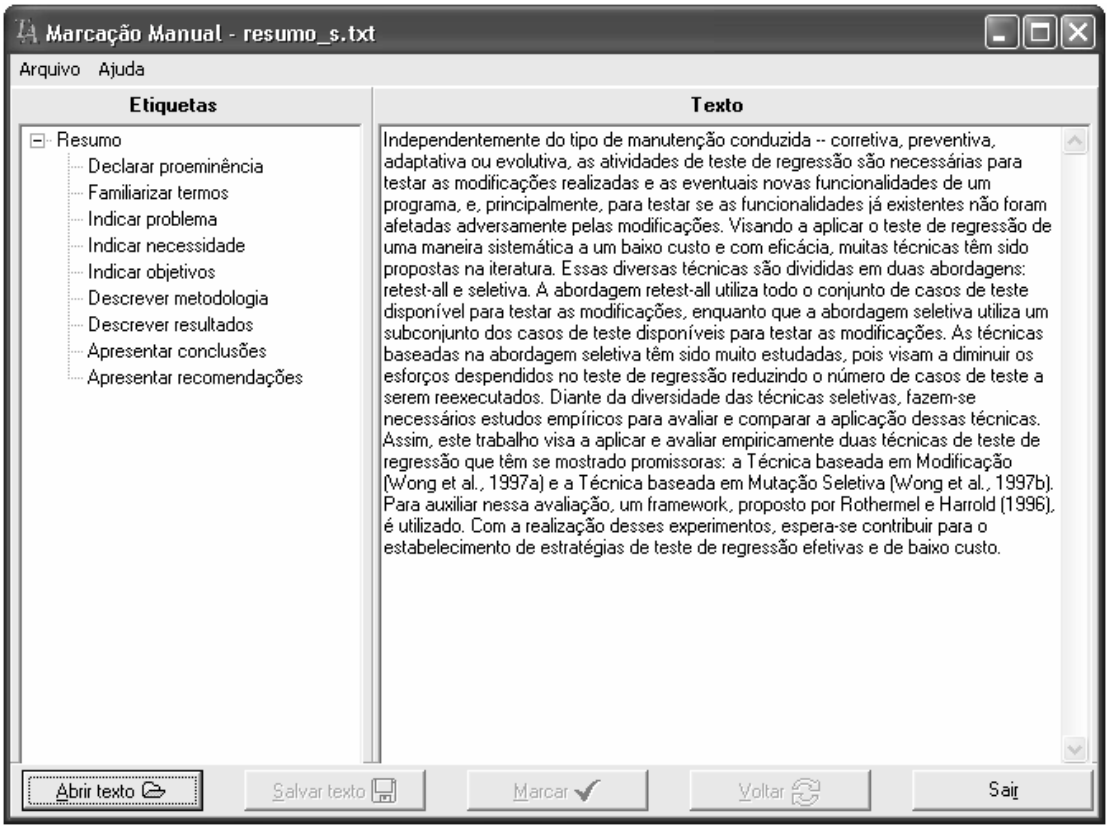


Figura 10. Tela do módulo de marcação manual com os arquivos de etiquetas e texto abertos.

Note na Figura 10 que o botão **Marcar** está desabilitado. Esse comando só é habilitado quando existe uma etiqueta selecionada e um texto carregado na janela **Texto**. No caso da figura, existe um texto carregado, mas nenhuma etiqueta foi selecionada. Essas condições são suficientes para habilitar o comando, mas para que uma etiqueta seja de fato inserida é necessário que o usuário selecione uma porção de texto (utilizando o *mouse* ou as teclas Shift + setas), mesmo que seja um espaço em branco, antes de acionar o comando **Marcar**.

Não existe uma ordem obrigatória entre a seleção do texto e da etiqueta. A seleção do texto funciona como em qualquer editor de texto. A seleção da etiqueta é realizada com o clique do mouse sobre a etiqueta desejada. Essa seleção é persistente, ou seja, só é modificada quando outra etiqueta é selecionada. Isso permite que várias marcações consecutivas sejam inseridas sem a necessidade de se clicar repetidamente na mesma etiqueta.

O comando Marcar pode ser acionado de 3 maneiras: por meio do **botão Marcar**, por meio do menu ***popup*** **Marcar** ou por meio do **clique duplo** sobre uma etiqueta. Em todos os casos a etiqueta só é inserida se houver uma porção de texto selecionada. A Figura 11 mostra um

exemplo em que a marcação está pronta para ser inserida. Note que a etiqueta “Declarar proeminência” está selecionada e a primeira sentença do texto também. Tanto o botão Marcar quanto a opção do *popup* estão habilitadas.

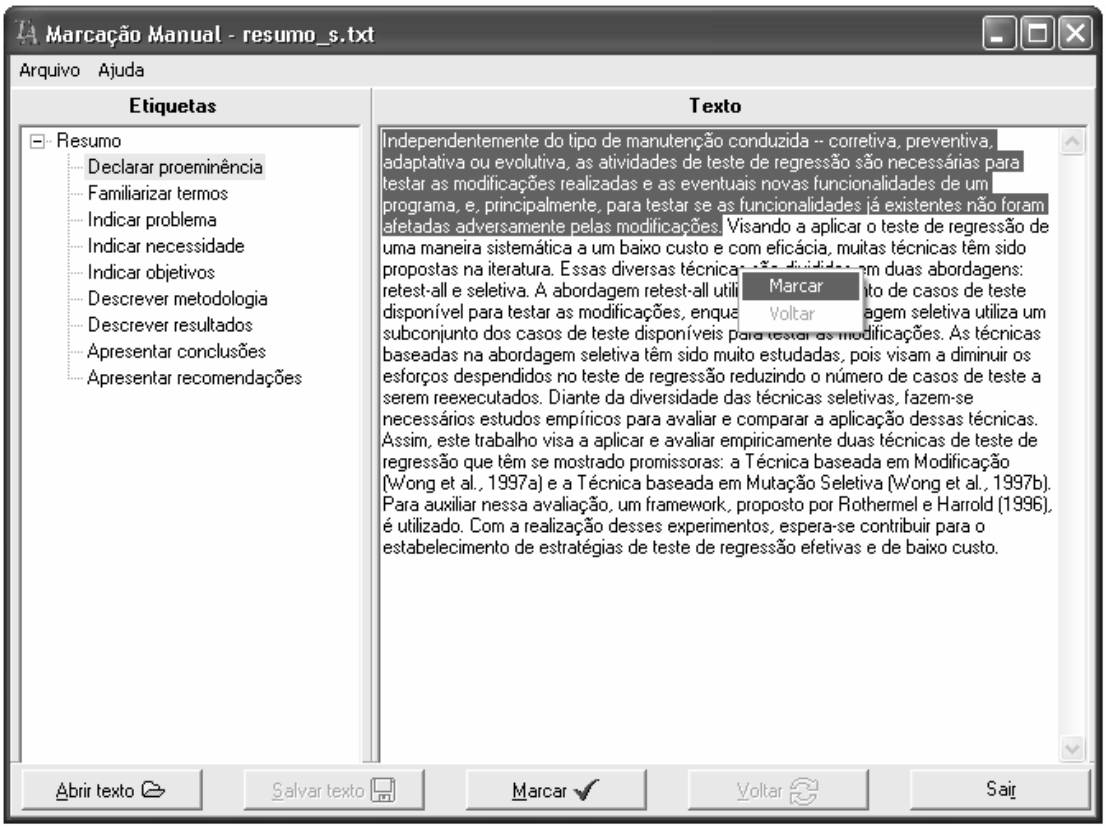


Figura 11. Tela do módulo de marcação manual mostrando a opção Marcar habilitada

Após ter sido inserida uma marcação, os botões **Voltar** e **Salvar texto** também serão habilitados, assim como as respectivas opções de menu. A opção **Voltar** permite que seja desfeita a **última** marcação realizada. A opção **Salvar texto** salva o texto que está sendo marcado com o nome já existente. Caso o usuário queira salvar o texto com outro nome pode utilizar a opção **Salvar como**, menu **Arquivo**. A Figura 12 mostra a tela do módulo de marcação manual com todas as opções habilitadas.

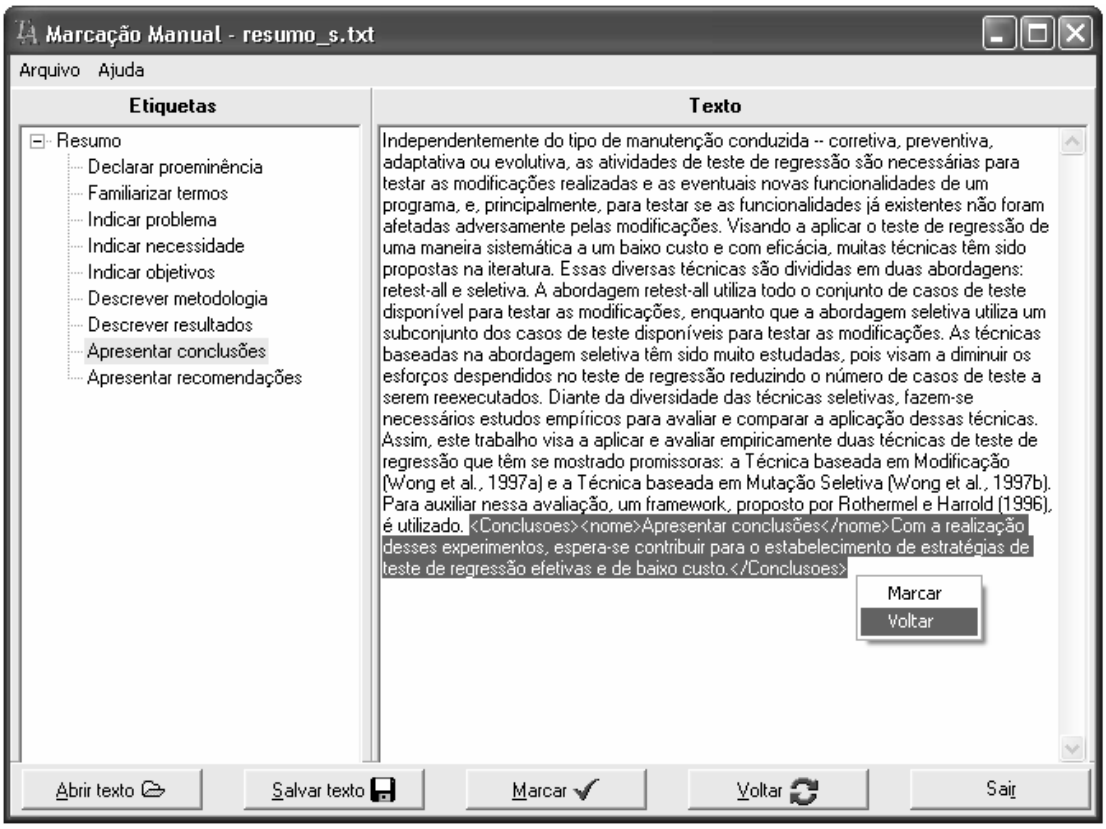


Figura 12. Tela do módulo de marcação manual com todas as opções habilitadas

Para retornar a tela principal do TagAlign pode-se utilizar o botão **Sair** ou a opção **Sair** do menu **Arquivo**. Caso o arquivo da janela **Texto** não tenha sido salvo, o programa irá perguntar se o usuário deseja fazer o salvamento. Caso contrário o módulo será encerrado e retornará a tela principal do TagAlign.

2.1.2.1 Definição do Arquivo de Etiquetas

O arquivo que define o conjunto de etiquetas a ser utilizado pela ferramenta deve ser do tipo Texto sem Formatação (.txt) e conter obrigatoriamente a definição de uma etiqueta por linha, seguindo a seguinte sintaxe:

etiqueta nível {*comentário*} {| *nome público*}

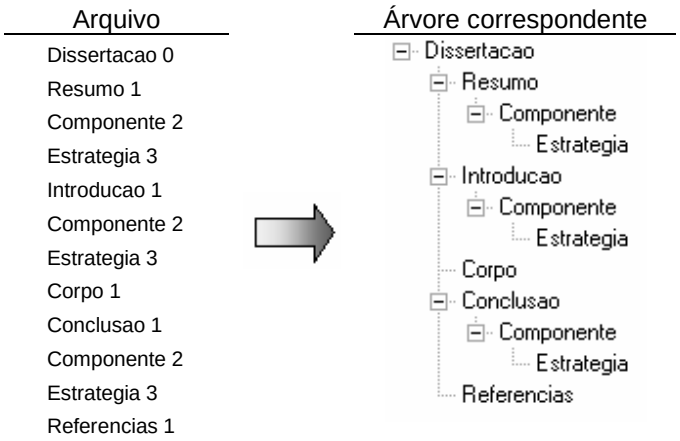
onde:

- *etiqueta* é uma string, sem espaços em brancos, e indica o nome da etiqueta que será inserida no texto marcado. A string será automaticamente delimitada pelos símbolos “<>” e “</>”, uma vez que as etiquetas seguem o padrão XML. Veja o exemplo:

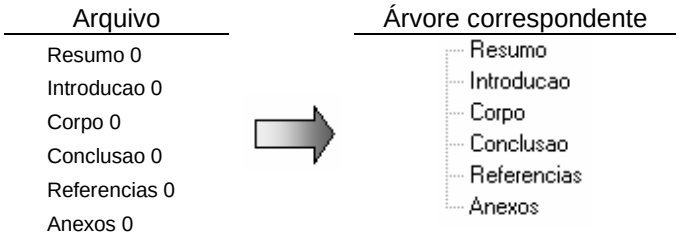
Definição da etiqueta		Marcação correspondente
Dissertacao 0	→	**<Dissertação>** texto **</Dissertação>**

- *nível* é um número inteiro e indica o nível da etiqueta dentro de uma estrutura do tipo árvore, sendo que o nível 0 (zero) indica a raiz da árvore (Exemplo 1). Caso se queira representar o conjunto de etiquetas como uma lista simples, o nível de todas as etiquetas deve ser 0 (Exemplo 2). A organização em árvore serve apenas para melhorar a visualização do conjunto de etiquetas e não influencia a marcação em si. É importante destacar que é recomendável iniciar a organização dos níveis em zero e seguir uma seqüência crescente de aninhamento dos níveis. Uma organização inconsistente de níveis poderá causar a leitura incorreta das etiquetas.

Exemplo 1 (Árvore):



Exemplo 2 (Lista):



- *comentário* é uma string **opcional**, podendo conter espaços em brancos, e indica um comentário que será inserido imediatamente após a etiqueta de abertura. Veja o seguinte exemplo de definição:

  proposito 0 ***<Nome>Indicar objetivos</Nome>***

  No exemplo, o *comentário* está representado em negrito. Dessa forma, o resultado da marcação com a etiqueta *proposito* será o seguinte:

  *<proposito>****<Nome>Indicar objetivos</Nome>*** texto texto texto *</proposito>*

  A definição de um comentário para uma etiqueta não modifica sua apresentação na janela **Etiquetas**, aparecendo apenas na marcação em si, ou seja, na janela **Texto** e no arquivo final.

- *nome público* é uma string **opcional**, sempre precedida pelo caractere |, podendo conter espaços em branco, e indica um "rótulo" que substituirá a etiqueta na sua apresentação na janela **Etiquetas**. O *nome público* serve para melhorar a visualização e a identificação das etiquetas, não influenciando na marcação em si, isto é, no texto marcado aparecerá a string definida como *etiqueta*. O rótulo definido como *nome público* será exibido apenas na janela **Etiquetas**. Veja o seguinte exemplo de definição contendo um *comentário* e um *nome público*:

  proposito 0 <Nome>Indicar objetivos</Nome> | Indicar objetivos

  O resultado da marcação com a marcação da etiqueta propósito será o seguinte:

  <proposito><Nome>Indicar objetivos</Nome> texto texto texto *</proposito>*

Note que o resultado é o mesmo apresentado no exemplo anterior, sobre o uso de *comentários*. No entanto, a apresentação na janela **Etiquetas** será diferente, pois em vez de ser mostrada a etiqueta real ─ *propósito*, aparecerá o rótulo *Indicar objetivos*, conforme representado na figura abaixo.

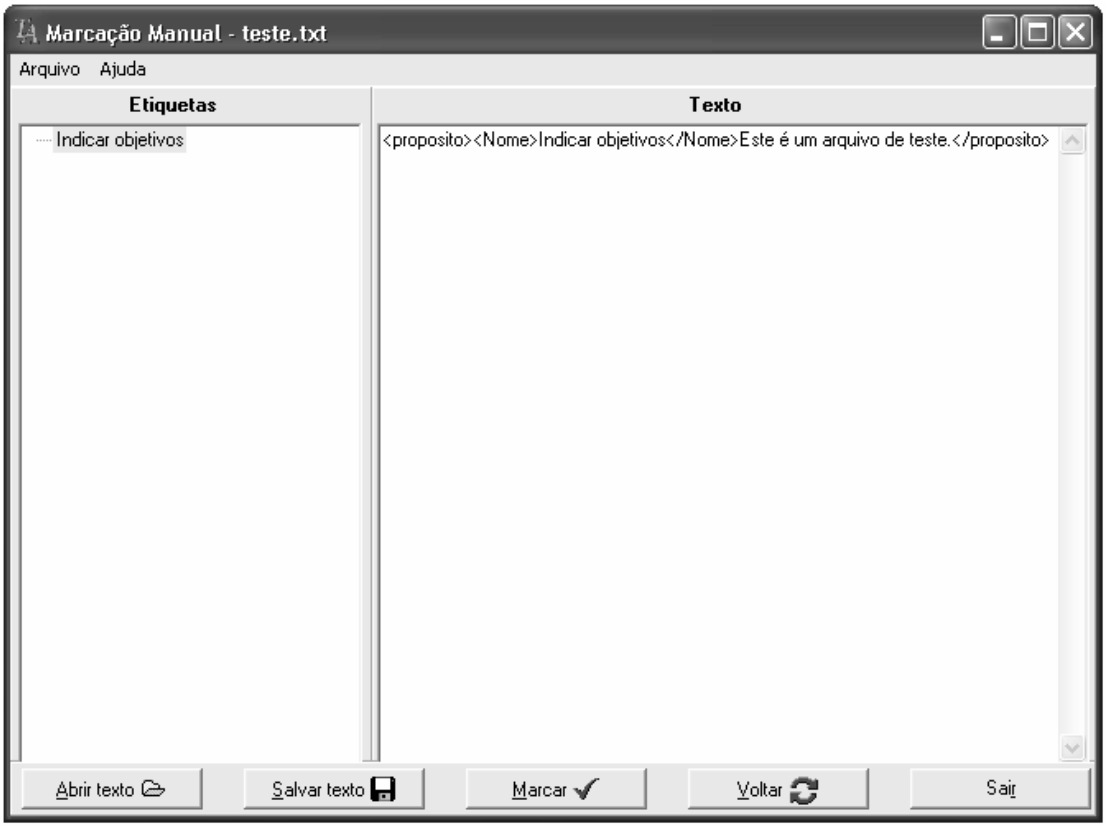


Figura 13. Exemplo mostrando a utilização de uma etiqueta com nome público

A utilização de *comentário* e *nome público* possibilita que etiquetas repetidas com comentários distintos sejam apropriadamente visualizadas e inseridas no texto. Veja o exemplo de definição abaixo:

Resumo 0
Contexto 1 <nome>Declarar proeminencia</nome> | Declarar proeminência
Contexto 1 <nome>Familiarizar termos</nome> | Familiarizar termos
Gap 1 <nome>Indicar problema</nome> | Indicar problema
Gap 1 <nome>Indicar necessidade</nome> | Indicar necessidade
Proposito 1 <nome>Indicar objetivos</nome> | Indicar objetivos
Metodologia 1 <nome>Descrever metodologia</nome> | Descrever metodologia
Resultados 1 <nome> Descrever resultados</nome> | Descrever resultados
Conclusoes 1 <nome>Apresentar conclusões</nome> | Apresentar conclusões
Conclusoes 1 <nome>Apresentar recomendações</nome> | Apresentar recomendações

Note que existem duas definições para a etiqueta *Contexto*: uma com o comentário *<nome>Declarar proeminência</nome>* e outra com o comentário *<nome>Familiarizar termos</nome>*. Essa definição permite que seja feita a marcação de dois tipos diferentes de *Contexto*. Para que a etiqueta correta seja identificada na janela **Etiquetas** são usados nomes públicos semelhantes aos comentários. A definição acima é exemplificada na Figura 14, sendo que os exemplos de marcação com as etiquetas *Contexto* aparecem destacadas em vermelho e azul. O destaque em cores é apenas ilustrativo.

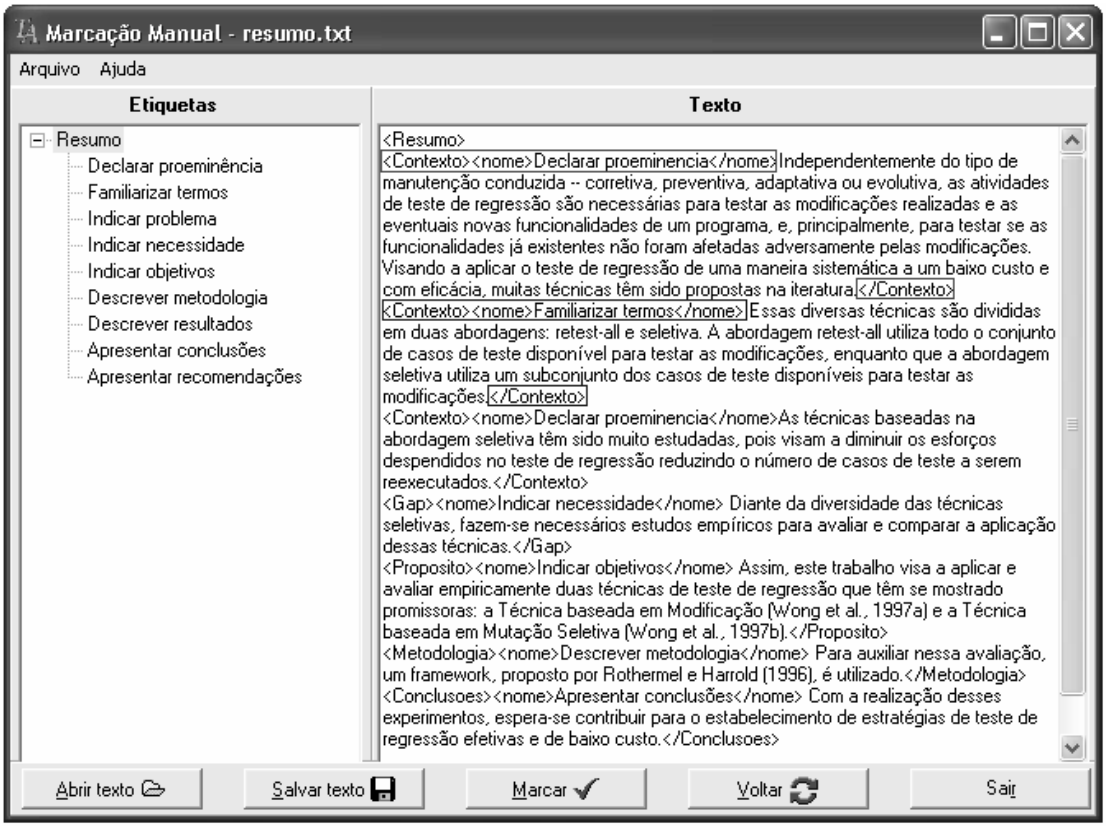


Figura 14. Exemplo de marcação usando etiquetas iguais com comentários diferentes

É importante destacar que, embora a o módulo de marcação manual também siga o padrão XML na grafia das etiquetas, o mesmo não garante que a hierarquia das etiquetas inseridas esteja de acordo com a hierarquia esperada em documento XML, uma vez que, nesse módulo, a inserção da marcação é totalmente controlada pelo usuário.

2.2 Alinhamento

O Alinhamento de Textos Paralelos é uma subárea de PLN que visa produzir recursos para diversos segmentos dessa área como: tradução automática, extração de terminologia, construção de léxicos bilíngües, entre outros. O ponto de partida para o alinhamento é um par de textos paralelos, ou seja, um texto escrito em um idioma (texto fonte) e sua tradução para outro (texto alvo). O alinhamento de um par de textos paralelos, ou um bitexto, é a identificação das partes correspondentes nesses textos. Por exemplo, no caso do alinhamento sentencial identifica-se qual(is) sentença(s) no texto alvo corresponde(m) a qual(is) sentença(s) no texto fonte.

O módulo de Alinhamento da TagAlign está inserido nesse contexto, mais especificamente no alinhamento sentencial de textos paralelos, com previsão de expansão para outros níveis de resolução como o alinhamento de palavras. Até o momento, a ferramenta possui apenas a opção de alinhamento sentencial manual, mas diversos métodos automáticos estão sendo implementados e testados no projeto PESA e poderão futuramente ser acrescentados à TagAlign.

O módulo de alinhamento pode ser acessado pela opção **Alinhamento** no menu principal, como mostrado na Figura 15.

Figura 15. Módulo Alinhamento.

A entrada do módulo de alinhamento é, necessariamente, dois arquivos do tipo .txt, um com o texto original e outro com a tradução deste para um idioma diferente. A saída será os dois textos de entrada com as marcações de correspondências entre suas sentenças ou um arquivo com

as sentenças de ambos alinhadas. Mais detalhes sobre o **Alinhamento Sentencial Manual** de textos paralelos são apresentados na próxima subseção.

2.2.1 Alinhamento Sentencial Manual

O alinhamento sentencial manual tem como objetivo auxiliar o usuário no processo de inserção de marcações de correspondência entre as sentenças do texto fonte e as sentenças do texto alvo e pode ser acessado pela opção **Sentencial Manual** no menu principal do módulo Alinhamento mostrado na Figura 15.

O alinhamento sentencial manual foi projetado para auxiliar a geração de um *corpus* de textos paralelos alinhados sentencialmente. Esse *corpus* possui papel fundamental na análise e avaliação dos métodos automáticos de alinhamento sentencial de textos paralelos que estão sendo implementados no projeto PESA.

No alinhamento sentencial, as correspondências entre as sentenças dos textos fonte e alvo são determinadas por dois atributos inseridos nas etiquetas de início de sentença: o atributo id, um identificador único da sentença, e o atributo corresp, que contém o(s) identificador(es) da(s) sentença(s) correspondente(s) à sentença identificada por id. Esses atributos foram escolhidos por estar de acordo com o padrão XML e também por já terem sido utilizados em trabalhos na área de alinhamento, como o de Santos e Oksefjell (2000) e o de Hofland (1996).

O valor do atributo identificador único da sentença (id) possui o seguinte formato:

[nome_arquivo]**.**[número_parágrafo]**.s**[número_sentença]

Por exemplo, a terceira sentença do segundo parágrafo do texto armazenado em art1.txt teria como valor de identificador único: art1.2.s3. A tela inicial do Alinhamento Sentencial Manual é mostrada na Figura 16.

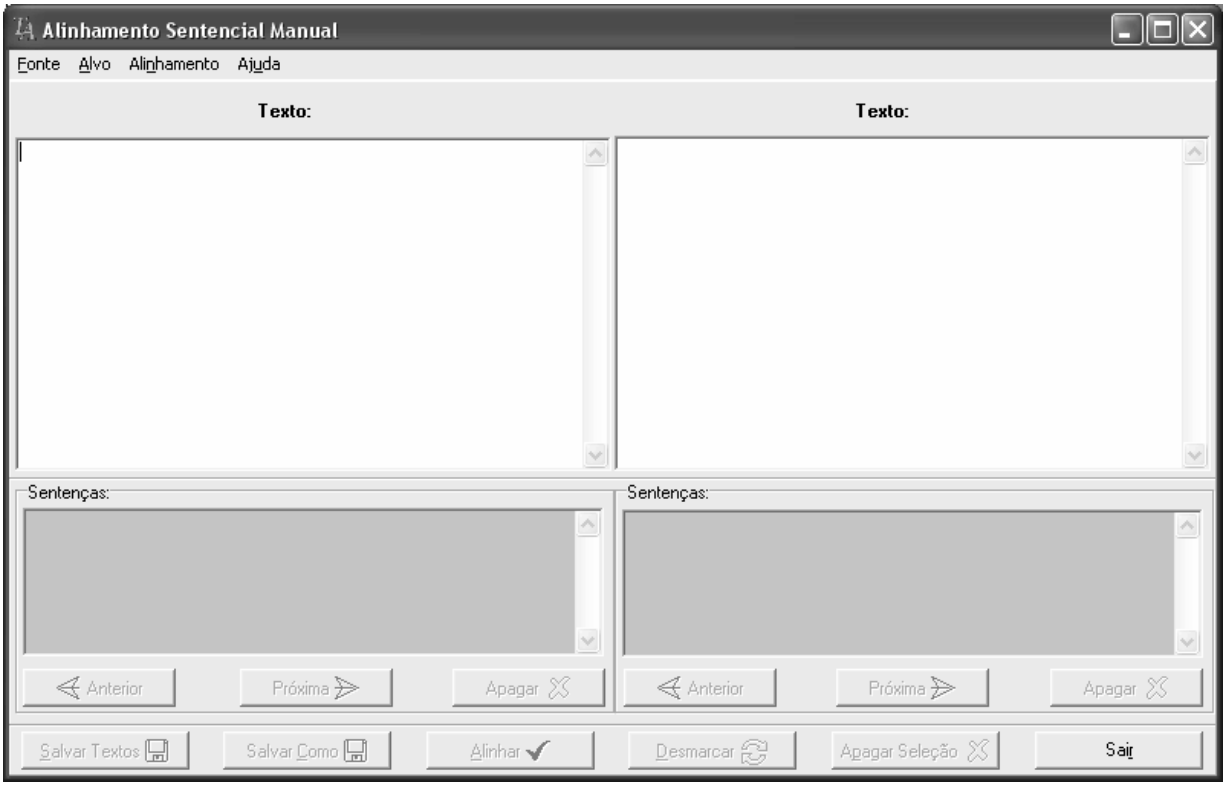


Figura 16. Tela inicial do Alinhamento Sentencial Manual.

Nessa figura observa-se que o menu do Alinhamento Sentencial Manual possui três opções:

1. **Fonte** – refere-se às atividades que podem ser desempenhadas com o arquivo do texto fonte: abrir, salvar e salvar como. Também há a opção de sair do módulo atual e retornar à tela principal da TagAlign.
2. **Alvo** – refere-se às mesmas atividades do item Fonte, mas com relação ao arquivo do texto alvo. Nesse item, a opção **Abrir** só estará habilitada se o texto fonte já estiver aberto. A escolha do texto alvo correspondente ao texto fonte já aberto fica a cargo do usuário, sendo sua responsabilidade escolher um texto que seja de fato a tradução do texto fonte.
3. **Alinhamento** – refere-se à visualização das sentenças alinhadas nos textos fonte e alvo de uma forma mais legível, como pode ser observado na Figura 21. Há também a opção de salvar as sentenças da forma mostrada em um arquivo de texto. A opção **Sentenças Alinhadas** só estará habilitada se todas as sentenças dos dois textos possuírem marcações de correspondência.

Após a abertura dos textos fonte e alvo utilizando as opções correspondentes do menu (Figura 16), os textos são mostrados lado a lado, cada um com uma área específica para exibir as sentenças selecionadas pelo usuário. A Figura 17 ilustra o que foi descrito.

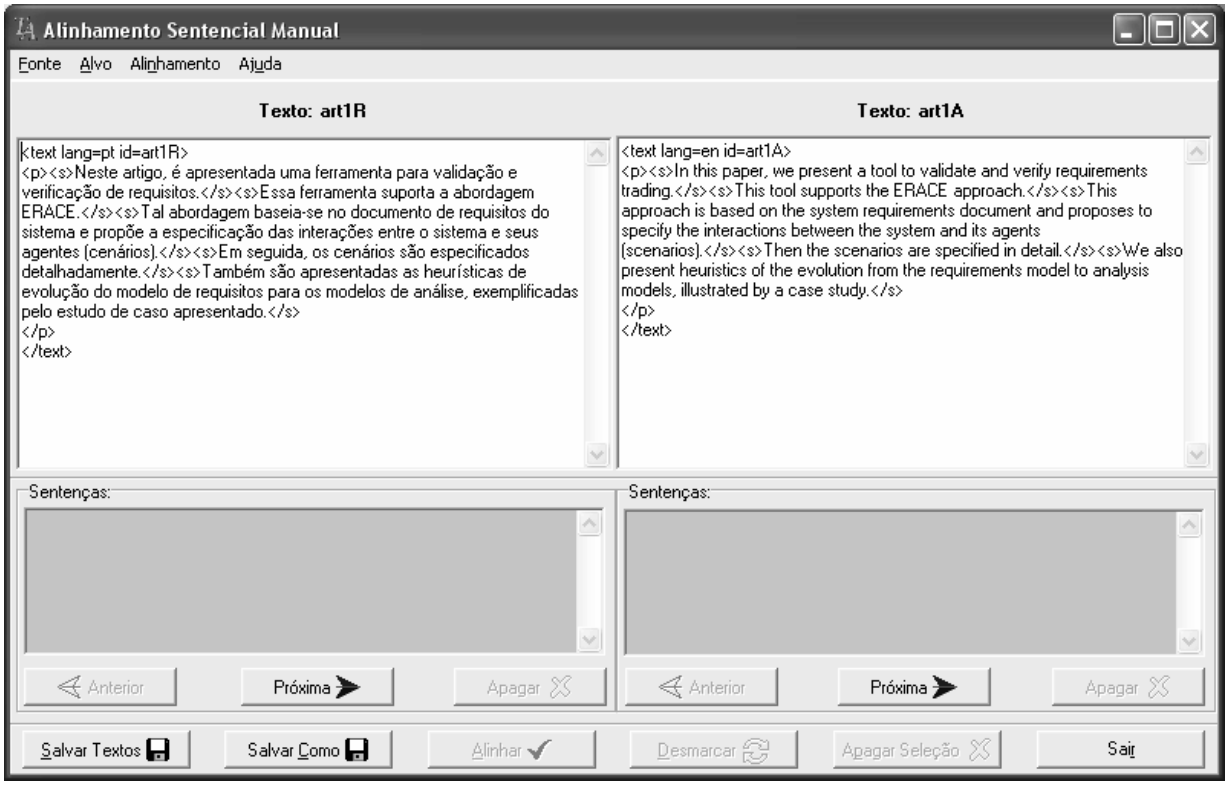


Figura 17. Arquivos fonte e alvo abertos.

O processo de alinhamento sentencial manual começa com a seleção de uma ou mais sentenças do texto fonte (ou alvo) e a seleção de zero ou mais sentenças do texto alvo (ou fonte). Para selecionar uma sentença basta clicar em qualquer lugar entre as suas etiquetas de início e fim (<s> e </s>, por exemplo). A sentença “clicada” é copiada para a área de sentenças selecionadas do texto correspondente e também é realçada no texto, como pode ser visto na Figura 18.

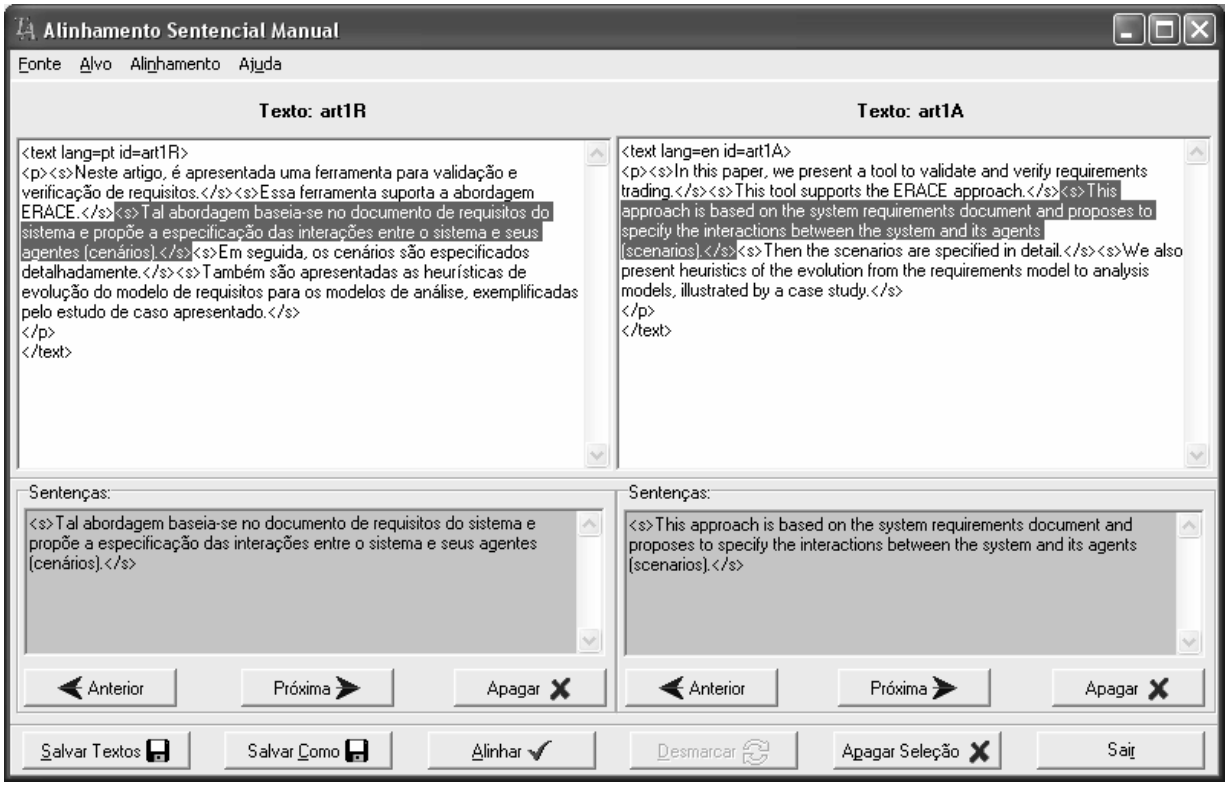


Figura 18. Seleção de sentenças.

A inserção de correspondências entre as sentenças selecionadas é feita clicando-se no botão **Alinhar**. Essa ação, considerando a situação mostrada na Figura 18, irá inserir os atributos id=art1R.1.s3 e corresp=art1A.1.s3 na etiqueta inicial da sentença selecionada no texto fonte e id=art1A.1.s3 e corresp=art1R.1.s3 na etiqueta inicial da sentença selecionada no texto alvo. As sentenças serão apagadas da área de sentenças selecionadas. Esse resultado é mostrado na Figura 19.

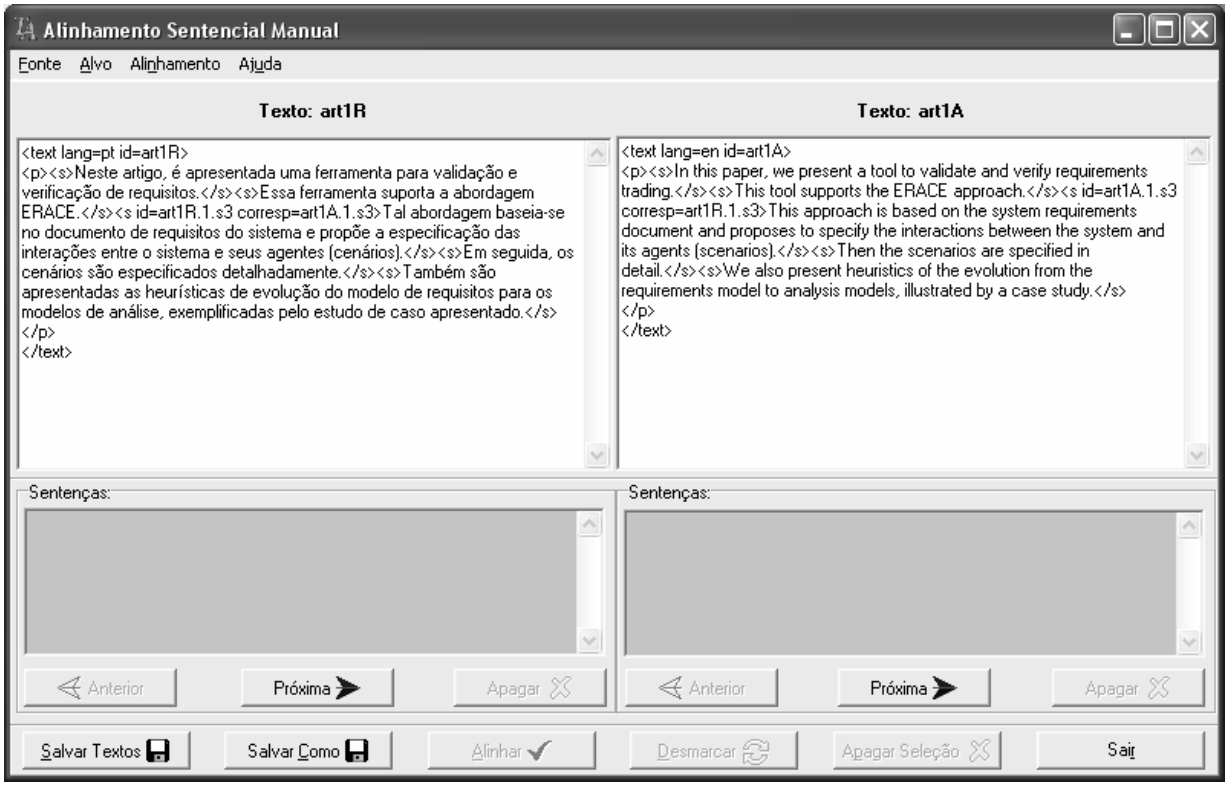


Figura 19. As sentenças selecionadas na Figura 18 foram alinhadas e reescritas nos textos.

A remoção das correspondências entre as sentenças segue um processo semelhante. A única diferença, nesse caso, é que após selecioná-las o usuário deve clicar no botão **Desmarcar** ao invés de Marcar. O resultado desse processo serão as sentenças selecionadas sem os atributos id e corresp em suas etiquetas iniciais.

A seleção das sentenças para a remoção das correspondências deve ser feita com muita cautela, pois qualquer deslize pode causar incoerência no alinhamento dos textos fonte e alvo. Ao selecionar uma sentença fonte o usuário deve, necessariamente, selecionar as correspondentes a ela no texto alvo. Esse cuidado é fundamental, pois a ferramenta não faz nenhuma verificação para garantir que os atributos sejam retirados em ambas e não apenas em uma delas.

Dessa forma, a remoção da correspondência entre a primeira sentença do primeiro parágrafo do texto art1R.txt e a primeira sentença do primeiro parágrafo do texto art1A.txt é ilustrada na Figura 20.

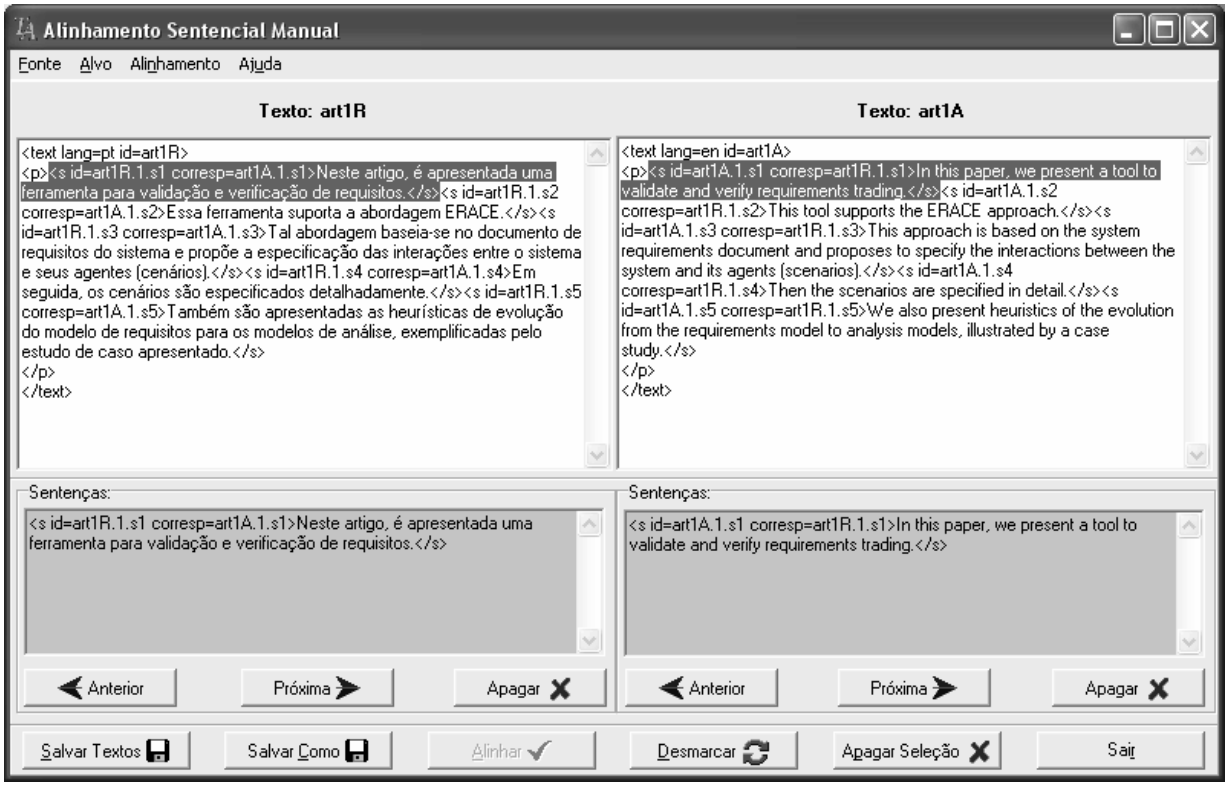


Figura 20. Seleção de sentenças para a remoção de correspondência. O usuário deve verificar se as correspondências serão removidas em ambos os textos quando este for o caso.

A última opção do alinhamento sentencial manual (Alinhamento → Sentenças Alinhadas) permite a visualização das sentenças alinhadas, uma após a outra, sem nenhuma marcação, como mostrado na Figura 21. As sentenças alinhadas, como aparecem nessa figura, podem ser salvas em um arquivo .txt para uma posterior conferência do alinhamento produzido.

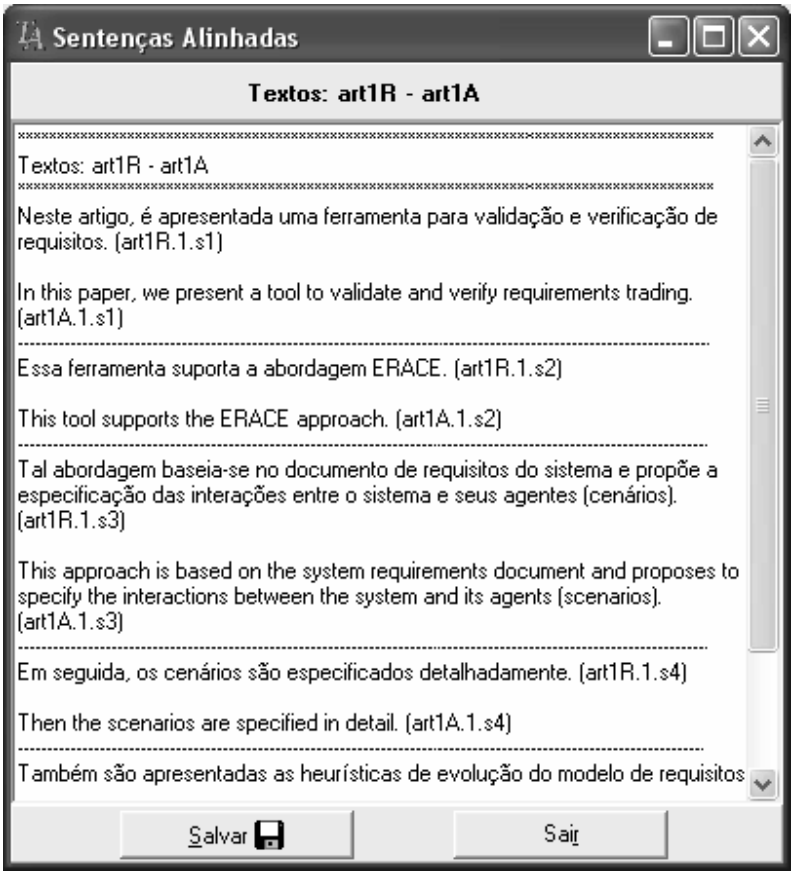


Figura 21. Visualização das sentenças alinhadas nos dois textos da Figura 19.

2.3 Ferramentas

Além dos dois módulos anteriormente citados, o de Marcação e o de Alinhamento, a TagAlign dispõe de algumas ferramentas de suporte que podem ser acessadas pela opção **Ferramentas** no menu principal (Figura 22). A primeira dessas opções, **Gerar Corpus Paralelo**, gera um arquivo texto com a identificação dos arquivos que formam cada par (texto fonte, texto alvo). A segunda, **Avaliar Corpus Alinhado**, compara cada bitexto alinhado com seu correspondente no *corpus* de referência calculando as métricas utilizadas na área de alinhamento para avaliar os métodos automáticos.

Figura 22. Ferramentas.

As duas ferramentas de suporte previamente citadas são apresentadas nas subseções 2.3.1 e 2.3.2, respectivamente.

2.3.1 Geração do Corpus Paralelo

Um *corpus* paralelo é um conjunto de textos paralelos dois a dois, ou seja, um conjunto de bitextos. Esse *corpus* é de grande importância para duas etapas do projeto PESA: o alinhamento e a avaliação conjunta de diversos bitextos. Por isso, foi implementada uma ferramenta para facilitar o processo de seleção dos textos que compõem cada um dos *corpora* paralelos. Tal ferramenta pode ser acessada escolhendo-se a opção **Gerar Corpus Paralelo** no menu Ferramentas mostrado na Figura 22.

A opção de gerar corpus paralelo cria um arquivo texto com o caminho completo (diretório\...\nome.txt) de cada um dos textos selecionados pelo usuário como fonte e alvo criando-se, assim, um meio mais fácil de manipular esse conjunto de textos em atividades como o alinhamento e a avaliação. É importante ressaltar que no arquivo de saída são armazenados apenas os caminhos para os arquivos que contém os textos fonte e alvo, não sendo tratadas situações geradas, por exemplo, pela remoção de um desses arquivos do local original.

A tela inicial dessa ferramenta é apresentada na Figura 23.

Figura 23. Geração do *corpus* paralelo.

Para gerar um *corpus* paralelo o usuário deverá, primeiro, selecionar os diretórios nos quais os arquivos com os textos fonte e alvo estão armazenados. Após a seleção, os arquivos .txt contidos nesses diretórios serão mostrados nas áreas nomeadas **Texto Fonte** e **Texto Alvo**, respectivamente. A partir daí o usuário pode criar os pares (bitextos) um a um selecionando os textos fonte e alvo e clicando na seta à direita da área **Texto Alvo** ou, se os textos aparecerem na mesma ordem nos dois diretórios, basta marcar a opção "Os textos paralelos aparecem na mesma ordem" e clicar na seta. O processo inverso – desfazer um par de textos paralelos – é efetuado selecionando-se o par de textos na área **Textos Paralelos** e clicando-se no botão "desfazer" abaixo da seta. A Figura 24 traz uma ilustração desse processo.

Figura 24. Geração de um *corpus* paralelo em andamento.

Há ainda a possibilidade de visualizar os textos disponíveis nas áreas **Texto Fonte** e **Texto Alvo**, para isso basta selecionar o texto desejado e clicar no botão "visualizar" acima da respectiva área de textos. O resultado desse processo para o texto fonte selecionado na Figura 23 é mostrado na Figura 25.

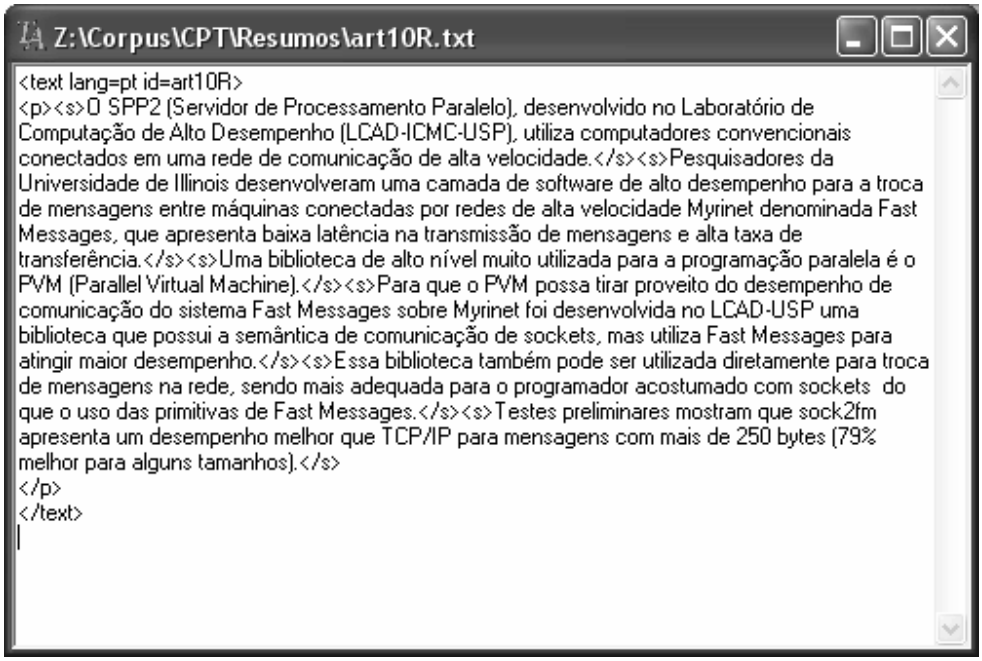


Figura 25. Visualização do texto fonte selecionado na Figura 23.

Futuramente pretende-se permitir que o alinhamento sentencial automático seja efetuado para o conjunto de textos paralelos presentes na área **Textos Paralelos** por meio do botão **Alinhar,** que, por enquanto, encontra-se desabilitado.

2.3.2 Avaliação do Corpus Alinhado

No projeto PESA, após o alinhamento dos textos paralelos por um método automático, o próximo passo é avaliar o desempenho do método comparando o *corpus* por ele alinhado com um outro considerado correto. Esse último *corpus*, denominado *corpus* de referência, é resultado de um processo manual de alinhamento dos mesmos textos contidos no *corpus* alinhado automaticamente.

Com o intuito de facilitar a avaliação dos métodos automáticos de alinhamento sentencial foi implementada uma ferramenta de suporte que pode ser acessada pela opção **Avaliar Corpus Alinhado** no menu **Ferramentas** (Figura 22). A tela inicial dessa ferramenta é apresentada na Figura 26.

Figura 26. Avaliação do *corpus* alinhado.

Como observado na Figura 26, a tela está dividida em três partes: *corpus* paralelo alinhado, *corpus* paralelo de referência e resultados. As duas primeiras contém os diretórios

(**Diretório**) e os arquivos (**Corpus Paralelo**) nos quais os *corpora* paralelos foram salvos, bem como os nomes de todos os arquivos nos quais textos paralelos que compõem os respectivos *corpora* estão armazenados (**Textos Paralelos**). A última parte (**Resultados**) contém os valores calculados para as métricas utilizadas na avaliação de métodos automáticos de alinhamento de textos paralelos.

São três métricas utilizadas para avaliar os alinhamentos gerados pelos métodos automáticos: *precision*, *recall* e *f-measure*. *Precision* é a porcentagem de alinhamentos corretos em relação a todos os que foram encontrados nos textos paralelos (1). *Recall* é a porcentagem de alinhamentos encontrados entre todos os possíveis (2). E *f-measure* é a medida de freqüência, calculada como o dobro da razão entre o produto *recall x precision* e a soma *recall + precision* (3) (Véronis & Langlais, 2000).

Dessa forma, a *precision* indica a capacidade do método de alinhamento em encontrar as correspondências corretas. Já o *recall* indica a capacidade do método de alinhamento em encontrar as correspondências. Por fim, o *f-measure* combina as duas anteriores em uma única métrica eficiente.

Assim:

$$precision = \frac{\text{NúmerodeAlinhamentosCorretos}}{\text{NúmerodeAlinhamentosPropostos}} \tag{1}$$

$$recall = \frac{\text{NúmerodeAlinhamentosPropostos}}{\text{NúmerodeAlinhamentosReferência}} \tag{2}$$

$$F = 2\frac{recall\ x\ precision}{recall + precision} \tag{3}$$

Existem duas formas de calcular essas métricas. A primeira é calculá-las para cada par de textos paralelos contido nos *corpora* e a segunda é calculá-las para conjuntos de textos paralelos selecionados aleatoriamente e considerar a média desses valores como a métrica final. A segunda opção pode ser considerada mais robusta uma vez que a aleatoriedade permite analisar o comportamento das métricas em uma situação do mundo real.

A motivação para a última opção surgiu da metodologia de avaliação de algoritmos que será aplicada aos métodos de alinhamento sentencial implementados no projeto PESA. Essa metodologia, muito empregada em Aprendizado de Máquina (AM), utiliza um estimador denominado *r-fold cross validation* para determinar o erro ou a precisão de um algoritmo (Freedman, Pisani & Purves, 1998 *apud* Baranauskas, 2001). No *r-fold cross validation* os bitextos são divididos aleatoriamente em *r* partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual a *n/r* bitextos. Um treinamento é efetuado com os bitextos contidos nos (*r*-1) *folds* e a hipótese induzida é testada no *fold* restante. Esse processo é executado *r* vezes e, em cada uma delas, um *fold* diferente é usado para teste. O erro é calculado como a média dos erros obtidos em cada um dos *r folds* e a partir daí outras métricas são calculadas para determinar se um dado algoritmo A é melhor do que outro algoritmo B.

A metodologia de avaliação dos métodos de alinhamento sentencial não faz parte do escopo deste relatório e poderá ser abordada em outra oportunidade. Por hora, a Figura 27 traz uma ilustração da avaliação de um *corpus* alinhado usando o método que calcula as métricas para

cada partição e a

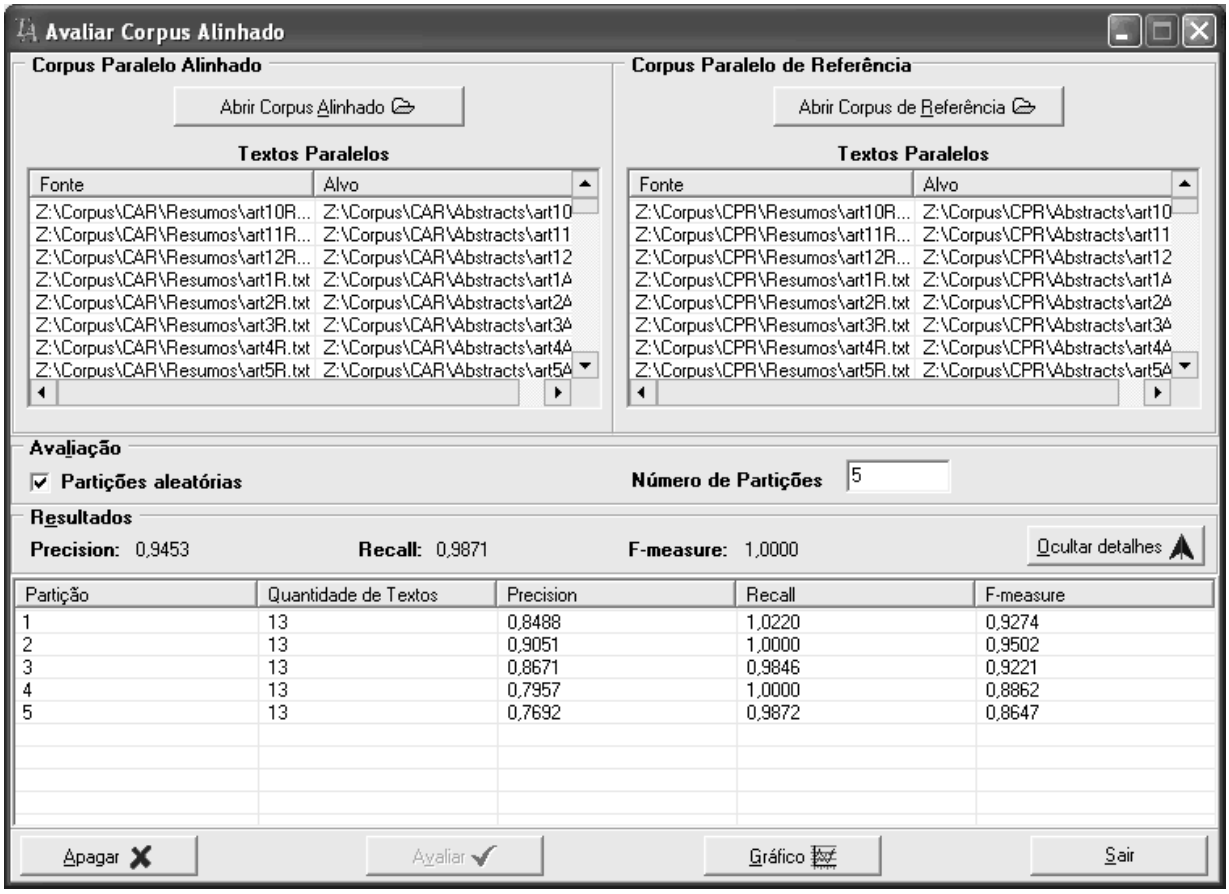


Figura 28

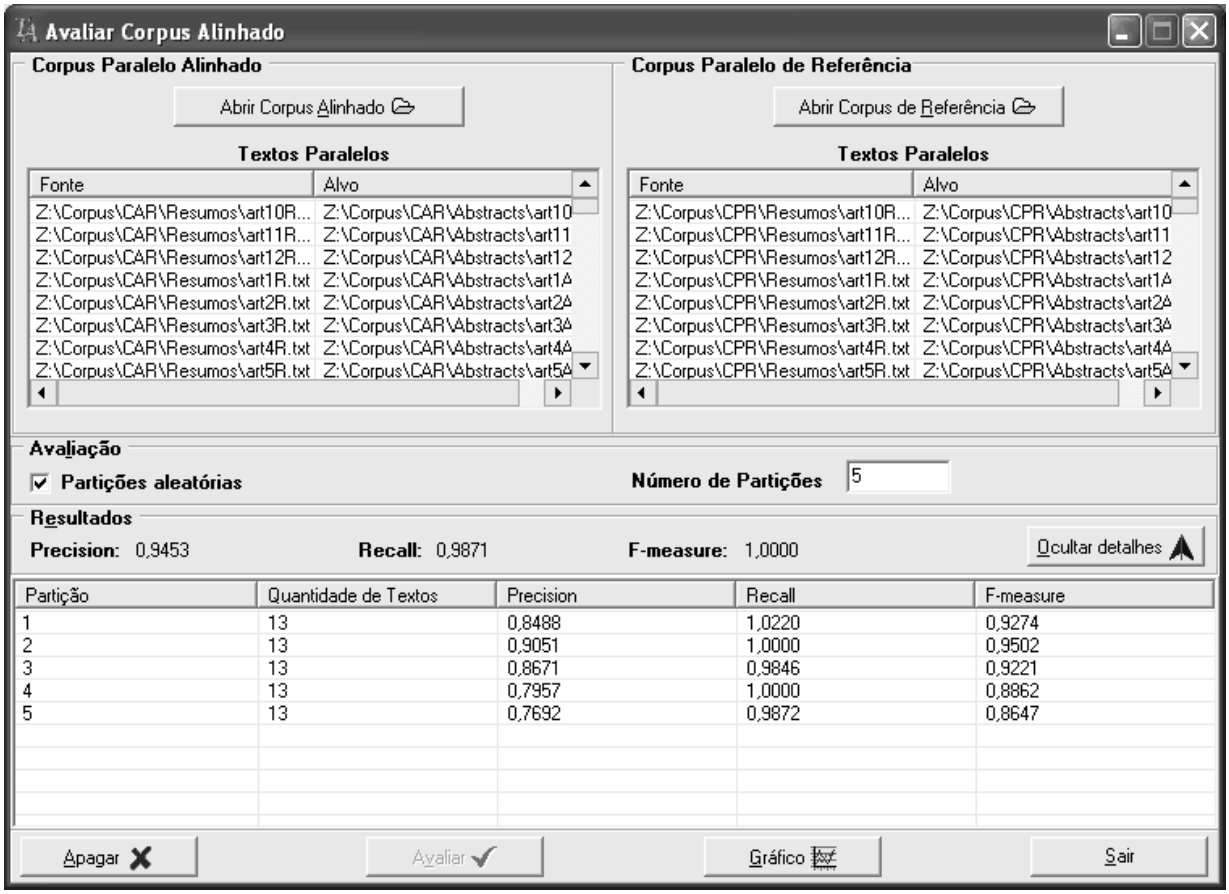


Figura 28, mostra os resultados detalhados dessa avaliação.

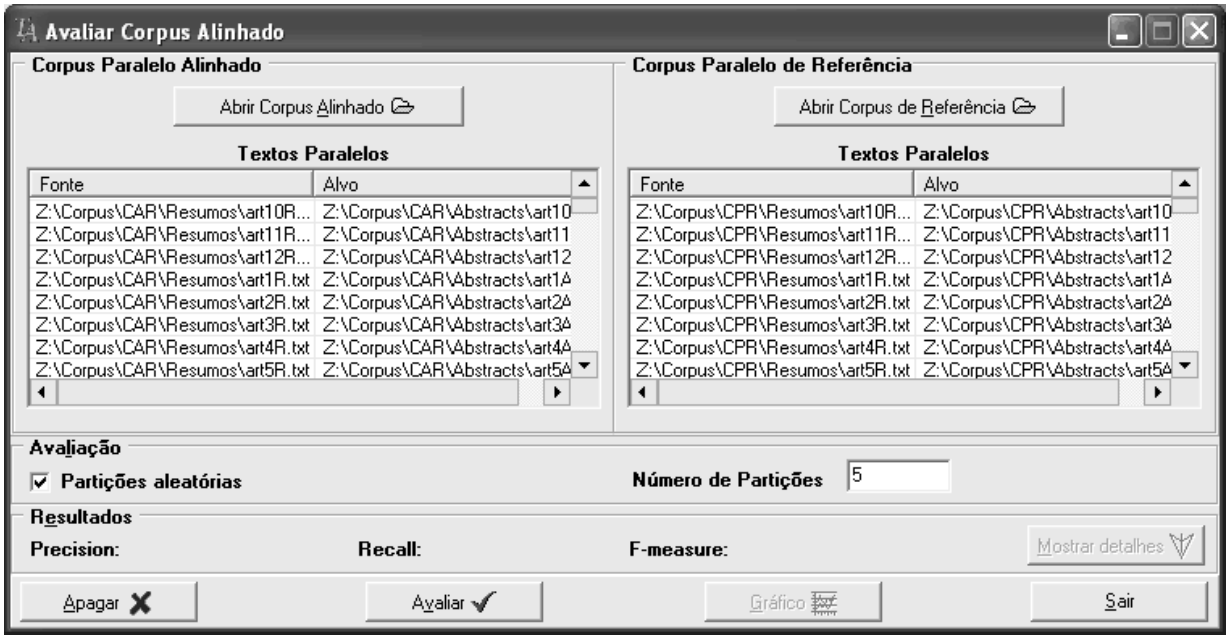


Figura 27. Avaliação para cada partição aleatória.

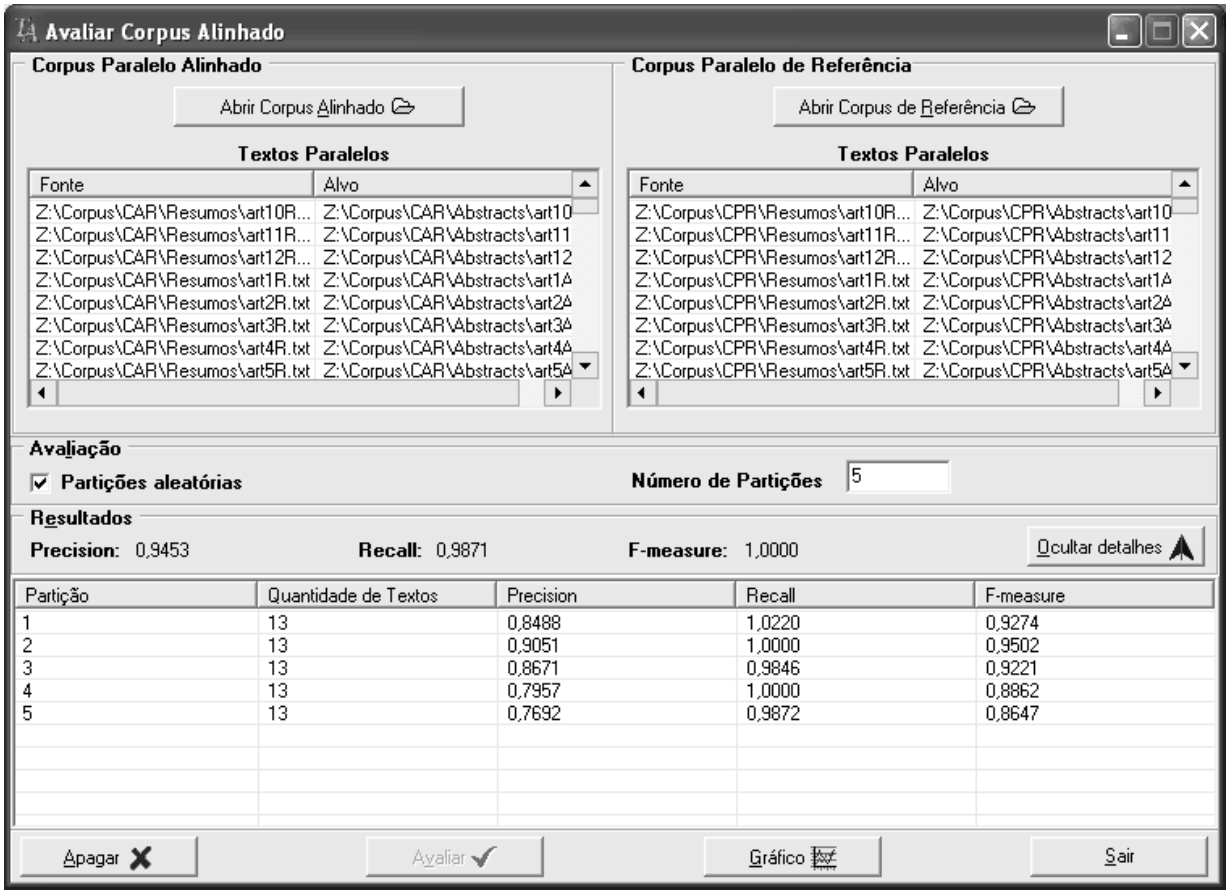


Figura 28. Resultados da avaliação para cada partição aleatória.

Os resultados apresentados nas figuras anteriores podem ser visualizados graficamente. A Figura 29 traz uma parte do gráfico gerado para os resultados da Figura 27. Nela os valores de *precision*, *recall* e *f-measure* para cada texto são mostrados em cores diferentes – verde, vermelho e azul, respectivamente – e podem ser separados selecionando-se os botões correspondentes. Há também a opção de salvar o gráfico em um arquivo com formato .bmp.

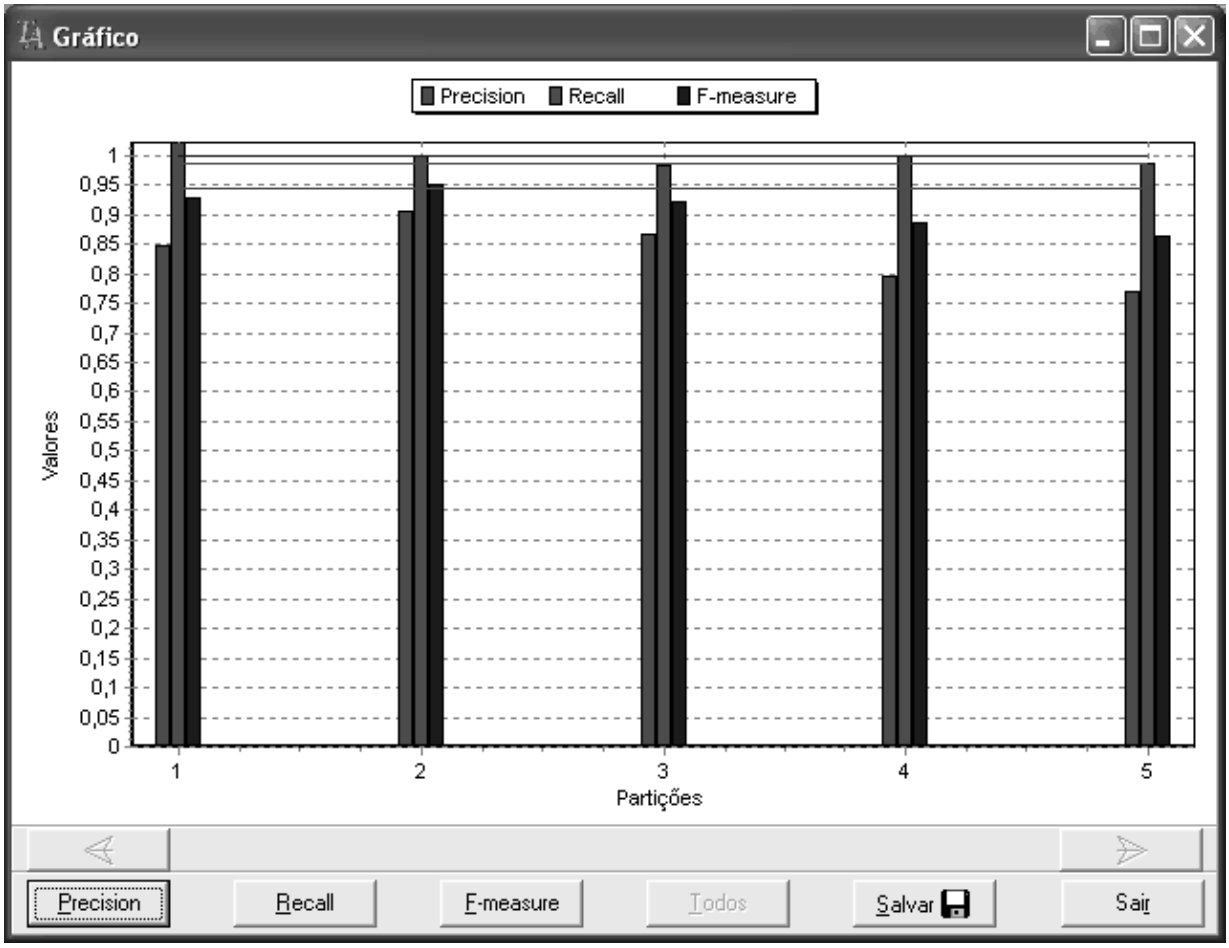


Figura 29. Gráfico dos resultados apresentados na Figura 28.

2.4 Ajuda

Uma outra funcionalidade da TagAlign é a Ajuda na qual o usuário poderá encontrar explicações sobre os principais tópicos da ferramenta. Essa funcionalidade, ainda em fase de implementação, visa facilitar a utilização da TagAlign e apresentar ao usuário os pontos mais importantes de todas as funcionalidades disponíveis.

3 Conclusão

A TagAlign é uma ferramenta de pré-processamento de textos com ampla funcionalidade para os projetos atuais e futuros do NILC. Foi projetada inicialmente para suprir apenas as necessidades dos projetos PESA e SciPo, mas com o decorrer de seu desenvolvimento foi generalizada e aperfeiçoada para ser útil também em projetos futuros na área de Processamento de Linguagem Natural.

A ferramenta possui basicamente duas funções: a marcação automática e manual e o alinhamento manual de sentenças em textos paralelos. Embora seja uma ferramenta simples, a TagAlign foi desenvolvida de uma forma que permite a adição de novas funcionalidades de acordo com as necessidades de outros projetos que venham a utilizá-la. Entre as extensões já previstas para a TagAlign podemos citar: a inserção de uma ou mais técnicas de alinhamento sentencial automático de textos paralelos e opções para o alinhamento manual e/ou automático em outros níveis de resolução, como o alinhamento de palavras.

A TagAlign pode, assim, ser vista como um passo inicial para uma ferramenta mais complexa de marcação e alinhamento de textos paralelos que estenderá ainda mais sua funcionalidade e utilidade.

4 Referências Bibliográficas

Baranauskas, J. A. (2001). Extração automática de conhecimento por múltiplos indutores. Tese de Doutorado. ICMC-USP, São Carlos.

Caseli, H. M. (2002). Alinhamento Sentencial de Textos Paralelos Português-Inglês. Monografia de Qualificação. ICMC-USP, São Carlos, Março.

Feltrim, V. D. (2002). Suporte Computacional à Escrita Científica em Português. Monografia de Quaficação. ICMC-USP, São Carlos, Março.

Freedman, D.; Pisani, R.; Purves, R. (Eds.) (1998). Statistics (Third ed.). W. W. Norton & Company.

Hofland, K. (1996). A program for aligning English and Norwegian sentences. *In:* Hockey, S.; Ide, N. & Perissinotto, G. (eds.), *Research in Humanities Computing*. Oxford: Oxford University Press. p. 165-78.

Santos, D.; Oksefjell, S. (2000). An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese. *In*: Proceedings of the 12th "Nordisk datalingvistikkdager". Trondheim, Departmento de Lingüística, NTNU, p.191-205.

Véronis, J.; Langlais, P. (2000). Evaluation of parallel text alignment systems: The ARCADE Project. *In*: Véronis, J., ed. *Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers, p.369-88.