Uma análise da ferramenta de tradução assistida por computador OmegaT versão 2.2.0_2

Por Francisco Araujo da Costa

Entre as opções gratuitas, o OmegaT é a ferramenta de tradução assistida por computador (CAT) mais popular do mercado. Esta análise apresenta as principais características, vantagens e desvantagens do aplicativo, além de oferecer recursos aos leitores interessados em experimentar o programa, incluindo algumas explicações sobre o funcionamento do programa e links com mais recursos.

O texto não analisa as características técnicas do programa enquanto obra da engenharia de software; seu leitor ideal é o tradutor ou localizador, não o programador. A análise também não pretende abranger todas as características do aplicativo, o que seria redundante com o manual do usuário. Finalmente, ela não se detém nas vantagens e desvantagens de se utilizar o programa para traduções que envolvam qualquer par específico de línguas, tratando o leitor como um tradutor genérico; exemplos em inglês e português são usados apenas por conveniência.

Características do OmegaT

O OmegaT é uma ferramenta de CAT, ou seja, um aplicativo que apoia o trabalho do tradutor e oferece ferramentas de apoio. É importante diferenciar as ferramentas CAT da tradução automática (TA), que usa algoritmos para gerar traduções sem a participação direta do tradutor. Ferramentas de CAT oferecem recursos como o uso de memórias de tradução, gestão terminológica, corretores ortográficos, buscas, alinhadores, estatísticas textuais, concordanciadores e dicionários. Esses recursos podem incluir o acesso a ferramentas de TA, como no caso do OmegaT, mas este não é seu objetivo principal. As ferramentas de CAT existem para aumentar a eficiência, velocidade e qualidade do trabalho do tradutor humano, não substituí-lo.

Na versão 2.2.0_2, o OmegaT é desenvolvido na linguagem de programação Java e distribuído sob os termos da Licença Pública GNU (GPL). Isso significa que o programa é distribuído gratuitamente[1] e o código-fonte está disponível para o público, permitindo que os usuários com proficiência em Java personalizem e adaptem o aplicativo ao seu bel-prazer. O uso de Java também facilita o uso do programa em múltiplas plataformas: Windows, Mac OS X e Linux, o que facilita a colaboração entre equipes heterogêneas. A equipe do projeto descreve o espírito descentralizado de contribuição e colaboração que cerca o projeto como o de "anarquia delegada".[2] Os usuários têm liberdade para criar vídeos e manuais explicativos, divulgar e distribuir o programa, desenvolver versões personalizadas, traduzi-lo para outras línguas etc. A equipe coordena e incorpora alterações

1 A última versão do programa – na verdade, toda e qualquer versão do programa – está disponível no site do projeto em <http://sourceforge.net/projects/omegat> ou em <www.omegat.org>.

2 Ver http://www.omegat.org/en/philosophy.html

ao código-fonte, de modo que nem todas as contribuições se tornam, por assim dizer, canônicas. Os usuários podem sugerir mudanças através do processo de RFE (Request for Enhancement ou Request for Feature Enhancement).[3] Algumas funções são desenvolvidas gratuitamente por voluntários, outras financiadas por doações ao projeto como um todo e outras ainda patrocinadas por indivíduos e empresas que desejam o desenvolvimento preferencial de alguma melhoria específica.

O programa traduz arquivos de uma série de formatos, os principais para os fins desta análise sendo texto puro e XML. Em outras palavras, o programa trabalha com arquivos do bloco de notas, páginas da Internet, legendas criadas no formato SRT, documentos do OpenOffice.org (conhecido como BrOffice.org no Brasil) e Microsoft Office 2007-2010, entre outros. Documentos criados com as versões anteriores do Microsoft Office (extensão .doc em vez de .docx, .ppt em vez de .pptx, etc.) precisam antes serem convertidos para formatos aceitos pelo programa.

O trabalho de tradução é relativamente simples e direto: o usuário cria um projeto, constituído de uma pasta e arquivos identificadores, usando o comando "Projeto → Novo"; importa os arquivos a serem traduzidos usando a interface do programa ou salvando-os direto no diretório de arquivos fonte designado no menu de criação do projeto; traduz os segmentos de texto no próprio OmegaT, sem nunca precisar interagir com o aplicativo que criou o arquivo original (ou mesmo possuí-lo no seu sistema, se for o caso); depois de finalizar, a tradução, usa o comando "Projeto → Criar documentos traduzidos" para criar versões traduzidas dos documentos no diretório de arquivos traduzidos designado no momento da criação do projeto. Se o usuário não souber realizar nenhum outro passo além destes, ele ainda será capaz de usar o aplicativo nas suas atividades de tradutor.

Isso seria, é claro, um grande desperdício. O programa permite a utilização de múltiplos glossários e dicionários externos, memórias de tradução de outros projetos, correção ortográfica, tradução automática e mecanismos de busca, além de oferecer estatísticas textuais elementares e customização da interface gráfica. Descrever cada um desses recursos em detalhe está além do escopo desta análise, mas ofereço a seguir uma breve explicação dos mais salientes.

Os glossários são arquivos de texto puro divididos por tabulação, com o conteúdo organizado no seguinte formato:

> palavra ou expressão na língua de origem<tabulação> palavra ou expressão na língua de chegada<tabulação> comentário

Se o usuário tem ou recebe um glossário na forma de uma planilha eletrônica, como o Microsoft Excel, basta copiar o conteúdo para um editor de texto; as divisões de colunas se transformam em tabulação. Os arquivos de glossário devem ser salvos no diretório de glossários designado no

3 Gerenciado no endereço <http://sourceforge.net/tracker/?atid=520350&group_id=68187&func=browse>

momento da criação do projeto, usando a extensão .tab ou .utf8.[4]

Quando o usuário está em um segmento que contém a palavra ou expressão na língua de origem, o conteúdo relevante do glossário é mostrado na respectiva janela do programa (Figura 1, janela inferior direita). Por exemplo, se o segmento é *The book is on the table* e um dos arquivos de glossário contém a linha "book<tabulação>livro<tabulação>Exceto às quartas-feiras", a janela mostra o texto:

book = livro

1. Exceto às quartas-feiras

Termos presentes no glossário mas ausentes no segmento não aparecem, diminuindo a poluição visual e mostrando ao usuário apenas o que é relevante em cada momento. O programa ignora a diferença entre maiúsculas e minúsculas nos glossários, mas também interpreta qualquer diferença de grafia como indicando uma palavra diferente.

Caso o segmento seja semelhante a algum outro que o usuário já traduziu, o segundo é apresentado na janela "Correspondências imperfeitas" (Figura 1, janela superior direita), acompanhada da sua respectiva tradução. Todas as palavras diferentes entre um e outro são marcadas em azul para a conveniência do usuário. O aplicativo mostra até cinco correspondências imperfeitas, por ordem de semelhança com o segmento em questão; a semelhança é determinada pela porcentagem de palavras do segmento antigo que corresponde às do atual. Infelizmente, quando a diferença é a ausência de uma palavra ou sua troca de ordem, nenhuma marcação é apresentada, mas esses pontos afetam o escore do segmento. Nosso segmento exemplo, *The book is on the table*, poderia apresentar as seguintes correspondências imperfeitas.[5]

The book is on the table!

O livro está sobre a mesa!

THE PEN IS ON THE TABLE

A CANETA ESTÁ SOBRE A MESA

The book is in the fire

O livro está na fogueira

Cada item seria acompanhado da sua respectiva porcentagem de correspondência. O primeiro seria apresentado como 100%, mas é considerado imperfeito devido à diferença de pontuação. O segundo é apresentado antes do terceiro por ser mais próximo do original, mas observe que o uso de

4 A diferença em extensão tem a ver com a codificação do arquivo. Se o leitor não sabe o que isso significa, então muito provavelmente poderá ignorar a diferença e salvar o glossário no Bloco de Notas no formato .tab.

5 Nos primeiros dois itens acima, a palavra "table" estaria marcada em verde, destacando uma diferença de pontuação. Essa diferença foi eliminada por uma questão de simplificação.

maiúsculas e minúsculas não interfere no cálculo. As correspondências podem ser selecionadas e inseridas automaticamente com um comando do menu “Editar” ou o uso de uma tecla de atalho.

Observe que as correspondências imperfeitas só são apresentadas se já foram traduzidas antes pelo usuário anteriormente e se encontram na memória do projeto ou se estão em algum dos arquivos de memória de tradução[6] salvos no diretório de arquivos de tradução designado no momento de criação da projeto. As correspondências imperfeitas não têm nenhuma relação com os glossários e dicionários.

O programa também está integrado com três ferramentas automatizadas de tradução: Google Translate, Apertium e Belazar. Assim, se o usuário ativar a opção, sempre que um novo segmento é aberto, o aplicativo envia o texto para o serviço escolhido e recebe de volta a tradução do serviço. O texto fica disponível em uma das janelas do programa, marcada “Machine Translation” e pode ser inserida com um comando do menu ou o atalho Ctrl + M. Se a opção estiver desativada, o OmegaT 2.2.0_2 não se comunica com nenhum dos três recursos. Atualmente, não há nenhuma configuração que permita que a tradução automatizada seja inserida automaticamente assim que o segmento é aberto, como é possível realizar com correspondências imperfeitas.

Para o tradutor que utiliza algum desses recursos regularmente, a vantagem da opção é óbvia, mas não está claro o que os provedores dos três serviços de tradução automática fazem com os dados recebidos. Usuários que lidam com informações confidenciais, segredos comerciais ou simplesmente não querem que seus textos-fonte sejam arquivados por terceiros devem desabilitar essa opção. Questões relativas ao uso desses serviços e seu impacto futuro na tradução fogem ao escopo desta análise; por ora, basta apresentar a presença dessa função.

6 O OmegaT utiliza o formato Translation Memory eXchange (TMX), administrado pela Localization Industry Standards Association (LISA). Para mais informações, ver o site da instituição em http://www.lisa.org/

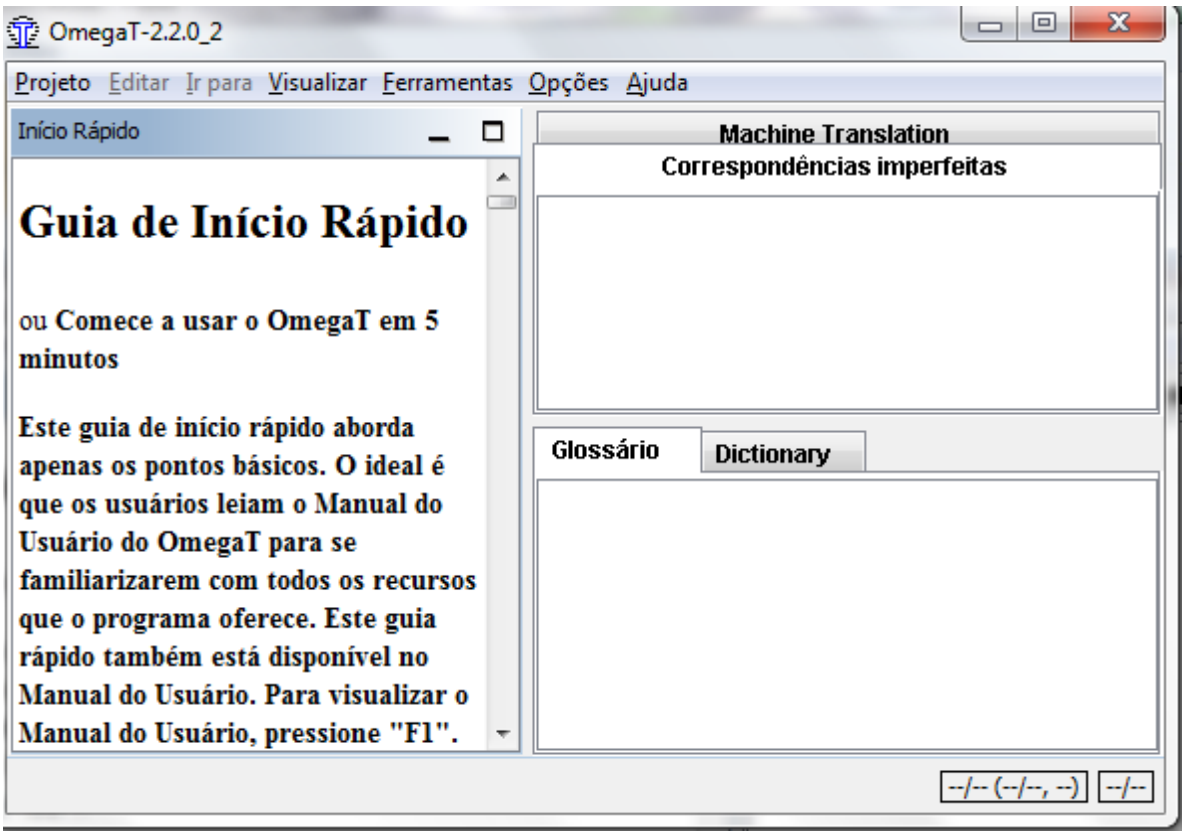

Figura 1. Tela de abertura do OmegaT versão 2.2.0_2 (reduzida para se adaptar à página).

Tags

O programa preserva a formatação do arquivo original com o uso de *tags*. Cada diferença de formatação é marcada por uma *tag* no segmento. O usuário reproduz as *tags* nos pontos apropriados da tradução para aplicar as mudanças de formação. Um exemplo ajuda a esclarecer sua função. No segmento abaixo

The <f0>book</f0> is on the table

… as *tags* "<f0>" e "</f0>"[7] marcam o começo e o final de uma mudança de formatação, que pode ser desde um texto em itálico a uma alteração simultânea de fonte, tamanho, cor e tipo. Infelizmente, se um mesmo segmento tem muitas alterações internas de formatação, o resultado pode ser desastroso. Por exemplo:

<f0>The</f0><f1> </f1><f2>book<f2></f2><f3> is</f3><f4> </f4><f5>on the table.</f5>

Os usuários chamam essa poluição visual de "floresta de *tags*", um lugar confuso e onde é muito fácil se perder. O sistema preserva toda a formatação tal e qual ela era apresentada no original, mas isso também significa que (a) dependendo do formato de arquivo utilizado, é impossível alterar a formatação do texto dentro do programa – por exemplo, para italicizar estrangeirismos; e (b)

7 O programa usa letras diferentes nas tags para designar tipos diferentes de mudanças do original – <br0/> para quebras de linha, <t0/> para quebras automáticas de página, <s0/> para múltiplos espaços e <a0></a0> para links, entre outros. Em geral, o usuário pode ignorar essas diferenças específicas; em caso de dúvida, consultar o original basta para esclarecer que elemento do documento cada uma está representando.

quando a formatação é irrelevante, a presença das *tags* pode ser um incômodo. Ambos os problemas são contornados pelos usuários sem dificuldade e, na minha experiência, raramente interferem com o trabalho. Caso o usuário apague ou altere uma das *tags* da sua tradução por acidente, os documentos criados no diretório de arquivos traduzidos ficam corrompidos; logo, é importante que todos os segmentos reproduzam fielmente na tradução todas as *tags* do texto de origem. Felizmente, o programa oferece uma maneira simples e rápida de confirmar que todos estão corretos ("Ferramentas → Validar Tags") e apresenta uma lista de todos os segmentos que precisam de correção.

Descrito dessa maneira, o sistema de *tags* pode parecer terrivelmente complexo. Na verdade, ele é uma simplificação tremenda da vida do tradutor. O uso de *tags* permite que o usuário não se preocupe com a formatação. Todas as diferenças ficam marcadas, mesmo aquelas que o leitor não perceberia a olho nu, e preservadas pelo programa. E quando o usuário fica perdido em uma "floresta de *tags*", ele sempre tem a opção de abrir o arquivo original, simplificar a formatação e recarregar o projeto.

Desvantagens

As três principais desvantagens do OmegaT 2.2.0_2 são: a impossibilidade de traduzir segmentos idênticos de modos diferentes em um mesmo projeto, a inflexibilidade das regras de segmentação e a ausência de um suporte técnico dedicado. Todas as três podem ser contornadas de alguma maneira e são pouco incômodas para o usuário inteligente, mas podem representar obstáculos significativos para alguns usuários especializados.

Dos três, o primeiro é o menos grave e o mais fácil de contornar: dentro de um mesmo projeto, quando um segmento é exatamente idêntico ao outro, os dois não podem ter duas traduções diferentes. Se o nosso exemplo *The book is on the table* aparece duas vezes no mesmo texto, ele não poderia ser traduzido na primeira vez como "O livro está sobre a mesa" e na segunda como "O livro é um fator nessa negociação". O problema não é frequente, mas acontece principalmente com segmentos que são frases curtas, itens em tabelas, índices remissivos e assemelhados. Essa característica também afeta casos em que um segmento deve ser traduzido e o outro não, por exemplo, o título em duas línguas diferentes de um artigo científico. A solução é editar o arquivo fonte e adicionar alguma marca ao segmento relevante, por exemplo, um asterisco no começo ou final, diferenciando os dois. Nesse caso, o resultado é:

The book is on the table
O livro está sobre a mesa

The book is on the table*

O livro é um fator nessa negociação

Obviamente, essa solução só é possível quando o usuário pode editar o arquivo fonte.

Quanto ao segundo problema, o OmegaT só permite a segmentação do texto em dois níveis: sentença e parágrafo. Não há nenhuma configuração intermediária. Se o usuário decide trabalhar por parágrafo, ele corre o risco de enfrentar segmentos excessivamente longos e encontra menos correspondências imperfeitas. Se, por outro lado, decide trabalhar com segmentação em nível de sentença[8], o usuário precisa enfrentar regras imperfeitas, que cortam as sentenças em pontos impróprios quando o texto esbarra em abreviaturas incomuns. O problema pode ser corrigido com a adição de novas regras pelo menu "Opções → Segmentação". Infelizmente, a versão atual do programa não contém um conjunto pronto de regras de segmentação para a língua portuguesa. Este pode ser criado[9] com a ajuda do manual de instruções à medida que o usuário encontra os casos que exigem novas regras, mas este trabalho de customização do software provavelmente está além da boa vontade de muitos usuários.

Mas o uso da segmentação em nível de sentença gera um risco mais grave e para o qual não há uma solução técnica: a perda de contexto. Com o texto dividido por sentenças, é difícil saber onde termina um parágrafo e começa o seguinte. Usuário se depara com um texto fragmentado e fragmentário. Em outras ferramentas de CAT, nas quais o usuário trabalha dentro da interface do editor de texto (ou outro aplicativo), a segmentação por sentença não causa esse efeito porque o tradutor tem acesso imediato ao documento em sua forma original; ele está sempre trabalhando com a estrutura original do texto. Com o OmegaT, isso não acontece. O usuário pode manter o original aberto em paralelo e consultá-lo periodicamente, é verdade, mas a solução é deselegante e trabalhosa. Na margem, o resultado é um texto menos coeso, ou pelo menos uma maior dificuldade em construir a coesão textual. O usuário ciente desse problema pode sempre optar por trabalhar com segmentação em nível de parágrafo, mas isto apenas retorna aos problemas apresentados anteriormente.[10]

Finalmente, a ausência de suporte técnico dedicado significa que o usuário que enfrenta problemas não tem como recorrer a uma equipe de atendimento especializada. Mas o usuário não está sozinho.

---

8 "Sentença" é o termo usado pelo OmegaT para se referir a frases, ou seja, sequências de caracteres que terminam em um ponto final ou quebra de parágrafo. Cada língua tem exceções, de modo que os segmentos não são interrompidos abruptamente quando contêm sequências como "Dr. Fulano" e "D. Pedro I". A lista de exceções não é completa, mas pode ser expandida e personalizada pelo usuário.

9 O usuário também pode instalar um conjunto de regras de segmentação pronto, já que o OmegaT adota o padrão Segmentation Rules eXchange (SRX), também administrado pela Localization Industry Standards Association (LISA). Para mais informações, ver http://www.lisa.org/Segmentation-Rules-eXchange-SRX.40.0.html

10 É possível alterar as regras de segmentação de um projeto durante o uso, mas elas sempre se aplicam a todo o projeto. Assim, se o usuário traduziu parte do texto com segmentação por sentença e depois decide usar a segmentação por parágrafo, todos os segmentos constituídos por mais de uma sentença mudam. A tradução anterior não desaparece, permanecendo armazenada na memória de tradução e aparecendo em buscas sob a marcação de "Strings órfãs", já que não se referem a nenhum segmento contido no projeto atual; em alguns casos, elas podem até aparecer como correspondências imperfeitas. Ainda assim, o usuário precisa reinseri-las manualmente. O problema oposto, não menos trabalhoso, ocorre quando o usuário decide alterar a segmentação de parágrafo para sentença.

Os desenvolvedores do programa, além de vários usuários experientes, participam da lista de discussão oficial do projeto.[11] Os participantes divulgam notícias, trocam dicas de uso, resolvem dúvidas técnicas e debatem o futuro do programa; a maior parte das mensagens está em inglês, mas o grupo inclui discussões ocasionais em russo, alemão, francês, espanhol e português. Participantes são encorajados a ler os arquivos da lista de discussão antes de fazer perguntas que já foram trabalhadas. Os usuários também podem participar de chats sobre o programa em um canal IRC,[12] mas nada garante que a sala sempre terá outro indivíduo para participar da conversa, quanto mais um capaz e disposto a resolver dúvidas na língua do usuário.

A natureza aberta e voluntária do projeto também faz com que o trabalho de localização (tradução do software) avance de modo irregular: os menus do aplicativo misturam inglês e português, com as funções mais novas podendo demorar bastante tempo para serem traduzidas. O resultado é que o menu "Ferramentas", por exemplo, apresenta as opções "Validar tags", "Statistics" e "Match Statistics", criando uma mistura que faz o aplicativo parecer permanentemente inacabado. O manual de instruções só cobre as funções introduzidas até a versão 2.0.5, mas o documento não está atualizado em todas as línguas. Essa última versão está disponível em inglês e espanhol, entre outras línguas, mas o texto em português vai apenas até a versão 1.6.1 e o francês para na versão 1.4.4, colocando as duas línguas no mesmo patamar que esperanto e albanês, respectivamente, enquanto leitores de húngaro, basco, holandês, russo, esloveno e tcheco têm acesso à última versão do documento.

Uma Não Desvantagem

Leitores mais experientes com ferramentas CAT devem ter percebido que esta análise não lista como desvantagens a ausência de um alinhador ou conversor de memórias de tradução nativos, a impossibilidade de se adicionar termos aos glossários através da interface do programa, tutoriais internos, etc. Estas características não qualificadas como desvantagens porque fogem do escopo do programa. Alguns usuários de ferramentas CAT preferem sua integração ao programa, mas a verdade é que todas essas funções estão disponíveis na forma de aplicativos, *scripts* e sites, o que mantém a simplicidade da interface e a leveza do programa e permite que o usuário customize sua experiência com recursos externos. O uso de padrões não-proprietários, como o formato TMX, significa que o usuário do OmegaT tem acesso a uma ampla variedade de recursos externos.

Por que esta análise já está obsoleta

É preciso destacar que esta análise muito provavelmente não está examinando a última versão do

11 Disponível no endereço http://groups.yahoo.com/group/omegat
12 Disponível no endereço http://java.freenode.net//index.php?channel=omegat

programa. O OmegaT muda pouco de versão para versão, mas uma nova é disponibilizada quase que todos os meses, sempre com correções de bugs ou a adição de um ou outro pequeno recurso. Por exemplo a última versão[13] incluiu correções no filtro para arquivos HTML, uma correção na tradução automática do chinês e uma atualização da localização em língua italiana; a anterior, corrigiu um *bug*, atualizou a localização em três línguas e agregou um parâmetro ao mecanismo de busca.

Todas estas versões são beta, ou seja, ainda em fase de teste, mas bastante estáveis e recomendadas pela equipe do projeto. Os desenvolvedores também disponibilizam uma "versão estável", atualizada com menos frequência – no momento da redação desta análise, esta era a versão 2.0.5. O aplicativo não baixa e instala as atualizações automaticamente, o que pode gerar desafios para os usuários menos adeptos ou em situações nos quais o programa precisa ser instalado em múltiplos sistemas. A solução é simples: visitar a página do programa com regularidade ou assinar o *feed* de RSS desta, que informa o usuário automaticamente de quaisquer atualizações do site.

Recomendo que os leitores experimentem o programa, especialmente se não forem usuários atuais de ferramentas de CAT ou se estudantes da disciplina ou novatos na profissão. O OmegaT 2.2.0_2 oferece os recursos mais importantes e permite que o usuário se familiarize com o uso de memórias de tradução, glossários e dicionários, tudo em uma interface simples e eficiente. Usuários experientes podem descobrir uma ferramenta robusta que, por ser software livre, reduz o custo de formar equipes de tradutores. Finalmente, mesmo os usuários que desgostarem do aplicativo estarão aprendendo mais sobre as ferramentas de CAT e estarão mais capacitados para avaliar concorrentes como Déjà Vu, Wordfast, o também software livre Anaphraseus e o líder SDL Trados.

13 As informações sobre o que mudou de versão para versão estão listadas no arquivo changes.txt, salvo na pasta principal do aplicativo. Em um computador usando o sistema operacional Windows 7 Home Premium, por exemplo, esta deve ser file://C:\Program Files (x86)\OmegaT.