2
Contextualização

Neste capítulo, apresentamos a tradução automática e a localização de software e comentamos algumas relações entre as duas áreas a fim de contextualizar nosso estudo de caso, que será apresentado no capítulo 4.

2.1.
Tradução automática

Breve histórico

A tradução automática é comumente entendida como a tradução gerada por computador. Alguns sistemas de tradução automática (daqui em diante também referida como *TA*) incluem etapas para participação humana, seja na manutenção ou no aprimoramento do sistema, com a inclusão de termos nos dicionários ou de regras gramaticais, seja durante o próprio processo de tradução, para seleção da melhor opção dentre aquelas apresentadas pelo sistema, por exemplo. Contudo, independentemente da etapa em que haja intervenção humana e de quão freqüente ela seja, o fator determinante na tradução automática é que a interpretação do texto de origem e a produção do texto traduzido são realizadas pelo computador.

A TA sempre foi um tema polêmico. As opiniões variam desde um ceticismo absoluto sobre a possibilidade de um computador realizar uma tarefa humana tão complexa até a convicção de que seja apenas uma questão de tempo até que os computadores sejam capazes de substituir os tradutores. Entre esses extremos, há aqueles que acreditam que a TA seja possível dentro de contextos específicos, outros que vivenciam um certo sentimento de ameaça, como relatado por alguns tradutores profissionais, ou um sentimento de insatisfação em relação à qualidade, como o público leigo e os usuários dos poucos sistemas disponíveis atualmente para português.

A história da TA costuma ser dividida em fases bem distintas na literatura (Martins&Nunes, 2005; Dayrell, 1999; Arnold *et al*, 1993; Hutchins&Sommers, 1992).

Na primeira fase, por volta da década de 50, a TA surgiu como um projeto ambicioso, em que o desconhecimento sobre a complexidade da linguagem humana deu margem a pretensões de substituição do tradutor humano e criação de um processo totalmente automatizado, capaz de funcionar para várias línguas e para qualquer assunto, com qualidade. As línguas eram entendidas como códigos e parecia fazer sentido aplicar também às línguas os métodos de criptografia dominados pelos matemáticos, para gerar traduções de qualidade com mais rapidez e menor custo.

Grandes investimentos foram feitos, especialmente pelos governos dos Estados Unidos e de alguns outros países, motivados pela possibilidade de utilizar os tradutores automáticos em assuntos de interesse militar. Contudo, na segunda fase, já em meados da década de 60, constatou-se que os resultados não foram os esperados, os problemas se multiplicavam e os investimentos pesados não apresentavam o retorno pretendido, fazendo com que o projeto praticamente tivesse sua morte decretada com o famoso relatório ALPAC (*Automatic Language Processing Advisory Committee*) em 1966. Apesar de ser voltado especificamente para o governo dos Estados Unidos e se restringir à avaliação dos resultados em relação à tradução automática de documentos de interesse militar escritos em russo, o parecer negativo do relatório teve um forte impacto nas pesquisas em tradução automática em geral, fazendo com que houvesse pouco investimento (e, conseqüentemente, pouco desenvolvimento) nessa área durante a década seguinte.

Apesar da redução do interesse direto em TA, as constatações e as recomendações do relatório tiveram impacto decisivo no desenvolvimento de outras áreas de interação entre linguagem e tecnologia: o relatório sugeria que os esforços e investimentos deveriam se concentrar em pesquisas mais básicas em lingüística computacional e a própria tradução automática, assim como outras ferramentas, deveriam ser desenvolvidas como forma de apoio à tarefa do tradutor humano, e não de substituição. As ferramentas de auxílio à tradução (ou *CAT, computer-aided tools*) desenvolveram-se muito a partir deste período. Além de dicionários eletrônicos, corretores gramaticais e ortográficos e ferramentas para

gerenciamento de terminologia, por exemplo, as *CATs* que mais ganharam expressividade no mercado nos últimos 20 anos foram os programas de memória de tradução, como o *Trados Workbench*, o *Wordfast*, o *Déjà Vu* e o *SDLX*, que têm fundamental importância no contexto deste trabalho pela sua aplicação no mercado de localização de software, como veremos mais adiante na seção sobre localização.

Numa terceira fase, por volta do início da década de 80, observou-se uma retomada de interesse sobre a tradução automática, ainda com motivação militar, durante a Guerra Fria, mas também com motivação comercial, já que nessa época os computadores pessoais, os aplicativos e a internet começavam a ser utilizados por milhões de pessoas em todo o mundo. Facilitar a comunicação no mundo globalizado e, principalmente, reduzir os custos e o tempo gastos com tradução voltou a ser prioridade.

Apesar de manter vários dos principais objetivos iniciais, como a redução dos custos e do tempo para a realização das traduções, a retomada das pesquisas em tradução automática teve características muito diferentes daquelas do projeto inicial da década de 40. Além do amplo desenvolvimento tecnológico, as pesquisas sobre inteligência artificial, linguagem e processamento de linguagem natural (PLN) já estavam mais avançadas e era possível saber, também com base nos resultados dos primeiros projetos, que fazer com que uma máquina reproduzisse o processo tradutório humano era uma tarefa ambiciosa demais. O escopo dos projetos assumiu um perfil mais realista e a pretensão de desenvolver um programa capaz de traduzir com qualidade qualquer assunto em qualquer idioma deu lugar ao objetivo de utilizar a tradução automática em domínios específicos do conhecimento, muitas vezes em pares de idiomas e direções delimitados (somente do inglês para o francês, por exemplo). O projeto canadense MÉTÉO®, talvez a iniciativa mais bem-sucedida e famosa atualmente, mostrou ser possível utilizar com sucesso a tradução automática em domínios restritos, como as previsões meteorológicas. O projeto é utilizado de inglês para francês até hoje, desde 1977.

Restringir a tradução automática a domínios específicos foi um passo que representou várias vantagens práticas: em sistemas como o que foi utilizado na presente pesquisa, por exemplo, baseados em conhecimento lingüístico com

gramáticas e dicionários, é possível criar e ampliar constantemente dicionários com a terminologia específica do domínio, analisar os padrões lingüísticos do texto original e criar regras gramaticais mais específicas (referentes ao uso do imperativo em manuais, por exemplo). A essa linguagem específica empregada em um domínio, costuma-se dar o nome de *sublíngua*. Textos que tratam do mesmo tema ou têm forma ou função semelhantes, como é o caso dos manuais, e de certa forma, dos softwares, tendem a utilizar um vocabulário comum, estruturas sintáticas similares e, portanto, a identificação, a descrição e a formalização dessas semelhanças delimitam a tarefa, que antes tinha a ambição de formalizar a linguagem como um todo, e tendem a contribuir para a melhoria do desempenho dos sistemas.

Mesmo dentro da sublíngua dos domínios, é possível restringir ainda mais a linguagem, empregando uma *linguagem controlada*, cujo principal objetivo é reduzir a ambigüidade do texto original para melhorar a qualidade da tradução automática. Isso pode ser feito durante a própria redação do material original, pelo autor, ou na adaptação (pré-edição) de um texto já existente antes de submetê-lo à TA.

Atualmente, o desenvolvimento tecnológico garante uma capacidade de processamento de dados que era inimaginável décadas atrás. Com esses recursos, as questões de formalização de conhecimento lingüístico e enciclopédico assumiram o foco das discussões e dos estudos. Dentre elas, destacam-se as questões que envolvem semântica e desambiguação de sentidos, tanto no nível lingüístico quanto no nível cognitivo mais amplo, dos conhecimentos extralingüísticos ativados por um tradutor humano durante a tarefa de tradução, por exemplo.

Uma descrição mais detalhada sobre a história da tradução automática pode ser encontrada em Arnold *et al.* (1994) e Hutchins&Sommers (1992).

Principais características dos sistemas de tradução automática

Atualmente, vários tipos de sistemas de TA coexistem. Eles partem de premissas diferentes sobre o que sejam linguagem e tradução e são utilizados para diferentes propósitos.

Os sistemas variam de acordo com a quantidade de idiomas envolvidos, podendo ser bilíngüe ou multilíngüe (Martins&Nunes, 2005).

Outra variação diz respeito à direção da tradução. Um sistema desenvolvido para traduzir especificamente de inglês para português, por exemplo, é um sistema unidirecional. Já um sistema que ofereça a possibilidade de tradução na direção do inglês para o português e do português para o inglês, por exemplo, será chamado de bidirecional (Rino&Specia, 2002).

A utilidade da TA é em grande parte determinada pela capacidade dos sistemas disponíveis. No estágio atual de desenvolvimento, a maioria dos sistemas é capaz de produzir um desses dois tipos de tradução: a "rudimentar" (*rough*) ou a "crua" (*raw*) (Martins&Nunes, 2005). A tradução rudimentar apresenta uma qualidade inferior, servindo principalmente ao propósito de compreensão mais superficial de um texto (*gisting*). É muito utilizada na internet ou por empresas para determinar se um texto deve ou não ser traduzido por profissionais, por exemplo. Já a tradução crua em geral apresenta melhor qualidade e é utilizada para a obtenção mais rápida e mais barata de uma primeira versão traduzida do texto, para a posterior revisão (pós-edição) desse "rascunho". No primeiro caso, até mesmo os sistemas mais simples, que empregam uma tradução "palavra por palavra", podem conseguir atender ao objetivo. No segundo caso, vários tipos de sistemas são usados, recorrendo a diferentes métodos e recursos para garantir um nível de qualidade que justifique sua utilização. Esse é o caso, por exemplo, dos sistemas utilizados pela Comissão Européia e por empresas como a Siemens e a Toshiba (*ibidem*). Se a qualidade não fosse razoável, não se justificaria o investimento em TA, sendo provavelmente melhor recorrer diretamente a tradutores profissionais, apesar dos custos.

Em relação à intervenção humana, os programas podem ser interativos, contando com a participação do usuário durante o processo tradutório (para a seleção de uma dentre duas opções para uma frase ambígua, por exemplo) ou não-interativos. Alguns programas oferecem a possibilidade de intervenção em seus parâmetros e configurações, para inclusão ou alteração de regras gramaticais ou de palavras nos dicionários, por exemplo. Uma terceira forma de intervenção humana se dá em relação aos textos envolvidos no processo. Em relação ao texto original, pode haver a pré-edição (para alteração de características do texto a fim

de reduzir ambigüidades e facilitar a TA) e o uso de linguagem controlada. Na pré-edição de um texto já existente, são feitas adaptações diretas ou são incluídas marcações que visam fornecer informações metalingüísticas ao sistema. Já no caso da linguagem controlada, o texto é escrito de acordo com regras rígidas de redação, que refletem as características lingüísticas com as quais o tradutor automático já tem capacidade de lidar.

Finalmente, Martins&Nunes (2005) citam a pós-edição, na qual a intervenção humana se dá pela revisão, por tradutores ou revisores especializados, do texto gerado pela máquina. Atualmente, essa é a intervenção mais utilizada.

Specia&Rino (2002) definem os sistemas em relação a dois aspectos principais: seus métodos de tradução e seus paradigmas. Os métodos são a tradução direta (palavra por palavra) ou a tradução indireta, sendo que o último pode ser baseado em transferência ou em interlíngua. No método direto pode haver algum processamento morfológico e um reordenamento básico de palavras, mas não há processamento sintático. O método de tradução indireta baseado em transferência procura estabelecer correspondências principalmente entre as estruturas sintáticas e às vezes envolve aspectos semânticos das línguas envolvidas e, por isso, é específico para um determinado par de idiomas. Já o método indireto por interlíngua procura estabelecer uma linguagem independente das línguas envolvidas. Esse método procura desmembrar o processo de tradução em duas etapas separadas: “a de projeção ou representação do texto-fonte na língua intermediária; e a da geração do texto de saída, na língua-alvo, a partir desta representação intermediária” (Martins *et al*, 2004:41). Uma discussão mais detalhada sobre os métodos, suas vantagens e desvantagens e exemplos de implementação no Brasil e no exterior pode ser encontrada em Specia&Rino (2002) e Martins *et al* (2004). Na taxonomia FEMTI (Hovy *et al*, 2002) que aplicaremos no capítulo 4, os métodos são chamados de *modelos* e os paradigmas, de *metodologias*.

Sobre os paradigmas, Specia&Rino (2002) apresentam uma descrição bastante detalhada, dividindo-os entre fundamentais e empíricos. Os paradigmas fundamentais baseiam-se em teorias lingüísticas bem definidas sobre as línguas naturais envolvidas, enquanto os paradigmas empíricos “indicam técnicas experimentais para especificar o mecanismo de tradução apropriado ao contexto

em foco" (*ibidem*, p. 15), sem recorrer diretamente a teorias lingüísticas e baseando-se no uso do língua. Como observam as autoras, esses paradigmas ganharam um grande impulso nos últimos anos com os avanços tecnológicos que possibilitam o processamento rápido de grandes quantidades de dados.

Na estrutura de Specia&Rino (2002). os paradigmas fundamentais estão divididos da seguinte forma:

- TA baseada em regras – faz a tradução com base em regras lingüísticas, em geral por mapeamento em árvores sintáticas;
- TA baseada em conhecimento – também se baseia em regras e envolve conhecimentos lingüísticos. A diferença principal está na utilização de conhecimentos extralingüísticos (também chamados de *conhecimento de mundo* ou *conhecimento enciclopédico*), apoiados em ontologias ou modelos conceituais;
- TA baseada em léxico – nesses sistemas, as regras para o mapeamento entre as línguas envolvidas são determinadas pelos itens lexicais e sua estrutura de argumentos;
- TA baseada em restrições – nesse caso, as regras são definidas por restrições lingüísticas que determinam as relações entre as estruturas da língua-fonte e da língua-alvo;
- TA baseada em princípios – nesse paradigma, em vez de regras lingüísticas específicas, o sistema utiliza conjuntos de princípios mais abstratos sobre morfologia, gramática e léxico. Um exemplo apresentado pelas autoras é o sistema *Princitran*, baseado nos princípios sintáticos da Teoria da Regência e Ligação (Chomsky 1981, *apud* Specia&Rino 2002);
- TA *shake and bake* – esse paradigma baseia-se na transferência entre itens lexicais. Uma vez que essa transferência é feita, o sistema emprega um algoritmo não-convencional para combinar os itens lexicais da língua-alvo, envolvendo apenas conhecimentos monolíngües tanto na interpretação da frase original quanto na geração da frase traduzida.

Os paradigmas empíricos descritos pelas autoras são os seguintes:

- TA baseada em estatística – emprega técnicas estatísticas de ocorrências de palavras e estruturas na língua-fonte e na língua-alvo para resolver as tarefas envolvidas na tradução, como a desambiguação lexical, por exemplo. A qualidade da tradução produzida depende diretamente da abrangência e da qualidade do *corpus* de pesquisa e da exatidão dos modelos probabilísticos;
- Ta baseada em exemplos – também chamada de TA baseada em casos. Toma como base um *corpus* paralelo (que contém textos originais alinhados a suas respectivas traduções). Um algoritmo compara o texto original com os textos do *corpus*. Ao encontrar uma sentença semelhante, ele adota a tradução que foi dada a esse texto como modelo para a nova tradução. A semelhança entre as sentenças é "determinada pela distância semântica entre as suas palavras, a qual pode ser calculada com base na distância entre essas palavras em uma hierarquia de termos e conceitos provida, em geral, por um *thesaurus* ou uma ontologia" (Specia&Rino, 2002:16). A quantidade de dados necessária para esses sistemas é tão grande que pode dificultar o armazenamento e as buscas. Uma combinação desse paradigma a algum conhecimento lingüístico pode simplificar o processamento;
- TA baseada em diálogo – esses sistemas estabelecem uma interação com o autor do texto original, identificando pontos de ambigüidade que o autor deve esclarecer durante o processo de tradução ou antes de submeter o texto à tradução. Devido à quantidade de informações necessárias, em geral eles ficam restritos a domínios específicos;
- TA baseada em redes neurais – Ainda não há sistemas totalmente baseados em redes neurais, mas essa abordagem conexionista vem sendo utilizada nas funções de *parser* (análise sintática), desambiguação lexical e aprendizado automático de regras.

Por fim, Specia&Rino (2002) comentam os benefícios da utilização de sistemas híbridos, que podem combinar as vantagens dos paradigmas fundamentais e empíricos.

Como vemos, as possibilidades em tradução automática são inúmeras e as escolhas teóricas e práticas dependerão em grande parte do contexto de utilização pretendido. No caso do presente trabalho, o contexto de uso é o mercado de localização, descrito na próxima seção.

2.2. Localização

Localização de software e outros produtos

Cada vez mais empresas querem divulgar e vender seus produtos e pesquisadores e pessoas comuns querem compartilhar suas descobertas e idéias. Para todos eles, o mercado e o público não se restringem mais à própria comunidade ou ao seu país, tendo se expandindo potencialmente para o mundo inteiro pela internet. Porém, para que produtos e idéias possam atingir um público maior, é preciso que eles sejam *localizados*. Assim, poderão ser entendidos e consumidos por um maior número de pessoas em mercados variados.

Mas o que significar *localizar*? A LISA (*Localization Industry Standards Association*) define localização como um processo que envolve a adaptação lingüística e cultural de um produto para o local (país/região e idioma) no qual ele será usado e comercializado (Rieche, 2004). Para que um software possa ser vendido em outros países, é preciso que ele seja traduzido e adaptado a cada um desses países, a cada língua e cultura. As adaptações podem incluir desde mudanças nos formatos de hora, data, endereços e nomes até alterações nas cores e símbolos, na organização interna e no conteúdo do produto, nas leis de direitos autorais e na garantia e na estrutura do software. Um exemplo são os idiomas com caracteres especiais, como as línguas asiáticas, o russo e o árabe (este último, além das diferenças nos caracteres, é escrito da direita para a esquerda). Outro exemplo citado no mercado é o da cor branca, que no Brasil e em outros países representa a paz, mas na China é associada à morte.

Originalmente, o termo aplicava-se somente a programas de computador, mas hoje em dia engloba outros produtos, como telefones celulares, microondas, aparelhos de fax e de DVD, equipamentos médicos e industriais, *websites*, material de *e-learning*, CD-ROM, jogos. O que faz com que softwares, páginas da internet, CD-ROMs, microondas e telefones celulares sejam todos considerados produtos *localizáveis* é o fato de combinarem tecnologia e linguagem e envolverem mais de um *meio* de transmissão de dados (multimídia, neste sentido), fazendo com que seu processo de adaptação para outros países tenha características específicas, como veremos mais adiante. Como comenta Thomas Wassmer no artigo "Tools for localizing multimedia applications" (2003:121), o principal desafio "em todos os tipos de mídia ou em qualquer combinação de mídias (multimídia) é ter acesso ao texto escrito e falado para facilitar a localização e depois pôr essas traduções de volta na mídia" (nossa tradução)[1].

Como mencionado anteriormente, os produtos envolvem tecnologia e linguagem específicas e a tradução e a adaptação de seus conteúdos dependem de muitas etapas. Eles são dinâmicos e as opções do usuário e os textos inseridos por ele provocam ações diferentes, fazendo com que o produto realize diferentes operações.

A seguir, será apresentado, em linhas gerais, o processo de localização.

O processo de localização

O processo de localização pode começar numa empresa especializada ou na própria empresa fabricante do produto. Alguns fabricantes têm um departamento dedicado ao gerenciamento do processo de localização e chegam a fazer parte das tarefas internamente. No entanto, por questões relacionadas a custos decorrentes da complexidade do processo de localização, essa atividade passou a ser terceirizada para empresas especializadas na localização do produto em vários idiomas.

[1] "[...] it is the common problem in all types of media and in any combination of those (multimedia) to have written and spoken language accessible for an easy localization and then get those translations back into the media."

Como a maioria dos fabricantes dos softwares utilizados no Brasil é estrangeira, o processo começa ainda no exterior. Por isso, normalmente os clientes das empresas de localização brasileiras não são diretamente os fabricantes, como a *Microsoft* e a *Symantec*. Esses fabricantes contratam grandes empresas de localização para gerenciar o processo geral e terceirizar cada idioma para empresas especializadas nos países onde o produto será comercializado.

O sucesso do processo de localização depende muito das etapas de globalização e internacionalização do produto, realizadas pelo fabricante. Essas etapas estratégicas visam o planejamento da comercialização do produto em outros países (globalização) e seu desenvolvimento voltado para a localização, para que ele já seja desenvolvido de forma a facilitar o processo (internacionalização) (Rieche, 2004).

Uma das principais tarefas do processo de localização é ter acesso ao texto falado ou escrito para traduzi-lo e depois inseri-lo de volta no produto. Para que essa tarefa seja possível, a engenharia é uma etapa fundamental. Dentre outras tarefas, cabe aos engenheiros de software extrair todo o texto traduzível, protegendo os códigos de programação, e converter os arquivos do software em formatos mais acessíveis para os tradutores, como *arquivos rtf*, editáveis nos principais processadores de textos, como o *Microsoft Word*.

Ao gerente de projeto cabem as tarefas de fazer o cronograma do projeto, considerando sempre o prazo estabelecido pelo cliente, selecionar e contactar a equipe, organizar o material necessário (arquivos para tradução, material de referência, instruções, arquivos de memória de tradução) e enviar os arquivos para a engenharia e posteriormente para os tradutores e revisores, além de fazer todo o acompanhamento e a conclusão do projeto. Quando a engenharia conclui a preparação dos arquivos, o gerente de projeto elabora o cronograma e envia o material para a equipe de tradutores. Normalmente, o serviço de tradução é terceirizado para tradutores autônomos ou que atuam como pessoa jurídica, para evitar os encargos e custos de manter um grande número de funcionários contratados pela empresa.

A etapa seguinte à tradução é a revisão. Nessa fase, profissionais em geral mais experientes no processo de localização e na área de especialidade do software (informática, finanças ou medicina, por exemplo) comparam o texto

original e o traduzido, verificando a precisão da tradução em relação ao original, o estilo e a correção gramatical e a padronização terminológica do texto traduzido. Após a revisão, os engenheiros convertem o material de volta para o formato original e atualizam a memória de tradução. Em relação ao material impresso (normalmente o guia do usuário e a embalagem), a etapa final é a editoração eletrônica, ou *DTP* (*Desktop Publishing*), para formatação final. Geralmente depois disso acontece a etapa de *proofreading*, de leitura final do material em português sem referência ao texto original, apenas para garantir a fluência e a correção gramatical do texto traduzido. A fase de compilação do software no novo idioma freqüentemente é realizada no país de origem. Após a compilação, começa a fase de teste, em que são verificados a visualização e o funcionamento de todas as opções, caixas de diálogo e mensagens. É essencial que todos os componentes possam ser corretamente visualizados pelo usuário e que os recursos estejam funcionando corretamente na versão *localizada*.

Após a conclusão de todas essas etapas, o projeto é submetido a um processo de controle de qualidade (*QA*, *Quality Assurance*), cujos critérios podem variar de empresa para empresa.

As ferramentas de tradução

No mercado de tecnologia, os avanços são rápidos. Novas versões dos produtos são lançadas freqüentemente para oferecer novas opções aos usuários e as empresas de localização estão sempre em busca de soluções que reduzam o tempo de realização do projeto e seus custos, sem comprometer a qualidade. Para atender a essas necessidades, várias ferramentas vêm sendo desenvolvidas e utilizadas no mercado de localização. Algumas dessas ferramentas visam dinamizar ou até mesmo automatizar o gerenciamento do processo. Outras se dedicam ao gerenciamento de terminologia, como o *Trados MultiTerm* e o *SDLTermBase*. As ferramentas mais amplamente utilizadas são as memórias de tradução.

Para compreender o processo de localização e a utilidade das memórias de tradução é preciso saber que o volume de texto de um software costuma ser muito alto, muitas vezes ultrapassando 500 mil ou até um milhão de palavras. Para se ter

um termo de comparação, esta dissertação de mestrado contém menos de 35.000 palavras. A contagem de volume de texto é feita por número de palavras do texto original porque a estrutura de um software é diferente da de um texto corrido, como um livro ou um contrato, não sendo possível fazer a contagem por páginas ou laudas, por exemplo.

O volume de palavras também determina o número de pessoas envolvidas no projeto. A equipe de tradutores e revisores precisa ser composta por várias pessoas, algumas vezes chegando a mais de 30 profissionais, caso contrário, o prazo para lançamento do produto levaria muitos meses, o que é inviável nas áreas de tecnologia.

Além do volume de texto, outro dado fundamental é a quantidade de repetições, seja dentro de uma mesma versão ou entre versões. A cada atualização de um produto, em geral apenas algumas modificações são feitas e poucas funções novas são acrescentadas. A maior parte do conteúdo se mantém inalterada. Não seria razoável fazer novamente a tradução do material completo, não somente devido aos recursos e ao tempo consumidos, mas também porque haveria diferenças entre as versões, comprometendo a familiaridade que os usuários das versões anteriores já têm com o software.

É nesse contexto que as ferramentas de tradução surgem como uma solução satisfatória em termos de economia de tempo e dinheiro e de garantia de homogeneidade de terminologia e estilo numa equipe tão numerosa. Esses programas segmentam os textos originais em trechos mais ou menos equivalentes a uma oração e duplicam esse mesmo trecho no arquivo processado, fazendo com que a mesma frase apareça duas vezes no arquivo. O tradutor, então, insere sua tradução na segunda ocorrência, de modo que o arquivo passa a apresentar um trecho no idioma original seguido de sua tradução. Assim, cria-se um banco de dados para armazenamento de todas as traduções já feitas, que podem ser reaproveitadas. Alguns exemplos dessas ferramentas são o *Trados Workbench*, o *SDLX*, o *Wordfast*, o *Déjà Vu* e o *StarTransit*. Com elas, cada trecho traduzido vai sendo armazenado, o que é especialmente útil em textos repetitivos como softwares, manuais e contratos. Na próxima ocorrência daquele trecho ou de um trecho semelhante, o programa apresenta a tradução armazenada e o tradutor pode, então, optar se deseja utilizar a mesma tradução ou fazer alterações. Quando os

tradutores estão trabalhando no mesmo local (isto é, *in-house*, o que significa “no escritório da empresa” no jargão de localização) e é utilizada uma versão completa da ferramenta (para vários tradutores), todos os profissionais têm acesso às traduções já feitas pelos colegas, o que agiliza o trabalho e ajuda a garantir a homogeneidade do material. Sem esse recurso, cada tradutor provavelmente traduziria a mesma frase de forma diferente na primeira vez em que se deparasse com ela. Alguns programas de uso exclusivo, não comercializados, já oferecem essa possibilidade remotamente, pela internet, o que elimina a necessidade de os tradutores estarem trabalhando no mesmo local, como é o caso do *Logoport*, da empresa *Lionbridge*. Local ou remotamente, quando a tradução é concluída, todos os trechos estão armazenados num arquivo que é um banco de dados com os trechos originais seguidos de suas respectivas traduções.

As ferramentas apresentam uma opção de “limpeza” dos arquivos, para que o texto original seja apagado, restando somente o texto traduzido para ser submetido ao controle de qualidade e entregue ao cliente. Caso a memória não tenha sido compartilhada simultaneamente pelos vários tradutores, nessa etapa também será feita sua atualização, com todos os segmentos de texto traduzidos. Quando uma nova versão do produto é enviada para tradução, os arquivos são comparados com a memória de tradução. Todos os segmentos que existiam na versão anterior e se repetem na nova versão são identificados pela ferramenta, que indica o percentual de aproveitamento do material que já está armazenado. A ferramenta indica os segmentos exatamente iguais (que apresentam *100% de match*, ou correspondência), bem como segmentos em que há poucas diferenças (*fuzzy match*). Assim, as traduções já existentes podem ser aproveitadas, os tradutores fazem as alterações necessárias nos segmentos que apresentaram alguma semelhança e apenas os segmentos para os quais não havia nenhuma correspondência na memória (*no match*) precisam efetivamente ser traduzidos.

Alguns softwares têm um volume grande de palavras e várias versões *beta* (de teste) antes do lançamento, o que faz com que as memórias sejam úteis também porque possibilitam que o processo de localização seja iniciado antes de o software ter sido finalizado pelo fabricante. Assim, ganha-se mais tempo para o processo de localização e diminui-se o intervalo entre o lançamento de um produto no país de origem e nos países para os quais ele está sendo *localizado*.

Quando a versão final do software é enviada para localização, basta utilizar a memória, traduzir apenas o conteúdo acrescentado e incorporar as alterações realizadas no original.

É importante observar que as memórias não realizam traduções automáticas. Elas armazenam trechos de texto e suas respectivas traduções feitas por tradutores humanos. A determinação da semelhança entre o texto original já armazenado e o novo texto a ser traduzido é feita estatisticamente, sem nenhum critério semântico. Uma descrição detalhada do funcionamento das memórias de tradução e da importância de revisão e controle de qualidade do seu conteúdo pode ser encontrada em Rieche (2004).

Componentes do processo de localização

Em termos de conteúdo e formato de arquivos, o processo de localização se divide em três componentes principais: as opções de interface com o usuário, a ajuda e a documentação. As opções de interface, também conhecidas como *UI* (de *GUI*, *graphic user interface*) são compostas pelos textos extraídos dos menus, das janelas, caixas de diálogo e mensagens, exibidos para o usuário durante a utilização do programa. O sistema de ajuda contém o texto exibido nos menus de ajuda do software, e a documentação geralmente é composta pelo material impresso que acompanha o software (em geral, um manual), o material de referência disponível em páginas da internet, além do texto de embalagens e encartes. A tradução de cada um desses tipos de arquivo apresenta especificidades e pode utilizar ferramentas de tradução diferentes.

Para a tradução dos arquivos de *UI*, algumas empresas, como a *Microsoft*, possuem uma ferramenta própria, exclusiva, não disponível para venda. Das ferramentas disponíveis no mercado, duas das mais utilizadas são o *Alchemy Catalyst* e o *Multilizer*. Além de possibilitarem o processamento de arquivos de diferentes formatos, com essas ferramentas os tradutores visualizam as caixas de diálogo, as mensagens e as janelas do software como elas são na versão original, podendo, em seguida, visualizar a interface com o texto traduzido. Essa visualização é especialmente relevante por dois motivos: compreender o contexto e respeitar as restrições de espaço. Em relação ao contexto, é complexo e às vezes

impossível traduzir palavras como *display* sem saber se estamos falando de um *visor* ou usando a forma imperativa do verbo *exibir*, ou *key*, sem compreender se é uma *chave*, uma *senha* ou a forma adjetiva *fundamental* (*key factor*, por exemplo).

Os arquivos da ajuda costumam ser criados em formato html. A empresa *Trados*, por exemplo, oferece a ferramenta *Tag Editor* para edição desses arquivos (que também pode ser usada para outros formatos), possibilitando a utilização da memória de tradução, a proteção dos códigos de formatação (as *tags*) e a visualização prévia da página traduzida no formato final. É comum também que os arquivos .html sejam convertidos para um formato editável no *Word*, como *rtf*, para facilitar o processo. Quando a ajuda é criada em *Word*, aplica-se o mesmo processo utilizado para os arquivos de documentação.

A documentação costuma ser criada em programas próprios para publicação, como o *Adobe Framemaker*. Para facilitar a tradução e a utilização da memória de tradução, os arquivos de documentação geralmente são convertidos para o formato rtf, editável no *Word*.

Esses três componentes variam também em relação às características lingüísticas. A interface com o usuário apresenta uma linguagem mais concisa e direta, para facilitar a utilização do software. Os menus apresentam comandos, como “Salvar” e “Imprimir”, e entidades, como “Arquivo” e “Ferramentas”, no *Microsoft Word*. As mensagens em geral contêm frases simples, com instruções ou perguntas sobre determinada ação, como “Deseja salvar as alterações em localização.doc?”. A linguagem da ajuda tenderá a ser um pouco mais concisa do que a da documentação pela limitação das janelas do software, mas normalmente tanto a ajuda quanto a documentação têm o papel de instruir o usuário a instalar e utilizar o software (entre outras tarefas, como solucionar problemas e fazer atualizações). Ambas fazem referência às opções de interface e é fundamental que tanto o original quanto a tradução empreguem a mesma terminologia usada nelas. Para explicar como salvar um arquivo do *Microsoft Word*, por exemplo, a ajuda e o manual deverão mencionar a opção de interface “Salvar”, não podendo referir-se a ela como “Gravar” ou “Armazenar”, o que poderia comprometer a compreensão do usuário ou confundi-lo. A possibilidade de confusão seria ainda maior se a tradução “Armazenar” tiver sido usada para outra função, como “Store”, por exemplo.

Vale lembrar que freqüentemente são lançadas novas versões dos softwares e que o processo de localização envolve grandes equipes de tradutores e revisores. Garantir a homogeneidade de estilo e terminologia entre versões e entre tantos profissionais exige um grande esforço de padronização. Nesse sentido, o material de referência desempenha um papel fundamental de apoio à tarefa tradutória. Em geral, ele é composto por glossários, guias de estilo, *sites* de internet e outros documentos e produtos do fabricante. Os *sites* e a documentação servem principalmente de referência para esclarecer questões técnicas ou o contexto de uso do produto. Já os glossários e o guia de estilo contêm as diretrizes lingüísticas específicas que devem ser consultadas e seguidas rigorosamente. Nos glossários estão definidas a terminologia a ser empregada no projeto, a terminologia já empregada nas versões anteriores do produto e as traduções específicas das opções de interface (quando ela já foi traduzida antes da ajuda e da documentação). Outra referência importante é a terminologia empregada no sistema operacional em que o software será utilizado (*Windows XP* ou *Windows 2000*, por exemplo). Já os guias de estilo definem as padronizações referentes a títulos, uso de maiúsculas, siglas, como fazer referência às opções de interface e nomes de produtos e serviços, entre outros.

Controle de qualidade no mercado de localização

Como o processo de localização foi estruturado para a indústria de software, os padrões de qualidade adotados internacionalmente para o processo e para os produtos localizados em geral são os mesmos adotados para a fabricação dos softwares originais (Rieche, 2004:87). O modelo mais empregado no mercado de localização é o da LISA, que apresenta “diretrizes para o controle de qualidade de todos os componentes de um produto localizado, incluindo aspectos lingüísticos, de formatação e de funcionalidade” (*ibidem*, p. 88). Muitas empresas fazem suas próprias adaptações do modelo, mas as categorias principais são preservadas. As categorias apresentadas a seguir são baseadas no modelo utilizado pela empresa de localização proprietária do sistema de TA avaliado nesta pesquisa especificamente para o controle da qualidade lingüística.

As principais categorias são **precisão** (tradução incorreta, erros semânticos, omissão, erros de referência cruzada e texto não-traduzido), **terminologia** (não-observância a glossários e planilhas de dúvidas respondidas, não-incorporação dos comentários do cliente referentes à amostra revisada do trabalho, falta de padronização e inadequação ao contexto), **correção gramatical** (erros de ortografia, de pontuação e gramática), **estilo** (inadequação do estilo geral, não-observância ao guia de estilo do cliente final e da agência de localização, acréscimos desnecessários, registro ou tom inadequados, uso impróprio de variantes lingüísticas e uso de gírias), **padrões locais** (formato de hora e data, símbolo de moeda, *layout* do teclado, inadequações culturais e padrões da empresa, como *copyright*, suporte técnico e garantia). Existe ainda a categoria **erro funcional** que, apesar de não ser especificamente lingüístico, pode ser provocado pelo tradutor ou pelo revisor ao lidar com o texto e inclui erros de formatação original (como não manter as fontes, os negritos e itálicos) e de formatação de *tags* (códigos inseridos no texto) ou *hiperlinks*, modificação ou exclusão de texto oculto e erros referentes a procedimentos técnicos, como instruções sobre passagens a serem mantidas em inglês, por exemplo.

Todos os erros incluídos nessas categorias recebem ainda uma classificação como erro grave (*major*) ou menos relevante (*minor*). Os erros serão considerados graves se comprometerem o lançamento do produto, como uso de declarações que possam ser ofensivas, erros que afetem a integridade dos dados ou a segurança dos usuários, erros em passagens recorrentes ou de grande visibilidade (como menus do software ou a capa do manual) e erros que demonstrem claramente que as expectativas e exigências da empresa de localização e do cliente final não foram atendidas.

Em geral, o controle é feito por amostragem durante o projeto de localização. Como comenta Rieche, os modelos são uma “tentativa de estabelecer critérios objetivos para avaliação da tradução” (2004:90) e são utilizados para gerar métricas quantificáveis para aferir a qualidade do processo e dos profissionais envolvidos, que têm implicação direta na qualidade do produto final. Na prática, observamos que muitas dessas categorias são subjetivas e os profissionais aprendem os critérios específicos com a experiência de trabalho e com as avaliações de desempenho que recebem de cada cliente.

O mercado de localização cresceu e se estruturou voltado para a localização de software. Contudo, como vimos, ele se expandiu e hoje abrange a localização de muitos produtos de tecnologia, como *websites*, equipamentos médicos e câmeras fotográficas. Apesar de o processo de localização estar bem-estruturado e consolidado, cada um desses vários tipos de produtos apresenta suas especificidades tanto lingüísticas quanto tecnológicas. No presente trabalho, iremos nos concentrar na localização de aparelhos móveis, mais especificamente, telefones celulares.

A localização de aparelhos móveis

A localização de aparelhos móveis ainda é uma área relativamente recente de pesquisa, com poucas publicações. Em 2001, A LISA (*The Localization Industry Standards Association*) publicou um artigo de apresentação da área, que ainda é a única referência sobre o assunto disponibilizada pela instituição. O artigo, escrito pelo engenheiro de localização Shailendra Musale, especialista em aparelhos móveis da empresa F-Secure Corporation, na Finlândia, trata especificamente de computadores móveis, como *palmtops* e *handhelds*. O foco da nossa pesquisa é a localização de telefones celulares, mas até a data deste estudo, não foram encontradas referências específicas sobre o tema. Contudo, muitas das características dos computadores móveis se aplicam aos telefones celulares e é cada vez maior a convergência entre esses aparelhos. Muitos telefones celulares já oferecem acesso à internet e possuem câmeras digitais. Os *smartphones*, por exemplo, são telefones celulares inteligentes, que combinam muitas das vantagens dos telefones celulares e dos computadores de mão.

Musale (2001) ressalta que as principais diferenças entre a localização de software e computadores pessoais convencionais e de aparelhos móveis são o ciclo de vida mais curto dos aparelhos móveis e a maior influência dos clientes para personalização dos produtos, o que exige prazos ainda mais curtos e mais flexibilidade no processo de localização. Outra diferença básica diz respeito à condução do processo de localização. Diferentemente dos grandes fabricantes de software convencional, que em geral adotam os padrões já estabelecidos no mercado e terceirizam a localização para empresas especializadas, muitos

fabricantes de software para aparelhos móveis ainda utilizam recursos internos, sem experiência em localização.

Um dos principais desafios na localização de aparelhos móveis é a restrição de espaço, devido ao tamanho reduzido das telas. No caso dos computadores pessoais, muitos softwares permitem o redimensionamento das caixas de texto para que, onde em inglês constava a palavra “share”, possa caber a tradução “compartilhamento”, por exemplo. No entanto, essa possibilidade não existe para os aparelhos móveis. Nesses casos, o tradutor é informado sobre o número máximo de caracteres aceitos em cada tela e cabe a ele alterar a tradução e utilizar paráfrases, formas mais concisas ou abreviações, situação equivalente, de certa forma, ao que é vivenciado na tradução de legendas. Nesse mercado, a importância da internacionalização fica ainda mais evidente. Como vimos anteriormente, na indústria de software existe uma preocupação em já desenvolver o produto procurando facilitar sua localização para outros mercados. Em alguns aparelhos, no entanto, observamos que, devido às restrições, muitas vezes um mesmo texto original é aproveitado para diferentes situações, como a palavra “All” por exemplo. Em inglês, isso não provoca problemas de concordância de gênero e número, mas em português teremos de utilizar a tradução “Todas” se o contexto se referir a “chamadas” e “Tudo”, se o usuário desejar apagar todo o conteúdo de uma mensagem, por exemplo. Assim, a utilização dessa mesma *string* de software (o mesmo segmento de texto) representa um problema, muitas vezes insolúvel, para a localização. Outra questão relacionada às *strings* é que, em geral, elas são enviadas para tradução isoladamente, sem contexto. O tradutor é obrigado a empregar uma tradução mais genérica, como “Todos” ou “Tudo”, que pode se revelar inadequada em certos contextos de uso. Também por esse motivo, assim como com os softwares convencionais, é fundamental submeter o aparelho localizado, ou pelo menos reproduções das imagens das telas do aparelho (*screenshots*), a um teste lingüístico, em que serão verificadas a correção e a adequação das traduções aos devidos contextos.

Outra questão relevante é que, enquanto a maioria dos computadores pessoais usados no Brasil utiliza o sistema operacional *Microsoft Windows*, os aparelhos móveis utilizam sistemas diferentes, como o *Microsoft Pocket PC*, o *Palm OS* e o *Symbian OS*, o que acarreta diferenças de terminologia e

funcionamento com as quais os *localizers* (tradutores envolvidos no processo de localização) devem estar familiarizados.

Nessa seção apresentamos um breve panorama do mercado de localização. Uma descrição abrangente pode ser encontrada no livro *A Practical Guide to Localization*, de Bert Esselink (2000), que é a principal referência publicada na área.

2.3. Tradução automática no mercado de localização

No mercado de tecnologia, o ciclo de vida dos produtos é muito curto. Novas versões de softwares e produtos são lançadas com freqüência e a demanda por redução de prazos de lançamento e de custos é muito intensa. Essa pressão se reflete diretamente no mercado de localização. O prazo e o custo da localização não podem comprometer questões estratégicas, como o lançamento simultâneo do produto em vários países e antes da concorrência. Nesse contexto, o uso de TA pode ser um atrativo para a redução de prazos e custos, mas a desinformação sobre o tema, as questões de qualidade, os custos diretos e indiretos e os critérios para a escolha do melhor programa para cada contexto ainda aparecem como obstáculos para a utilização mais ampla dessa ferramenta no mercado de localização.

Esselink (2000), no livro que é a principal referência publicada sobre localização, dedica apenas uma página ao tema, afirmando que a TA só trará retorno se grandes investimentos forem feitos no uso de linguagem controlada no texto original, na pós-edição e na alimentação do sistema com a terminologia do domínio dos textos. Contudo, a visão de TA nesse mercado vendo mudando nos últimos anos. É possível perceber um aumento de interesse pelo tema, refletido no número de edições especiais e artigos sobre TA em publicações da área. Muitos autores apresentam estudos de caso (Allen, 2004; Dillinger&Lommel 2004; Guerra, 2004;) de implementações bem-sucedidas de TA não somente no mercado de localização, seja do ponto de vista das empresas que são o cliente final, que precisam traduzir grandes volumes para vários idiomas, ou do tradutor profissional que pretende aumentar sua produtividade. Os estudos de Guerra

(2004) e Allen (2004) relatam o processo de implementação e o ganho de produtividade inclusive com material de marketing, domínio cuja linguagem normalmente é percebida como mais criativa do que em materiais de suporte técnico, por exemplo.

A LISA (*Localization Industry Standards Association*) publicou em 2004 um guia em que faz uma introdução à TA, explica os vários modelos, compara custos, discute vantagens e desvantagens e apresenta estudos de caso, sempre com foco no mercado de localização. A publicação apresenta o uso da TA não apenas como uma possibilidade vantajosa, mas como uma realidade imprescindível financeiramente, especialmente para projetos que envolvam grandes volumes de texto.

Como discutido na seção 2.2, as ferramentas de memória de tradução são uma peça-chave no processo de localização. O aproveitamento das memórias e a busca de formas de integração entre elas e os sistemas de TA têm sido tema de pesquisas recentes (Hodasz *et al*, 2004; Bruckner&Plitt, 2001) e será o tema da conferência anual da EAMT (*European Association for Machine Translation*) em 2006.

Neste capítulo, procuramos apresentar informações gerais sobre o processo de localização e a tradução automática, que são os temas de nossa pesquisa. Antes de analisarmos especificamente a utilização de TA no mercado de localização no Brasil, discutiremos algumas pesquisas já produzidas sobre avaliação de TA, no próximo capítulo.