Universidade Federal de Santa Catarina
Bacharelado em Letras-Libras na Modalidade a Distância

Lincoln Paulo Fernandes
Lautenai Antonio Bartholamei Junior

Estudos da Tradução II

Florianópolis

2009

APRESENTAÇÃO

> A tecnologia está se desenvolvendo a um passo assustador e as demandas feitas ao tradutor não demonstram nenhum sinal de enfraquecimento. De fato, o tradutor está se tornando cada vez mais dependente da tecnologia da informação e se ele não se adaptar às mudanças, poderá tornar-se obsoleto (Samuelsson-Brown, 1996, p. 280, nossa tradução).

Como Geoffrey Samuelsson-Brown observou na epígrafe desta apresentação, a tecnologia é agora uma realidade que não podemos escapar, assim como uma absoluta necessidade na formação do tradutor. De fato, para se manter competitivo e poder lidar com as pressões do mercado global, é necessário que o tradutor tenha familiaridade com Programas de Auxílio à Tradução (PATs). Esses programas se tornaram um pré-requisito básico para que os profissionais da tradução sejam capazes de enfrentar os desafios e sobreviver à concorrência do século XXI.

O objetivo desta disciplina é, portanto, fornecer uma introdução básica dos principais tipos de tecnologia e ferramentas que tradutores provavelmente encontrarão e acharão úteis ao longo de seus trabalhos. A disciplina ESTUDOS DA TRADUÇÃO II – Tradução e Tecnologia dá continuidade à disciplina ESTUDOS DA TRADUÇÃO I, no sentido de oferecer uma visão geral de como o trabalho do tradutor está inserido em um ambiente tecnológico. Desta forma, a disciplina está dividida em quatro unidades, a saber:

Unidade 1: Sistemas de Memória de Tradução;
Unidade 2: Sistemas Tradução Automática;
Unidade 3: Sistemas de Gerenciamento Terminológico;
Unidade 4: Corpora Eletrônicos e Tradução.

Unidade 1 – **Sistemas de Memória de Tradução** – discute os sistemas de memória de tradução (MT), como já apresentados na disciplina de Introdução aos Estudos da Tradução. A unidade está dividida em cinco partes principais: (i) histórico – onde apresentamos uma breve contextualização história dos sistemas de MT (ii) definição – onde oferecemos uma definição de trabalho do que vem a ser um sistema de MT; (ii) funcionamento – onde descrevemos o processo básico de funcionamento de um sistema de MT; (iii) tipos – onde rapidamente discutimos alguns dos principais sistemas de MT disponíveis no mercado; e (iv) aplicabilidade – onde mostramos a aplicação prática de um sistema de MT com referência a um projeto de tradução voltado aos Estudos Surdos.

Unidade 2 – **Sistemas de Tradução Automática** – esta unidade explora os Sistemas de Tradução Automática (Sistemas de TA), como já visto na disciplina de Introdução aos Estudos da Tradução. A unidade está dividida em quatro partes principais: (i) um breve apanhado histórico do surgimento dos sistemas de TA; (ii) uma definição do que vem a ser um sistema de TA; (iii) o seu funcionamento através dos vários tipos de abordagens empregadas e, consequentemente, os problemas enfrentados por estes sistemas; e, finalmente, (iv) algumas discussões que permeiam a utilização destes recursos no que diz respeito às línguas de sinais.

Unidade 3 – **Sistemas de Gerenciamento de Terminológico** – esta unidade tem por objetivo discutir sobre os Sistemas de Gerenciamento Terminológico (Sistemas de GT), e está organizada da seguinte forma: (i) um breve apanhado histórico dos SGT, a partir do surgimento do estudo da terminologia; (ii) uma definição do que são os SGT; (iii) o processo de funcionamento dos SGT, através dos vários métodos utilizados; e, por fim, (iv) a utilização de SGT no âmbito da organização terminológica em línguas de sinais.

Unidade 4 - **Corpora Eletrônicos e Tradução –** explora as diferentes características de ferramentas para análise de corpora. A unidade está dividida em cinco partes principais: (i) histórico – onde apresentamos uma breve contextualização histórica dos estudos da tradução em corpora (ii) definição – onde oferecemos uma definição de trabalho do que vem a ser um corpus; (ii) desenho – onde descrevemos o processo básico de criação de um corpus eletrônico; (iii) tipos – onde rapidamente sugerimos uma tipologia baseada em Baker (1995); e (iv) aplicabilidade – onde mostramos a aplicação prática de alguns corpora online à tradução.

Essa última unidade fecha o ciclo da disciplina, que procura mostrar a importância do conhecimento tecnológico para o desenvolvimento da competência tradutória do profissional-em-formação.

UNIDADE I

Sistemas de Memória de Tradução

Esta unidade discute os sistemas de memória de tradução (MT) e está dividida em cinco partes principais: (i) histórico – onde apresentamos uma breve contextualização histórica dos sistemas de MT (ii) definição, vantagens e desvantagens – onde oferecemos uma definição de trabalho do que vem a ser um sistema de MT e suas principais vantagens e desvantagens; (ii) funcionamento – onde descrevemos o processo básico de funcionamento de um sistema de MT; (iii) tipos – onde rapidamente discutimos alguns dos principais sistemas de MT disponíveis no mercado; e (iv) aplicabilidade – onde mostramos a aplicação prática de um sistema de MT com referência a um projeto de tradução voltado aos Estudos Surdos.

Histórico

No passado, a automação do processo de tradução era geralmente associada ao uso de tradutores automáticos (ver Unidade II). Nos dias de hoje, entretanto, a situação tem mudado consideravelmente. Os programas de apoio à tradução (PATs), principalmente os sistemas de memória de tradução (MT), vem desempenhando um papel central na atividade tradutória profissional. De fato, este tipo de ferramenta tornou-se um pré-requisito básico tanto para tradutores que trabalham em grandes empresas de tradução quanto para tradutores free-lance que colaboram com agências de tradução.

Historicamente, o conceito de memória de tradução não é algo recente. MELBY e WARNER (1995) observaram que a ideia originou-se nos anos 70 e as primeiras implementações surgiram a partir dos anos 80. Mas somente nos anos 90 é que este tipo de ferramenta eletrônica tornou-se um produto comercial amplamente disseminado entre os tradutores (p. 187). O motivo desse interesse deve-se às inúmeras vantagens que este tipo de PAT pode oferecer ao usuário (ver seção abaixo), mas antes de discuti-las, vamos, primeiramente, tentar entender o que vem a ser uma memória de tradução.

Definição, Vantagens e Desvantagens

Para Austermühl (2001), memórias de tradução são bancos de dados linguísticos que armazenam textos traduzidos juntamente com seus textos originais correspondentes. O sistema de memória de tradução, então, permite que o tradutor recupere *unidades* ou *segmentos* armazenados no banco de dados para a reutilização dos mesmos em uma nova tradução (p. 135). É essa reutilização ou "reciclagem", a ideia principal por trás de um sistema de MT, pois proporciona uma redução do tempo e custos, assim como o aumento da qualidade e consistência do texto traduzido (Bowker, 2002, pp. 92-93).

No que se refere às vantagens dos sistemas de MT, HEYN (1998) mostra que essas vantagens podem ser classificadas sob a égide de seis fatores principais e que, por sua vez, podem ser utilizados para distinguir entre as diferentes necessidades dos usuários, a saber: (i) repetição;

(ii) consistência; (iii) referência; (iv) concordância; (v) terminologia; e (vi) criação de recursos (pp. 124-125).

Repetição: as MTs encontram sua principal aplicação na tradução de material textual repetitivo, pois essa repetição permite que o tradutor re-utilize partes de uma tradução já traduzida anteriormente. Por sua vez, isso fará com que o tradutor traduza textos mais rapidamente, aumentando, assim, sua produtividade e, consequentemente, seus lucros. Entretanto, HEYN (IBID.) mostra a importância de distinguir entre repetições internas de um texto propriamente dito e repetições externas, onde as repetições são inerentes a um tipo específico de texto ou gênero.

Consistência: os sistemas de MT oferecem uma maior consistência na tradução quando integrados a um banco de dados terminológicos, pois permitem ao tradutor obter uniformidade na terminologia a ser empregada, diminuindo, assim, a necessidade da revisão terminológica do texto traduzido.

Referência: toda unidade de tradução pode vir acompanhada por vários tipos de informação, por exemplo: nome do usuário, período de criação, data de atualização, código do assunto, observações, etc. Essas informações levam a determinado melhoramento da qualidade da tradução, pois fraseados revisados e aprovados são re-utilizados. É como utilizar a tradução de uma referência autorizada, e, assim, atingir certo nível de padronização das atividades tradutórias.

Concordância: sob uma ótica linguística, uma MT pode ser descrita como um corpus paralelo bilíngue (ver UNIDADE IV). No caso dos sistemas que permitem mais de uma língua fonte e de uma língua alvo, podemos falar de corpora paralelos multilíngues. Esses corpora podem ser utilizados para recuperar uma unidade de tradução, buscando uma ou várias palavras-chave. MTs podem ser vistos como uma fonte valiosa de terminologia implícita (em contraste à terminologia explícita armazenada em bancos terminológicos). Neste sentido, MTs competem de certa forma com os bancos terminológicos.

Terminologia: o reconhecimento terminológico (RT), que é a busca automática em um banco de termos por um equivalente em uma unidade de tradução fonte, desempenha um papel fundamental nos PATs. O RT não dever ser confundido com extração terminológica, que significa a extração automática de terminologia a partir do material textual. Reconhecimento terminológico dispensa a necessidade de se fazer buscas manuais em bancos de dados, já que o sistema de MT automaticamente chama a atenção do usuário aos termos relevantes.

Criação de Recursos: os PATs podem automaticamente criar recursos de três formas: (i) gerando uma memória de tradução a partir de textos paralelos existentes em um processo conhecido como alinhamento de sentenças; (ii) gerando uma lista de candidatos a termos em um língua a serem introduzidos em um sistema de banco terminológico (extração terminológica monolíngue); e (iii) gerando uma lista de candidatos a pares-terminológicos de textos fonte e alvo a serem introduzidos em um sistema de banco terminológico (alinhamento por palavra ou extração terminológica bilíngue).

Outra vantagem que não é contemplada pela classificação de HEYN (1998) estaria relacionada à integração dos sistemas de memória de tradução com outras ferramentas.

Segundo BOWKER (2002), a maior parte desses sistemas é integrada a ferramentas de suporte à tradução. Por exemplo, os sistemas de MT mais populares no mercado (ver seção abaixo) incorporam sistemas de gerenciamento terminológico, concordanciadores bilíngues e sistemas de tradução automática. Além disso, esses sistemas trabalham de forma simbiótica com processadores de texto (e.g. *MS Word* ou *Word-Perfect*), o que de certa forma reduz a curva de aprendizado do tradutor, já que o mesmo poderá continuar a trabalhar com um aplicativo já conhecido. Em outras palavras, esta integração dos sistemas de MT com processadores de texto, sistemas de gerenciamento terminológico, concordanciadores bilíngues e sistemas de tradução automática cria uma espécie "bancada" ou "estação de trabalho" do tradutor.

Já com relação às desvantagens em se utilizar um sistema de MT, BOWKER (2002) aponta algumas das quais quatro merecem ser discutidas: (i) dificuldades relacionadas à língua e conjunto de caracteres; (ii) atitudes; (iii) remuneração; e (iv) propriedade.

(i) Dificuldades relacionadas à língua e conjunto de caracteres – segundo Bowker (IBID.), algumas línguas são mais fáceis de serem processadas do que outras, por isso, é importante certificar-se que o sistema de memória de tradução selecionado será capaz de processar o par linguístico sendo utilizado. No caso da língua de sinais, como língua espaço-visual, necessita de dois bytes para armazenar cada caractere (assim como o japonês, chinês e coreano) ao passo que na maioria das línguas um caractere pode ser armazenado utilizando um byte (i.e. uma unidade de armazenamento). Felizmente, hoje em dia, a maioria dos sistemas de MT disponíveis no mercado utiliza o padrão de codificação de caracteres Unicode que permite codificar línguas cujos caracteres necessitam de dois bytes para serem armazenados. Outra dificuldade adicional está relacionada à segmentação da língua. Ao se criar uma memória de tradução o sistema deve ser capaz de dividir a língua fonte em segmentos. Isso significa que o sistema deve reconhecer quais elementos indicam o fim de um segmento (e.g. pontuação). Neste caso, alguns sistemas de MT têm dificuldade em identificar onde um segmento termina e o outro começa. A maioria dos desenvolvedores de MT reconhece tais problemas e está trabalhando para resolvê-los.

(ii) Atitudes – no passado, ferramentas computacionais eram frequentemente vistas como uma ameaça aos tradutores. Mas nos dias de hoje a conscientização por parte de tradutores e clientes sobre os benefícios potenciais de se utilizar essas ferramentas computacionais está aumentando constantemente. Entretanto, ainda existe uma necessidade real de educar esses dois grupos sobre as potencialidades dos PATs. A confiança dos tradutores precisa ser renovada quanto ao fato de que tais sistemas de MT podem ajudá-los em suas tarefas eliminando todo o trabalho maçante e repetitivo. E no caso dos clientes, eles têm que conhecer as limitações de tais ferramentas. Embora, essas ferramentas permitam uma maior agilidade e rapidez na entrega das traduções, os clientes precisam ser lembrados que a tradução não é realizada pelo computador. Os tradutores ainda realizam uma tarefa valiosa e desafiadora e, desta forma, merecem ser tratados com respeito e remunerados adequadamente pelo trabalho que realizam.

(iii) Remuneração – a ideia de reutilizar traduções levantou questões relacionadas aos valores a serem pagos aos tradutores que utilizam MTs. Alguns sistemas de MT vem equipados com um módulo de análise de repetições (às vezes chamado de modulo de análise de alavancagem)

que compara um novo texto fonte com uma MT antes de se iniciar a tradução. Isso é feito com o objetivo de computar o número de combinações que provavelmente serão encontradas, assim como o número de repetições internas contidas no texto fonte. Alguns módulos podem também calcular o número de palavras e as unidades de tradução contidas no texto, ignorando elementos tais como rótulos HTML ou códigos de programas que possam influenciar a contagem de palavras. A análise de repetição tem um papel importante na negociação de preços do trabalho de tradução. Ela é também útil em auxiliar clientes e tradutores na estimativa de tempo para entrega de trabalhos de tradução. Devido a essa nova tendência, alguns tradutores estão cobrando seus clientes por hora ao invés de cobrar por caractere, palavra, linha ou página, já que existe trabalho extra envolvido na utilização de MTs (e.g. pré-processamento, conversão de arquivos e manutenção de bancos de dados).

(iv) Propriedade – outra questão muito importante surgiu com advento das MTs, isto é, a quem pertence uma MT. Levando em consideração o fato de que uma MT pode ser um recurso valioso, tanto tradutores quanto clientes parecem ficar ansiosos em reivindicar posse da mesma. Tradutores argumentam que por terem realizado o trabalho, eles deveriam ser os proprietários, já que se não tivessem criado a MT, a mesma não existiria. Clientes, por sua vez, querem proteger suas propriedades intelectuais e não querem que seus concorrentes se beneficiem do trabalho de tradução que pagaram para fazer. Esses clientes argumentam que por terem contratado e pago pelo serviço, eles deveriam ter a posse da MT. Os dois argumentos têm seus méritos, e por se tratar de um conceito relativamente novo na área, não há precedentes legais regendo essas questões. Consequentemente, a posse de uma MT está às vezes sujeita à negociação e deve ser tratada de forma clara e objetiva em contratos para que ambas as partes conheçam seus direitos e deveres.

Funcionamento de um sistema de Memória de Tradução (MT)

Esse tipo de tecnologia funciona através da comparação automática de um novo texto fonte com um banco de dados de textos que já foram traduzidos. Quando o tradutor tem um novo segmento para traduzir, o sistema de MT consulta o banco de dados para verificar se este segmento corresponde a um segmento traduzido anteriormente. Se um segmento correspondente é encontrado, o sistema de MT apresenta ao tradutor uma tradução já realizada daquele segmento. O tradutor pode consultar esta tradução prévia e decidir se irá incorporá-la ou não a sua nova tradução (Bowker, 2002, p. 94). Segue abaixo uma explicação mais detalhada do funcionamento de um sistema de MT.

Segmentação – na maioria dos casos a unidade básica de segmentação de um sistema de MT é a sentença. Isso explica o motivo pelo qual as MTs são conhecidas como memórias de sentenças. Entretanto, nem todos os textos são escritos na forma de sentenças. Cabeçalhos, itens de uma lista e células de uma tabela são elementos familiares de um texto, mas eles podem não ser estritamente considerados sentenças. Portanto, muitos sistemas de MT permitem que usuário defina outras unidades de segmentação além de sentenças. Estas unidades podem ser fragmentos de sentenças e até mesmo parágrafos inteiros.

Combinações – a maioria dos sistemas de MT apresenta ao usuário um número de diferentes tipos de combinações de segmentos. Os tipos mais comuns de combinações são exatas, completas, difusas, de termos, e de subsegmentos.

Combinação exata: as mais óbvias combinações são conhecidas como combinações exatas ou perfeitas. Uma combinação exata é 100% idêntica ao segmento que o tradutor está traduzindo, tanto linguisticamente quanto em termos de formatação.

Quadro 1 – Exemplo de uma combinação exata recuperada de uma MT

Novo segmento fonte	The book is on the table.
Unidade armazenada na MT	EN: The book is on the table. PT: O livro está sobre a mesa.

Qualquer segmento do novo texto fonte que não combinar precisamente com um segmento armazenado na MT não produzirá uma combinação exata. No caso acima, temos um exemplo claro de combinação exata ou perfeita.

Combinação Completa: uma combinação completa ocorre quando um novo segmento fonte difere de uma unidade da MT somente no que diz respeito aos assim chamados elementos variáveis, também conhecidos como "colocáveis" (*placeables*) ou "entidades designadas" (*named entities*). Elementos variáveis incluem números, datas, horas, moedas, medidas e, algumas vezes, nomes próprios. Estes elementos necessitam de algum tipo de tratamento especial no texto. Por exemplo, a maioria dos nomes próprios e nomes de empresas não são geralmente traduzidos, ao passo que datas e horas podem ter seus formatos alterados (e.g. DD/MM/AA pode se tornar MM/DD/AAAA ou 4:00 p.m. pode se tornar 14:00). Em qualquer caso, o número ou nome próprio exato que aparece no segmento do texto fonte, geralmente não afetará como o resto do segmento será traduzido.

Combinação Difusa: nem todos os trechos que um tradutor encontrar terá sido expresso exatamente da mesma maneira em um texto prévio ou diferirá de um texto prévio somente em relação aos elementos variáveis. No entanto, trechos que são similares podem ainda vir a ser úteis. Por esse motivo, muitos sistemas de MT são capazes de localizar combinações difusas, às vezes conhecidas por combinações aproximadas ou parciais. Uma combinação difusa recupera um segmento que é similar, mas não idêntico ao novo segmento fonte.

Quadro 2 – Exemplo de uma combinação difusa recuperada de uma MT

Novo segmento fonte	The specified file is not valid.
Unidade armazenada na MT	EN: The specified file is not a valid file. PT: O arquivo especificado não é um arquivo válido.

O grau de similaridade em uma combinação difusa pode variar de 1 a 99% e o usuário pode estabelecer o limiar de sensibilidade (*sensitivity threshold*) para permitir que o sistema de MT localize segmentos previamente traduzidos que possam divergir levemente do novo segmento textual fonte ou de segmentos que variam muito.

Combinação de Termos – a maioria dos sistemas de MT opera juntamente com bancos de termos. Utilizando programas de gerenciamento terminológicos compatíveis (ver UNIDADE IV), um tradutor pode construir um banco bilíngue de termos e o sistema de MT irá comparar os termos individuais contidos em segmentos do texto fonte em relação aos termos contidos no banco de termos. Este processo é também conhecido como reconhecimento terminológico ativo.

Combinações de Subsegmentos – esses tipos de combinações ficam entre uma combinação difusa e uma combinação de termos. A diferença reside no fato que no caso da combinação de subsegmentos, os elementos comparados são amostras menores dos segmentos. Isso significa que uma combinação pode ser recuperada entre duas pequenas amostras dos segmentos, mesmo se o segmento completo não possua um alto nível de similaridade.

Combinações inexistentes – devido ao fato de que dois textos não podem ser completamente idênticos, haverá provavelmente segmentos onde nenhuma combinação útil será recuperada. Nesses casos, o novo segmento fonte deverá ser traduzido pelo tradutor, embora seja possível que equivalentes para alguns dos termos possam ser localizados em um banco de termos associado. Outra opção seria utilizar um sistema de tradução automática (ver UNIDADE II) para traduzir as partes do texto fonte onde nenhuma combinação foi encontrada na MT. Independente do método utilizado, uma vez que um segmento do texto fonte seja traduzido, o mesmo poderá ser adicionado à MT para que fique disponível para uma eventual reutilização futura.

Sistemas de Memórias de Tradução Disponíveis no Mercado

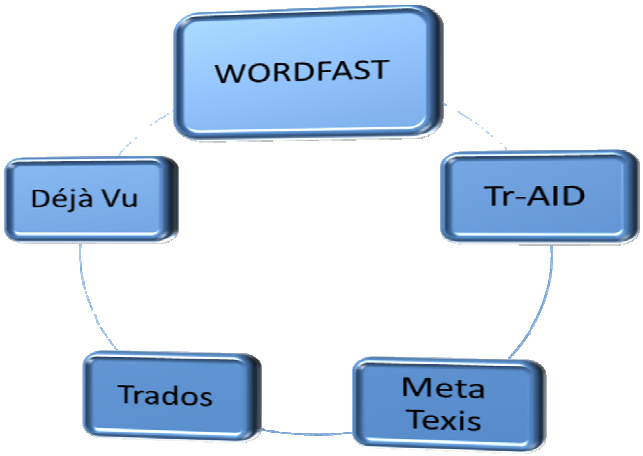


Figura 1 – Os cinco sistemas de MT mais conhecidos no mercado.

Dentre os vários sistemas de memória de tradução disponíveis no mercado, nesta seção, gostaríamos de centrar nossa atenção a cinco desses sistemas (Ver Figura 1 acima). As informações fornecidas sobre esses sistemas são apenas de caráter informacional e, portanto, não pretendem avaliar cada um desses sistemas de MT. Para maiores informações sobre os mesmos, o leitor terá que se referir aos sites dos respectivos PATs fornecidos abaixo.

Déjà Vu

Um dos precursores dos Sistemas de Memória de Tradução, o Déjà Vu surgiu no início dos anos 90. Conhecido por oferecer um conjunto de ferramentas praticamente completa para a tarefa de auxiliar o tradutor, o Déjà Vu contém desde ferramentas básicas para trabalhar como memórias de tradução, como também um conjunto bastante grande de ferramentas que operam em conjunto.

Dentre as principais características estão: a seleção de segmentos pré-traduzidos, em que o sistema escolhe entre a melhor tradução existente na memória de tradução; a propagação, onde se não existe nenhum segmento já traduzido no banco de dados, ele procura por todos os segmentos idênticos no texto e automaticamente inclui a tradução que você realizará para este primeiro segmento; a possibilidade de compartilhar o projeto de tradução entre os tradutores membros do projeto em execução; o gerenciamento terminológico, que faz com que a terminologia seja uniformizada na tradução; possui ainda, uma ferramenta para alinhamento de textos paralelos que podem ser convertidos em memórias de tradução; e é capaz de compartilhar memórias de tradução entre os principais sistemas existentes, com uma função importar e exportar bastante avançada.

Para que o usuário possa utilizar esse sistema de memória de tradução, os requerimentos necessários, de acordo com o fabricante, são: o uso do sistema operacional Windows 98/ME/NT4/2000/XP/Vista em um computador com uma configuração mínima de Pentium III com processador de 600 MHz e 256 MB de RAM.

(Fonte: http://www.atril.com/)

Trados

Um dos programas mais conhecidos entre a comunidade de tradutores, o Trados teve sua grande evolução no final dos anos 90. Por utilizar-se de um ambiente mais limpo para o usuário final, ele obteve seu espaço dentre as maiores agências de tradução e logo tradutores autônomos passaram a utilizá-lo. Assim como outros sistemas de memória de tradução, o Trado traz recursos essenciais para auxiliar o tradutor, como também um conjunto grande de ferramentas que podem ser integradas a ele, para que seja possível ter um maior aproveitamento da ferramenta como um todo.

Uma das principais características do Trados é a possibilidade de se trabalhar em conjunto com sua ferramenta de gerenciamento terminológico, o MultiTerm. Com isso, ele é capaz de unir as funções avançadas de "combinação difusa" entre os segmentos existentes na memória

de tradução, e aliar às buscas em um banco terminológico preparado para um projeto específico de tradução, garantindo maior rendimento e produtividade à tradução e assegurando a qualidade da tradução. Como os outros sistema de memória de tradução, o Trados traz as ferramentas de alinhamento, o WinAlign, e ferramentas que possibilitam trabalhar com arquivos que contenham códigos específicos, como HTML, XML e outros.

O requisitos necessários para a utilização do Trados são: sistema operacional Windows 98/ME/NT4/2000/XP/Vista em um computador com uma configuração mínima de um Pentium II com pelo menos 64 MB de RAM e o programa de processamento de textos Microsoft Word.

(Fonte: http://www.trados.com/en/)

MetaTexis

Um nome bastante conhecido entres os tradutores que utilizam sistemas de memória de tradução é o MetaTexis. Desenvolvido por um tradutor para tradutores, o MetaTexis é uma ferramenta de auxílio à tradução que compreende um conjunto enorme de recursos tais como a tradução propriamente dita, a revisão, o alinhamento e o gerenciamento terminológico que envolve um projeto de tradução.

As principais características do MetaTexis são: a interface de trabalho, que é executada dentro do ambiente do Microsoft Word; a compatibilidade com os vários formatos de arquivos disponíveis e, também, entre os arquivos utilizados pelo Trados e Wordfast; a possibilidade de importar e exportar arquivos de memória de tradução no formato padrão TMX, compartilhados pela maioria dos sistemas de memória de tradução; uma ferramenta de alinhamento de textos paralelos, como no outros sistemas apresentados, possibilitando criar memórias de tradução a partir de textos já traduzidos; um recurso importantíssimo para que em casos acidentais de apagar alguns códigos inseridos pelo programa durante a tradução, seja possível a recuperação dos mesmos; e, além de tudo, a possibilidade de se trabalhar em modo de servidor on-line.

Para que seja possível utilizar o MetaTexis, um computador com os requisitos mínimos de sistema operacional Windows 98/ME/NT4/2000/XP/Vista com uma configuração de Pentium III 700MHz com 256MB de RAM.

(Fonte: http://www.metatexis.com/)

Tr-AID

O Tr-AID nasceu a partir de muitas pesquisas realizadas para se conseguir obter uma plataforma em que fosse possível a realização do trabalho de tradução em um ambiente totalmente informatizado, ou seja, usando textos em formato eletrônico e com o auxílio do computador e trazendo como principais benefícios a qualidade, a consistência, a uniformização, a reciclagem e a produtividade da tradução.

O Tr-AID possibilita ao usuário tirar proveito dos mais variados recursos durante a tarefa

tradutória. Recursos esses que vão desde poder trabalhar com uma grande quantidade de textos; assegurar a qualidade das traduções realizadas, e, consequentemente, aproveitar os textos traduzidos, aumentando a produtividade; garantir a consistência e a uniformização da tradução por completo; e tornar o gerenciamento terminológico mais fácil de ser utilizado pelos tradutores. Todas essas características colocam o Tr-AID como um grande sistema de memória de tradução.

Para que o usuário possa utilizar o Tr-AID e seus recursos, ele deverá ter um computador que possua o sistema operacional Windows 98 e uma configuração razoável para que se possa trabalhar com o Microsoft Word.

(Fonte: http://www.ilsp.gr/traid_eng.html)

Wordfast

Desenvolvido em 1999, o Wordfast é um Programa de Apoio à Tradução baseado na plataforma Microsoft Office, sendo executada como uma macro.

O princípio básico deste sistema de MT é tornar acessível à utilização de um sistema de memória de tradução.

Apresentando uma interface simples, o Wordfast logo tomou lugar de importância dentre os outros concorrentes existentes no mercado.

O Wordfast utiliza o formato padrão de codificação de memórias de tradução, o formato TMX (Translation Memory eXchange), o que, por sua vez, possibilita a transferência de dados entre os diversos sistemas de memória de tradução. Assim, muitos usuários puderam, de forma simples, migrar para o Wordfast.

Trabalhando com memórias de tradução, sistema de terminologia com glossários ativos durante a atividade tradutória, o Wordfast oferece ao tradutor um conjunto de ferramentas que possibilita conseguir um resultado de qualidade.

Pelo fato de que o tradutor tem total autonomia durante o processo de tradução, consegue-se obter um ganho de produtividade em termos qualitativos e quantitativos.

Dentre os pontos principais de utilização do Wordfast destacamos:

- Memória de Tradução (Formato Padrão – TMX);
- Gerenciamento de Terminologia;
- Glossários Ativos;
- Preservação da formatação do texto;
- Dados estatísticos após a finalização da atividade de tradução;
- Texto final pronto para entrega ao cliente.

(Fonte: http://www.wordfast.com/)

Assim, com todos esses itens destacados acima, o Wordfast apresenta-se como uma ferramenta que consegue oferecer as principais funções exigidas pelo tradutor durante sua atividade. Além disso, devido ao seu baixo custo, aprendizado rápido – utiliza-se da plataforma Microsoft Word – e o total controle do texto fizeram deste PAT a ferramenta ideal para ser utilizada em um projeto de tradução sobre textos na área de língua de sinais. Oferecemos agora, uma ilustração prática da utilização deste sistema de MT no Projeto PROLIBRAS-TRAD.

O projeto PROLIBRAS-TRAD

O Projeto PROLIBRAS-TRAD foi um projeto de tradução que teve como objetivo traduzir 20 artigos sobre Línguas de Sinais selecionados e extraídos do **TISLR 9 (Theoretical Issues in Sign language Research 9) – 9º Congresso Internacional de Aspectos teóricos das Pesquisas nas Línguas de Sinais** – sediado pela Universidade Federal de Santa Catarina – UFSC, Florianópolis, SC, em dezembro de 2006.

Os textos tinham em média vinte (20) páginas, cerca de cinco mil (5.000) palavras e foram recebidos em formato eletrônico com a extensão de arquivo [.doc], arquivo do Microsoft Word, integrante do pacote MS-Office. Para visualizar os textos completos do congresso, acesse: http://www.editora-arara-azul.com.br/ebooks/catalogo/abertura.pdf.
O projeto de tradução foi realizado por uma equipe, formada por uma Coordenação Geral (CG), um assistente de CG, junto a coordenação geral estão dispostas três Coordenações de Equipe (C1, C2,C3), cada uma ligada a seus Gerentes de Trabalho (G1,G2,G3), que são responsáveis pelos Assistentes de Tradução (A1,A2,A3), Conforme representado na Figura: Organograma Funcional – PROLIBRAS-TRAD.

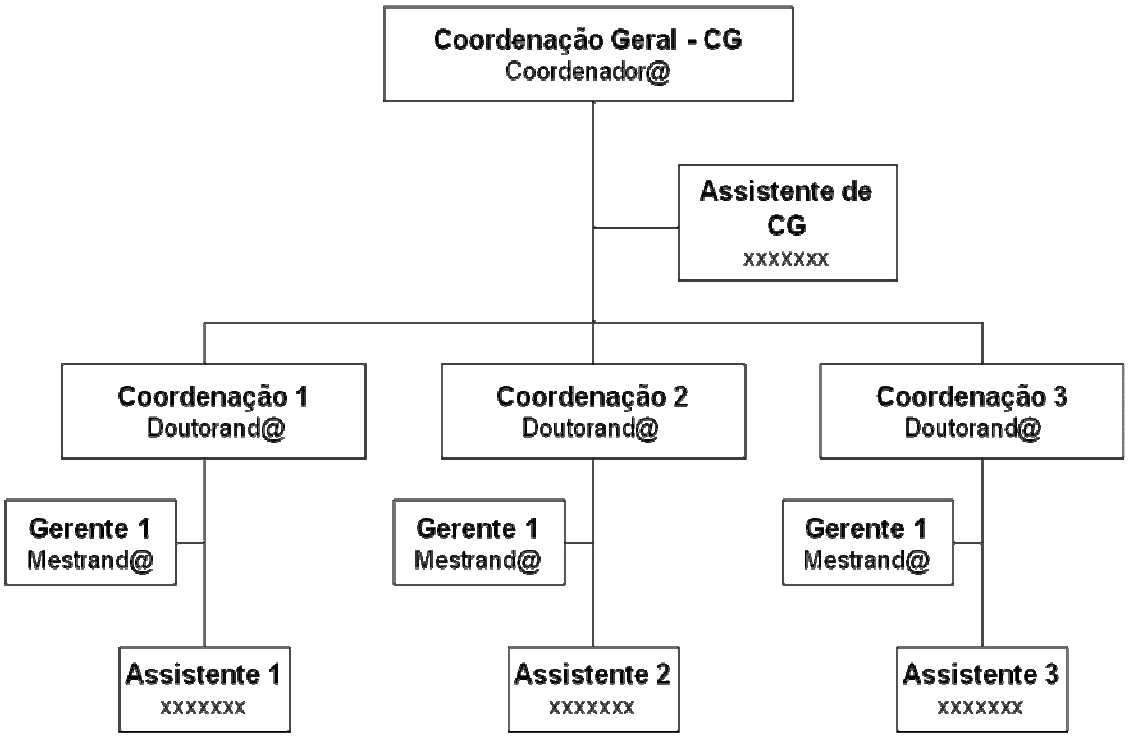


Figura: Organograma Funcional – PROLIBRAS

O tempo para execução do PROLIBRAS-TRAD foi de quatro meses, onde ao final do projeto, os artigos traduzidos tornaram-se um livro publicado pela Editora Arara Azul (http://www.editora-arara-azul.com.br/ebooks/catalogo/completo_port.pdf).

A necessidade da utilização de um PAT

Desde o início do projeto foi discutida a necessidade da utilização de um programa de auxílio à tradução. Devido à realização do trabalho de tradução ser realizado em equipe, o uso de um PAT poderia proporcionar melhor qualidade de trabalho, com relação à formatação dos textos e ainda propor um ambiente de trabalho confortável ao tradutor. Outro motivo é fato que por serem textos de conteúdo técnico, são textos que contém uma terminologia específica, e como o trabalho seria realizado em equipe, todos os tradutores deveriam trabalhar com a mesma terminologia, para que a tradução fosse coerente em seus resultados. A quantidade de textos a serem traduzidos versus ao tempo para serem realizados foi outro fator que impulsionou a utilização de um PAT.

A escolha do Wordfast

Vários PATs estão a disposição dos tradutores, dentre eles podemos encontrar ferramentas de diferentes classes, funções, preços e disponibilidade. Dentre elas estão:

PAT	**Preço**	**Configurações Mínimas**	**Tamanho do Arquivo**

TRADOS	US$ 995.00=R$1947,32	Windows-XP>Vista, Pentium IV, 512Mb>1GB.	120 Mb
WORDFAST	250.00 Euros=R$675.68 Versão Demo	Windows95>XP-, Word97>XP 120Mhz Processor, 128Mb.	550Kb
Déjà Vu X Standard	US$ 668.00=R$1307,29	Windows98>Vista, Pentium III, 600Mhz, 256Mb.	78,5 Mb

Muitos desses fatores devem ser observados na escolha de um PAT, são elementos fundamentais no planejamento de um projeto de tradução e requer uma ampla discussão sobre suas condições, preços funções e requerimentos necessários para um bom rendimento.

O fato que os computadores do laboratório disponível para o processo de tradução dos textos do Projeto PROLIBRAS-TRAD estavam equipados com o MS-Word versão 2003, e os arquivos com os textos originais a serem traduzidos estavam em formato [. *doc*], documento do próprio MS-Word, o Wordfast foi a escolha certa considerando os argumentos acima apresentados e, portanto, foi eleito como o Programa de Auxílio ao Tradutor para a realização das traduções durante o Projeto PROLIBRAS-TRAD. Dentre outros fatores estão o tempo que os participantes do projeto levariam para aprender a utilizar esta ferramenta, por contar com uma interface simples e já conhecida (baseia-se na plataforma Microsoft Word), e as funções disponíveis de uso no projeto, que correspondem a demanda do mesmo.

Pontos Principais

- Memórias de Tradução (MTs) alinham textos fonte e alvo e armazenam os seguimentos alinhados em um banco de dados;
- A ideia principal por trás de um sistema de MT é que ele permite a reutilização ou a reciclagem de segmentos traduzidos anteriormente. O sistema automaticamente compara um novo texto fonte com o banco de dados de traduções prévias;
- A combinação dos segmentos pode acontecer em diferentes níveis: combinação exata, combinação difusa, combinação de termos, ou combinação de subsegmentos;
- O uso das MT gerou algumas questões controversas em relação à propriedade e remuneração;
- As MTs podem ser integradas a outras ferramentas, como, por exemplo, processadores de texto, sistemas de gerenciamento terminológicos, concordanciadores bilíngues e sistemas de tradução automática, assim criando uma “estação de trabalho” ou uma “bancada” integrada para o tradutor;
- O Projeto PROLIBRAS-TRAD pode ser utilizado para ilustrar a utilização bem sucedida de um sistema de MT (*Wordfast*) em um projeto de tradução.

Referências

AUSTERMÜHL, Frank. Electronic Tools for Translators. Manchester, UK: St. Jerome Publishing, 2001.

BARTHOLAMEI, Lautenai. Wordfast: Utilização e Avaliação em um Projeto de Tradução. Monografia de Especialização em Língua Inglesa: Ênfase em Tradução. Chapecó, SC: UNOCHAPECÓ, 2008.

BOWKER, Lynne. Computer-Aided Translation Technology. A practical introduction. Ottawa: University of Ottawa Press, 2002.

HEYN, Matthias. Translation Memories: Insights and Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.). Unity in Diversity? Current Trends in Translation Studies. Manchester, UK: St. Jerome Publishing, 1998.

MELBY, Alan e WARNER, Terry C. The Possibility of Language: A Discussion of the Nature of Language with Implications for Human and Machine Translation. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

UNIDADE II

Sistemas de Tradução Automática

Esta unidade discute os Sistemas de Tradução Automática (Sistemas de TA), como já visto na disciplina de Introdução aos Estudos da Tradução. Está dividida em quatro partes principais: (i) um breve apanhado histórico do surgimento dos sistemas de TA; (ii) a definição do que vem a ser um sistemas de TA, (iii) seu funcionamento através dos vários tipos de abordagens empregadas e, consequentemente, os problemas enfrentados por estes sistemas, e finalmente (iv) algumas discussões que permeiam a utilização deste recurso no que diz respeito às línguas de sinais.

Histórico

Ideias iniciais sobre processos de TA se deram ainda no século XVII, quando filósofos como Leibniz e Descartes propuseram a codificação de palavras que se relacionavam entre as línguas, porém esse trabalho se deu apenas de cunho teórico e não influenciou o desenvolvimento dos sistemas de TA (Hutchins, 2001).

Os primeiros sistemas de TA surgiram ainda na década de 30. Em 1933, Smirnov-Trojanskij, apresentou um mecanismo que possibilitava a tradução entre diversas línguas de forma simultânea, porém linguistas russos que tratavam da tradução automática não consideraram o sistema desenvolvido por Smirnov-Trojanskij. Em 1946, alguns cientistas fizeram tentativas para a realização de uma automática utilizando calculadoras científicas. Essas calculadoras eram alimentadas por um conjunto de dados pequeno e proporcionavam apenas uma tradução palavra-por-palavra (MATEUS, M. H. M. et al., 1995).

Considerado como o pioneiro da TA, o *Weaver Memorandum* em 1949 também era um sistema que traduzia, automaticamente, segmentos entre a língua russa e a língua inglesa. Após esse marco, o americano Warren Weaver e o inglês Booth, criadores do sistema, se convenceram que a TA podia realizar o processo de tradução totalmente automatizada e alcançar os objetivos propostos.

O grande progresso da TA se deu a partir da década de 50. O primeiro experimento com tradução automática foi a realização da tradução de aproximadamente 60 sentenças entre a língua russa e a língua inglesa em 1954, conhecido como *Georgetown-IBM Experiment*. Esse experimento utilizou-se de um sistema lexicográfico bastante restrito, com aproximadamente 250 entradas e baseado em apenas seis regras gramaticais. Mesmo com todas essas limitações, o experimento foi realizado com sucesso e impulsionou o crescimento de pesquisas em TA.

O relatório realizado pelo ALPAC (John R. Pierce, John B. Carroll, et al., 1966) – Automatic Language Processing Advisory Committee [Comitê Assessor de Processamento Automático das Línguas] – na década de 60 avaliou de forma negativa a qualidade dos diversos sistemas de TA que existiam até o momento. Como consequência deste relatório, recursos que eram disponibilizados para a realização de pesquisa na área de TA foram cessados. As pesquisas em TA somente tiveram novos recursos a partir dos anos 80, onde houve muitas melhorias

nos sistemas desenvolvidos, dentre estes sistemas podemos citar o SYSTRAN e o EUROTRA.

A partir dos anos 90, a IBM apresenta o primeiro protótipo de um sistema de TA totalmente estatístico (ver Sistemas de TA: Abordagem Estatística), oferecendo, assim, novas direções para a TA. Atualmente, vários sistemas desenvolvidos com base em seus antecessores estão em desenvolvimento, sendo esses recursos são difundidos, na maioria dos casos, por meio da Internet.

Definição

Sistemas de Tradução Automática, como o próprio nome sugere, são sistemas capazes de realizar, por meio de um dispositivo computacional, uma tradução de forma automatizada, sem a necessidade de um agente humano durante a realização desta tarefa (Hutchins & Somers, 1992, p. 3). Os sistemas de tradução automática são conhecidos pelo acrônimo em inglês MT, já em português é designado como TA, para Tradução Automática, e Sistemas de TA, para Sistemas de Tradução Automática.

Segundo a *European Association of Machine Translation* (EAMT – Associação Europeia de Tradução Automática), entende-se por tradução automática a atividade tradutória que é realizada totalmente por um sistema computacional automatizado.

> A Tradução Automática (TA) é um programa de computador para a tarefa de traduzir textos de uma língua natural para outra. Uma das mais recentes atividades em ciência da computação, a TA provou ser um objetivo ilusório, porém atualmente diversos sistemas estão disponíveis, os quais produzem resultados que, senão perfeitos, são de qualidade suficiente para ser útil em diversos domínios específicos (EAMT, 1997, nossa tradução).

Inicialmente, esse tipo de tradução foi descrito não como *machine translation* [tradução por máquina], intitulado atualmente em inglês, mas sim nomeado como *automatic translation* [tradução automática], com seus equivalentes em francês, *traduction automatique*, em russo, *avtomaticeskii perevd*, e que também é nomeado em português como *tradução automática*. Porém, quando iniciados os estudos sobre a possibilidade de se traduzir de forma automática, estes sistemas ainda não incluíam bancos de dados com anotações complexas, bancos terminológicos com abordagens estatísticas avançadas e outros recursos que somente tornaram-se possíveis com a evolução da tecnologia.

Desde sua criação, os sistemas de TA passaram por diversas discussões com relação ao seu funcionamento, pois se acreditava, e ainda acredita-se, que sempre há a necessidade de revisão e edição posterior a realização de uma tradução realizada de forma automática. Este fato ocorreu devido à comparação feita com traduções realizadas por tradutores humanos, que tinham a possibilidade de tratar destes elementos durante a atividade tradutória, logo também sendo possível a re-edição e revisão do texto alvo. Desse modo, varias outras tentativas de nomear estes sistemas surgiram, dentre elas a Tradução Humana Auxiliada por Máquina (MAHT – *Machine-Aided Human Translation*) e a Tradução por Máquina Auxiliada por Humanos (HAMT – *Human-Aided Machine Translation*), que após certo período passaram a serem definidos como pertencentes à um único tipo de sistema nomeado Tradução Assistida por Computadores (CAT – *Computer-Aided Translation*), o qual tem a função de oferecer

sistemas capazes tanto de serem auxiliados por humanos ou por máquinas e vice-versa, assim com também ferramentas para gerenciamento terminológico e gerenciamento de projetos de tradução.

O fato de a tecnologia computacional e as pesquisas no campo do processamento natural da linguagem terem sido altamente desenvolvidas e executadas durante os últimos 60 anos e contribuído para o avanço dos Sistemas de TA, ocorreu por ocasionar um aumento de qualidade e precisão no resultado final de uma tradução realizado por este sistema (ver Histórico). Com o objetivo de se chegar à uma tradução o mais inteligível possível, porém não deixando de ser um sistema totalmente automatizado, hoje a TA é uma ferramenta que torna acessível, mesmo com todas suas deficiências, um grande conteúdo de informações dispostas, principalmente, em formato eletrônico serem distribuídas em diversos idiomas.

Abordagens dos Sistemas de TA

Durante aproximadamente 60 anos de pesquisa e testagem, os sistemas de TA encontraram muitos elementos que influenciaram diretamente seus resultados. Diversas abordagens foram utilizadas com esses sistemas e a evolução desses sistemas ocorreu a partir dos problemas detectados nas traduções geradas pela TA.

Desde a ocorrência de seus problemas iniciais, os sistemas de TA foram abordados de diferentes formas no que se refere à abordagem de tradução empregada na engenharia base do sistema, o qual é responsável pelo seu funcionamento. Inicialmente, o principal elemento observado se deu com relação às regras que diferenciavam os sistemas linguísticos de uma língua para outra, sendo que cada língua possui um conjunto linguístico regido por um sistema interno de regras. Com base nesse problema inicial que os sistemas de TA enfrentaram, podemos apontar algumas abordagens que atualmente são discutidas e auxiliam o processo de evolução dos Sistemas de TA, tornando-os mais eficientes. Dentre elas podemos destacar: a abordagem lexicográfica (dictionary-based approach), abordagem exemplária (example-based approach), a abordagem interlíngua (interlingua approach) e a abordagem estatística (statistical approach).

Um dos principais problemas enfrentados na TA foi o fato que a tradução não é apenas a transferência de palavras entre línguas, ou seja, transferência de uma palavra em língua fonte para outra palavra que seja equivalente da mesma em uma língua alvo (Arrojo, 1997, p. 22). Mesmo coma existência de um possível equivalente na língua alvo, o problema ainda não é resolvido no caso em que ocorre a inexistência de um possível equivalente na língua alvo, e ocasiona um problema ainda maior quando ocorre a existência de mais do que um único possível equivalente para uma mesma palavra na língua alvo.

Como base do processo de tradução, o primeiro elemento destacado é a *análise* do texto fonte, o qual o sistema se encarrega de decodificá-lo e prepará-lo para a realização da tradução. Logo, temos três abordagens, ainda empregadas na TA, para que seja possível a *geração* de um novo texto na língua alvo.

Das abordagens utilizadas, a primeira delas é a *tradução direta*, abordagem este que tem como principal ferramenta a utilização de dicionários para que a tradução seja realizada,

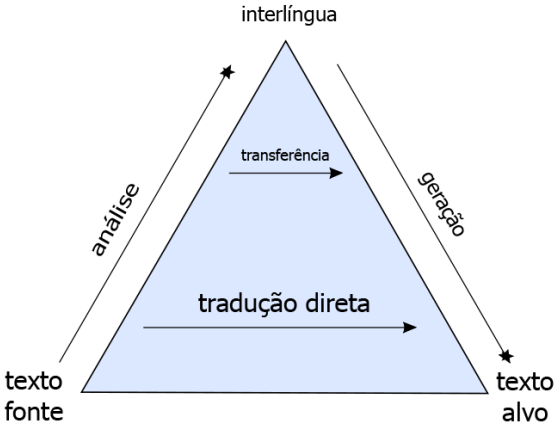
abordagem que é comumente conhecida pela utilização de um sistema *lexicográfico*, proporcionado pela utilização de dicionários. Discutido na seção Sistemas de TA: Abordagem de Tradução Direta, Lexicográfica.

Por segundo, temos a abordagem de tradução por *transferência*. Essa abordagem é responsável por buscar segmentos, fragmentos, já traduzidos em um determinado corpus e transferir estes segmentos semelhantes para a língua alvo. Comumente conhecido como a abordagem de tradução automática que utiliza um sistema *exemplário*, baseado em exemplos, essa abordagem vai além do sistema *lexicográfico*, considerando também segmentos e fragmentos, não apenas realizando a substituição de palavras isoladas. Esta abordagem será discutida na subseção Abordagem de Tradução por Transferência, Exemplaria.

Por fim, temos a abordagem de tradução *interlíngua*, abordagem que se utiliza de recursos estatísticos e compreende uma utilização mais elaborada das outras duas abordagens citadas anteriormente. Essa abordagem, geralmente utilizada em programas mais recentes, é a capacidade de realizar operações com base em um sistema interlíngua, onde a língua fonte é analisada e representada como uma língua independente e que, a partir desta representação, o texto alvo é gerado. Essa abordagem será discutida na seção: Sistemas de TA: Abordagem de Tradução Interlíngua.

Temos ainda outra abordagem, não destacada no diagrama, que se utiliza de estatística para a realização de uma tradução automática, esta abordagem é utilizada pelos sistemas mais conhecidos, por fazer uso de todas as abordagens já mencionadas, formando um conjunto capaz de processar a língua cuidadosamente em cada um de seus elementos. Essa abordagem será discutida na seção: Sistemas de TA: Abordagem de Tradução Estatística.

Podemos visualizar no diagrama abaixo como uma tradução é realizada por meio de um processo automático, destacando três abordagens inicialmente discutidas por Vauquois (1968) e utilizadas pelos sistemas de TA.



A ilustração de Vauquois (*ibid.*) contempla apenas três das diferentes abordagens discutidas: (i) o abordagem de tradução direta, (ii) a abordagem de tradução por transferência e (iii) a abordagem de tradução por interlíngua.

Para os sistemas de TA que tratam da tradução automática de línguas de sinais, devemos citar que além de estes sistemas serem repensáveis pela tradução, ainda existe a necessidade de um programa que torne possível a realização do processo de reconhecimento de voz, no caso de um não-sinalizante, e é exigido para que depois esta língua seja traduzida em uma língua de sinais, no caso em seja feita uma tradução para um sistema verbal.

Já para o caso de línguas de sinais escritas, destacado nesta unidade, concentramos em discutir a existência de métodos e programas disponíveis, como também a ocorrência de problemas na tradução das mesmas, com base nos principais problemas acima citados. Para apresentarmos os métodos utilizados para a tradução automática de línguas de sinais, descreveremos os principais recursos desenvolvidos para que se torne possível a tradução automática, o principal deles, o sistema *Sign Writing*, adotado por diversos países para a escrita de sinais, e o mais utilizado em sistemas de tradução automática. Todos esses métodos serão discutidos na seção: Sistemas de TA: Línguas de Sinais.

Sistemas de TA: Abordagem de Tradução Direta

Lexicográfica

Uma das abordagens utilizadas para realizar o processo de tradução em um dispositivo computacional é a abordagem lexicográfica, a qual transforma o texto (SANTOS, 2006). Em uma abordagem lexicográfica, cada palavra é interpretada com sendo uma palavra isolada, sem considerar o contexto e/ou outros elementos que constituem uma unidade de tradução.

O princípio dos sistemas lexicográficos de TA se dá pela ‘substituição direta’ das palavras em uma sentença, sendo simplesmente realizada a substituição de palavra-por-palavra, ou seja, uma palavra em um língua fonte é substituída por uma palavra, que conste na base do dicionário, na língua alvo. Geralmente, estes sistemas são empregados para realizar traduções de grande escala, onde o nível de processamento de busca é realizado em bancos de dados com um formato semelhante ao de dicionários, a partir de uma entrada, são substituídos por equivalentes diretos.

Os problemas que este abordagem apresentou foram constatados já em seus primeiros testes, ainda durante a Segunda Guerra Mundial, na tradução das mensagens entre os americanos e os russos. Um dos exemplos mais famosos é:

Original em Russo: *My trebuem mira*
Tradução Automática par o Inglês: *We require world*
Tradução realizada por Humano: “*We want peace*”

Na sentença acima, original em russo, a mensagem traduzida por este sistema para a língua inglesa como “We require world”, em português “Nós queremos o mundo”, ao invés de uma tradução mais correta realizada por um tradutor humano “We want peace”, em português, “Nós queremos paz”.

Como desde os primeiros testes esta abordagem apresentou várias deficiências, os sistemas lexicográficos de TA passaram por diversas mudanças. Entre as principais mudanças que ocorreram nesta abordagem, destaca-se a utilização de dicionários baseados em corpora, para

que os níveis de equivalência de uma palavra em uma língua fonte e alvo possam ter diferentes equivalentes potenciais em uma tradução. Os sistemas lexicográficos de TA são responsáveis pela base da evolução de outros Sistemas de TA, principalmente para a abordagem estatística.

Como exemplos de alguns sistemas que utilizam este tipo de abordagem, podemos citar o programa *Power Translator*, uma aplicação computacional que tem como base uma arquitetura de funcionamento por meio da tradução direta. O *Power Translator* está em sua 11ª edição e tem como principais vantagens a tradução de documentos, blogs, e-mails, páginas de Internet, mensagens instantâneas e outros. Possui a possibilidade de tradução entre 7 línguas: inglês, francês, alemão italiano, português, russo e espanhol. O *Power Translator* disponível em diversão versões, *Premium*, *Personal*, *Pro*, *Euro* e *World*, pode ser encontrado no sítio http://www.lec.com.

Sistemas de TA: Abordagem de Transferência

Exemplaria

Sistemas exemplários de TA, uma abordagem que se deu pela evolução dos corpora de tradução, que passaram a utilizar anotações complexas e desenvolver um papel de extrema importância na tomada de decisão, por parte do dispositivo computacional utilizado, em optar por uma ou outra determinada tradução de uma sentença, considerando as unidades de tradução já existentes em seu banco de dados. Este abordagem por sua vez procura fazer uma tradução por meio da analogia entres os segmentos nos textos (SANTOS, 2006).

Conhecidos como EBTM (Example-based Machine Translation) estes sistemas trabalham com o sistema de decomposição de uma sentença em pequenas expressões, assim o processo de tradução baseia-se nessas pequenas expressões, realizando buscas em um determinado corpus bilíngue de tradução para que, após encontrar possíveis equivalentes de tradução para estas expressões, as sentenças possam ser recompostas novamente e geradas no texto alvo. Portanto, o processo é, em um primeiro momento, a fragmentação de um grande período em pequenos períodos para facilitar a localização de uma possível tradução para cada fragmento e realizar a reconstrução do conjunto com todos estes fragmentos, realizada para que seja possível a reconstrução de uma unidade de tradução completa.

Esta abordagem foi apresentada pela primeira vez em 1984 por Nagao Makoto, que utilizou um corpus bilíngue de tradução (ver UNIDADE IV), composto de textos em língua japonesa e em língua ingleses devidamente preparados para o processo de tradução automática. O maior problema surgido pela utilização dessa abordagem se deu pelo fato que a língua japonesa contém elementos totalmente diferentes em sua estrutura com relação à língua inglesa, sendo necessária uma análise mais detalhada de cada segmento em uma sentença, para que a estrutura de cada língua possa ser interpretada durante o processo de fragmentação e também para que a reconstrução seja precisa na estrutura da língua alvo.

Dentre os sistemas de TA que utilizam desta abordagem podemos citar sistema distribuído pela *Lernout & Hauspie* com o programa T1 e o *Linguatec's Personal Translator*. Estes sistemas podem ser encontrados nos sítios eletrônicos http://www.lhs.com/tm/t1 e http://www.linguatec.de, respectivamente.

Sistemas de TA: Abordagem de Tradução Interlíngua

Nesta abordagem, o texto passa por um processo onde é transformado em interlíngua e a tradução (i.e. o texto alvo) é gerada a partir desse texto interlíngua, o qual se procura chegar ao sentido de cada elemento para a realização de uma tradução (SANTOS, 2006). A abordagem interlíngua utiliza-se principalmente de elementos de sistemas fornecidos com base em estudos de inteligência artificial.

A ideia de utilização desta abordagem surgiu em 1969, a partir das discussões do filósofo israelense Yehoshua Bar-Hillel que tinha como fundamentos uma tradução não baseada apenas em um processo mecânico, mas sim utilizando processos computacionais. Assim, o texto deixaria de ser apenas uma tradução direta ou uma transferência, mas passaria a ter uma relação de sentido entre as línguas envolvidas.

Por ser uma abordagem totalmente baseada em regras, esta abordagem encontrou muitas limitações, principalmente quando havia a necessidade de ser feita uma tradução de grande escala, necessitando da utilização de um corpus e um conjunto de regras baseado em um domínio geral. Porém, em situações na qual estas regras e o corpus eram exigidos a executar uma operação de tradução com um texto fonte de domínio de cunho específico, esta abordagem conseguia realizar essas operações mostrando um alto grau de qualidade e precisão.

Um dos principais exemplos que podemos citar desta abordagem é sua utilização na tradução de textos de domínio específicos no Japão, o sistema de tradução automática chamado Fujitsu. Por ser tratar apenas de textos de linguagem específica e restrita, o corpus preparado e o conjunto de regras tornou-se eficaz para a realização dessas traduções.

A abordagem interlíngua, por utilizar deste conjunto de regras fechadas, logo deixou de ser um dos principais objetos de estudo, devido surgimento de outras abordagens mais sofisticadas.

Sistemas de TA: Abordagem de Tradução Estatística

Sistemas de TA de abordagem estatística são os sistemas mais conhecidos e difundidos atualmente, diversos serviços disponíveis na Internet possuem sistemas de TA que adotam esta abordagem de funcionamento. Esta abordagem utiliza-se também de outras abordagens como a lexicográfica e a exemplaria, porém com uma preparação mais apurada e mais complexa. Além de se utilizar de funções dispostas por cada um dos outras abordagens já mencionadas, os sistemas de TA de abordagem estatística são capazes de fazer buscas altamente complexas em um determinado corpus bilíngue de tradução, tanto de domínio geral, quanto de domínios específicos.

Devido à alta complexidade das funções de busca, os sistemas de TA que utilizam abordagem estatística têm em sua composição corpora anotados e codificados, níveis de anotação tais como, classe gramatical, classificação morfológica e sintática, entre outros. Funções que possibilitam a realização de cálculos estatísticos em relação às palavras que possam constar

em um dicionário, como no caso da abordagem lexicográfica e exemplaria, onde se considera a existência de fragmentos de possíveis equivalentes em um corpus e que tem como principal função a opção de selecionar uma determinada palavra que possa corresponder à sua tradução para um determinado contexto. Perguntas como: qual é a palavra mais correta para esta situação? E por que a mesma tradução não corresponde a outro contexto? São os elementos principais destacados por este tipo de abordagem.

Isso somente torna-se possível devido à anotação específica de cada corpus, sendo possível uma análise minuciosa nos dados e para que o resultado possa ser melhor do que aqueles que são apresentados ou simplesmente pela substituição de palavra-por-palavra, ou a tradução por fragmentos únicos em uma sentença.

Sistemas de TA: Língua de Sinais

Sistemas de Conversão em Língua de Sinais

A língua de sinais, ainda como uma área de pesquisa recente, já possui alguns sistemas que são capazes tanto de traduzir uma determinada língua escrita para uma língua de sinais, utilizando a abordagem de conversão de caracteres e a tradução entre essas línguas por meio de uma linguagem de escrita denominada *Sign Writing*, como também sistemas que são capazes de proporcionar a tradução de uma língua falada em uma língua de sinais por meio de recursos computacionais que utilizam bonecos animados (avatares).

Nesta primeira seção destacaremos a tradução de línguas escritas, como o português e o inglês para línguas de sinais, utilizando da tradução/conversão do alfabeto latino no alfabeto da soletração manual da língua de sinais.

Sistemas de simples conversão do alfabeto latino foram desenvolvidos para que, ao menos, as letras fossem convertidas em um alfabeto de soletração manual, assim se deu os primeiros passos para o desenvolvimento de um sistema de tradução automática para línguas de sinais. Dentre os sistemas de conversão entre esses alfabetos podemos utilizá-los encontramos:

- Fingerspelling Translator – http://www.webstantaneous.ws/swfs/nofind/asl/translate.swf [Tradução/Conversão do alfabeto latino no alfabeto manual da ASL, utilização de desenhos que representam o alfabeto manual da ASL].

- Fingerspelling Machine – http://www.bsldictionary.com/bsvid/fsvids/fsbsl.swf [Tradução/Conversão do alfabeto latino para o alfabeto manual da BSL, utilização de um sinalizante realizando a representação de cada letra do alfabeto manual da BSL].

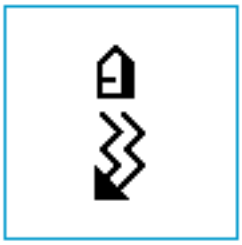
Com estas aplicações torna-se possível a tradução para o alfabeto manual, porém não possível uma tradução propriamente dita entre estas línguas, pois apenas faz a conversão desses elementos. Isso tornou o desenvolvimento de escrita de línguas de sinais e sua tradução de forma automatizada possível, logo diversos sistemas de escrita em línguas de sinais foram desenvolvidos, conseguindo um grau de complexidade na representação de cada elemento da língua, não utilizando apenas letras nesse modelo de tradução, mas com a possibilidade de

utilizar sinais que representam palavras e, até mesmo, expressões. Um desses sistemas é o sistema *Sign Writing*.

O Sistema *Sign Writing*

Desenvolvido para que fosse possível realizar a escrita da língua de sinais, o sistema *Sign Writing* http://www.signwriting.org, foi criado por Valerie Sutton e faz parte de um projeto chamado *Sutton Movement Writing & Shorthand*, e após um grande período de evolução, tornou-se um sistema utilizado em mais de 30 países, incluindo o Brasil. Temos aqui uma ilustração dos países que utilizam o sistema *Sign Writing*.

No sistema *Sign Writing*, os sinais são representados de acordo com a própria sinalização da língua de sinais, porém de maneira gráfica, na representação a seguir temos a transcrição da palavra Brasil:



Brazil

Originalmente codificado em ASL, o *Sign Writing* é capaz não apenas de converter os caracteres de nosso alfabeto latino, mas também utilizar um dicionário de sinais incorporado

em seu sistema. Com isso torna-se possível a construção de sentidos em um contexto utilizando sinais próprios de línguas de sinais naturais, não sinais criados para a transcrição e tradução por estes sistemas. Um dicionário da ASL codificado no sistema *Sign Writing* pode ser encontrado em http://www.signbank.org/dictionaries/pictdict/.

Um exemplo destacado é a transcrição de uma história infantil brasileira: *Uma Menina Chamada Kauana*, para o sistema *Sign Writing*.

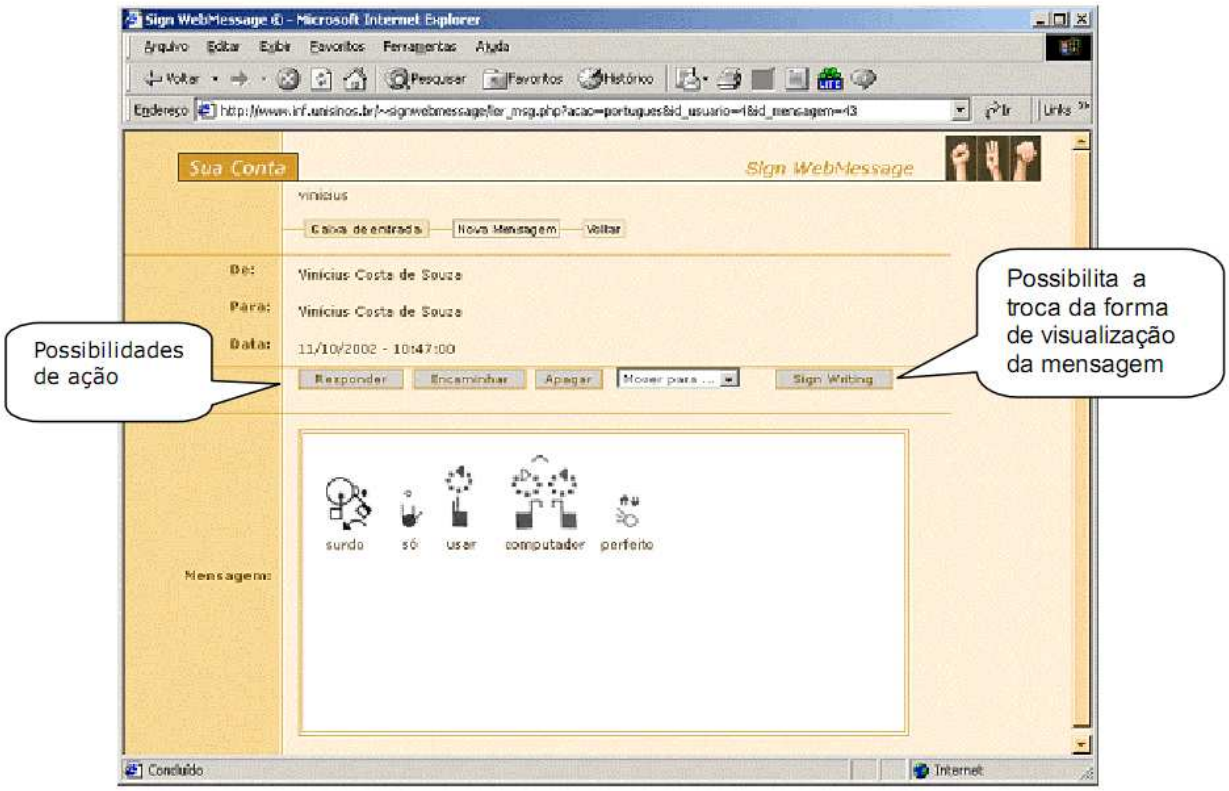
Pela utilização desta codificação, esse sistema é capaz de tornar a língua escrita em uma língua de sinais escrita. Assim, para que uma tradução seja realizada, o processo acontece com base em um corpus composto por sinais de uma determina língua de sinais e seus equivalentes em uma língua escrita.

Tradução e o Sistema *Sign Writing*

Com corpora compostos por sinais e devidamente codificados com informações lexicais, semânticas e morfológicas, nasce, porém ainda em pequeno porte, formas de tradução entre línguas verbais e línguas de sinais.

O surgimento deste processo se deu pela necessidade de comunicação entre sinalizantes surdos através das novas mídias, principalmente a Internet, pelo sistema de e-mails. Muitos

sistemas foram desenvolvidos com base na abordagem de escrita do *Sign Writing*, podemos citar, em Língua Brasileira de Sinais, o sistema chamado *Sign WebMessage*, que fora desenvolvido especialmente para a comunicação através de e-mails.



No sistema *Sign WebMessage*, a mensagem é visualizada em escrita *Sign Writing* de LIBRAS e logo seu significado em língua portuguesa. Em uma analogia, podemos dizer que este sistema é feito por meio da tradução direta, lexicográfica.

Outros sistemas que utilizam o *Sign Writing*, porém, de maneira mais aperfeiçoada, são os sistemas Sign Avatar. Estes sistemas são capazes de traduzir para língua de sinais, em ASL, textos e até mesmo diálogos falados, por meio de outro sistema de reconhecimento de voz (ver www.signavatar.com).

A IBM e a Língua de Sinais

Muito conhecida pelo desenvolvimento de diversos sistemas para processamento de língua natural, a IBM está engajada em um projeto específico para a tradução entre línguas verbais e línguas de sinais.

Chamado de *SISI* (*Say It, Sign It)*, este sistema é capaz de reconhecer a voz de um falante; decodificar a mensagem em modo texto e recodifica-la em língua de sinais; e através de um avatar, reproduzir esta língua de sinais em sinais representados por este sinalizante virtual animado. Um vídeo de testes realizados pela IBM como este programa pode ser acessado em http://www.youtube.com/v/RarMKnjqzZU&hl=en&fs=1.

Em um artigo apresentado pela *The Press Association*, esta tecnologia é capaz de tornar o modo de vida da comunidade surda mais fácil, pois poderá ser utilizado até mesmo em programas de televisão, os quais utilizarão este sinalizante automático. Além disso, sistemas de rádio em noticiários poderão ser apresentados em um monitor e distribuídos pela própria Internet, e até mesmo poderão ser utilizados para a tradução de e-mails escritos em línguas verbais para a sinalização em língua de sinais.

Acredita-se que com o desenvolvimento constante de ferramentas com o intuito de auxiliar no processo de comunicação, utilizando principalmente a tradução, o sistema de tradução de línguas de sinais em um curto período poderá ter ferramentas capazes de realizar todas as operações necessárias, porém os problemas clássicos da tradução automática ainda precisam ser mais explorados.

Programas de Sistemas de Tradução Automática

Nesta seção apresentamos os principais programas de tradução automática existentes no mercado. Dentre os programas de tradução automática destacam-se duas categorias, os programas baseados na Internet e os programas instaláveis. Destacaremos apenas os principais.

Programas baseados na Internet:

Yahoo Babel Fish – Disponível em http://babelfish.yahoo.com/, realiza tradução automática entre diversos pares linguísticos. Seu mecanismo de tradução é fornecido pelo *SYSTRAN*.

Google Translate – Disponível em http://translate.google.com/, realiza tradução automática entre 25 pares linguísticos, destaca-se o idioma Árabe, Russo e Chinês. Depois de muito tempo utilizando o sistema fornecido pelo *SYSTRAN*, atualmente o *Google Translate* utiliza seu próprio sistema de tradução automática. Utiliza-se de uma abordagem estatística.

Windows Live Translator – Disponível em http://www.windowslivetranslator.com/, também utiliza o sistema fornecido pelo *SYSTRAN*, portanto contem as mesmas características do *Yahoo Babel Fish*.

Programas Instaláveis:

SYSTRAN – Disponível em http://www.systransoft.com é o motor principal dos programas baseados na Internet, o *SYSTRAN* conta com um grandioso corpus devidamente preparado, ainda pode ser utilizado em conjunto com corpora específicos. Atualmente o SYSTRAN é declarado um dos melhores sistemas existentes para a tradução automática. O *SYSTRAN* conta atualmente com a tradução entre 52 pares linguísticos.

Power Translator – Disponível em http://www.lec.com é distribuído pela LEC (Language Engineering Company). O Power Translator é um sistema que compreende desde dicionários até tradutores totalmente automatizados, conta atualmente com a tradução para 21 línguas.

Pontos Principais

- Os sistemas de TA possuem um alto índice de produtividade quando comparados a tradutores humanos. Isso ocorre devido ao seu elevado grau de rapidez para a tradução de grandes volumes de textos;

- Quanto utilizado com um corpus específico para uma determinada área do conhecimento, o sistema de TA poder trazer benefícios além da produtividade, alcançando certo nível de qualidade, porém, não superando a tradução realizada por tradutores humanos;

- Com a pós-edição e revisão da tradução apresentada de forma automática, podem-se obter resultados significativos em um projeto de grande dimensão, pois o tradutor atuará também como revisor, assim problemas que ainda não foram solucionados nesses sistemas podem ser corrigidos manualmente;

- Não tirar conclusões precipitadas apenas utilizando sistemas de TA disponíveis gratuitamente na Internet, pois, para uma avaliação rigorosa, deve-se recorrer à aplicativos comerciais que possuem recursos completos para a realização da tarefa de tradução;

- O baixo nível de qualidade dos sistemas de tradução automática disponíveis gratuitamente na Internet, o que, consequentemente, faz com que tradutores iniciantes, como também clientes, sejam preconceituosos quanto aos benefícios que esses sistemas podem oferecer;

- Com o tratamento de corpora especializados e também a ampliação de ferramentas capazes de realizar análises linguísticas complexas, os sistemas de TA têm, a cada dia, evoluído consideravelmente, e logo poderemos contar com resultados mais eficientes e procurar pela qualidade da tradução desenvolvida pelos mesmos.

Referências

ARROJO, R. Oficina de Tradução: A teoria na prática. 3º edição. São Paulo: Editora Ática, 1997.

AUSTERMÜHL, F. Electronic Tools for Translators. Manchester: St. Jerome Publishing, 2001.

BAR-HILLEL, Y. Automatic Translation of Languages, 1960. Disponível em: http://www.mt-archive.info/Bar-Hillel-1960.pdf. Acessado em: 10/07/2008.

European Association for Machine Translation (EAMT). Disponível em: http://www.eamt.org/mt.html. Acessado em: 10/07/2008.

HUTCHINS, W.J. History of MT in a nutshell. A two-page sketch, from the beginnings to the present, 2001.

HUTCHINS, W.J; SOMERS, H.L. An introduction to machine translation. London: Academic Press. 1992.

MAKOTO, N. A framework of a mechanical translation between Japanese and English by analogy principle. In ELITHORN, A; Banerji, R. Artificial and Human Intelligence. Elsevier Science Publishers, 1984.

MATEUS, M. H. M. Tradução automática: um pouco de história. In Engenharia da Linguagem. Maria Helena M. Mateus e António Horta Branco (Orgs.). Lisboa, Edições Colibri, 1995, pp. 115-120.

Pierce, J. R; CARROLL, J.B; et al. Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

SANTOS, D. “Tradução automática”. Material de ensino na Escola de Verão da Linguateca, 2006. Disponível em: www.linguateca.pt/escolaverao2006/TA/TraducaoEscolaVerao.pdf. Acessado em: 10/07/2008.

SOUZA, Vinícius Costa. Sign WebMessage: um ambiente para comunicação via web baseado na escrita de Libras. Trabalho de conclusão - Unisinos. São Leopoldo, 2002.

VAUQUOIS, B. A survey of formal grammars and algorithms for recognition and transformation in machine translation, IFIP Congress-68 (Edinburgh), 1968, pp. 254-260.

UNIDADE III

Sistemas de Gerenciamento Terminológico

Esta unidade discute os sistemas de gerenciamento terminológico (GT) e está dividida em cinco partes principais: (i) um breve apanhado histórico dos sistemas de GT a partir do surgimento do estudo da terminologia, (ii) a definição do que são sistemas de GT, (iii) o funcionamento dos SGT através dos vários métodos utilizados; (iv) a utilização de sistemas de GT no âmbito da organização terminológica em línguas de sinais e, por fim, (v) alguns dos programas disponíveis no mercado.

Histórico

O congresso realizado em Copenhague em 1972 foi de grande importância para os Estudos da Tradução, por nele ser apresentado o texto considerado fundacional dos Estudos da Tradução, como já visto em Introdução aos Estudos da Tradução e Estudos da Tradução I, o artigo de James Holmes, The Name and Nature of the Translation Studies [O Nome e a Natureza dos Estudos da Tradução]. Mas, ainda neste congresso foi apresentado outro trabalho que tornou-se precursor dos estudos da terminologia e da teoria geral da terminologia, o artigo do engenheiro alemão Eugen Wüster, *Die Allgemeine Terminologielehre. Ein Grenzgebiet zwischen Sprachwissenchaft, Logik, Ontologie, Informatik und den Sachwissenschaften* [Teoria Geral da Terminologia. Um Limite entre o campo Linguístico, Lógico, da Ontologia, da Ciência da Informação e áreas afins].

Para a terminologia, este foi o marco inicial. Este trabalho realizado por Wüster (1972), portanto, deu inicio aos estudos relacionados à terminologia. Wüster em sua trajetória de estudos se concentrou em estabelecer uma padronização para gerenciamento de bancos terminológicos.

Inicialmente, o gerenciamento terminológico era realizado por meio de manuscritos devidamente estruturados no formato de enciclopédias (Austermühl, 2001). A utilização de sistemas mais complexos foi possível somente a partir do desenvolvimento das ferramentas computacionais. Iniciado pela utilização de programas para processamento de textos e, posteriormente, por programas específicos para realizar a coleta, o gerenciamento e busca nesses bancos terminológicos. Com a disseminação da Internet, vários programas baseados em sua interface foram desenvolvidos, tornando mais fácil a distribuição desses bancos terminológicos e, além disso, acessível a um grupo maior de usuários, devido ao fácil modo de circulação de informações.

A partir dos anos 80, com a criação dos primeiros dicionários eletrônicos, os SGT entraram em uma nova dimensão. Ao passar a utilizar sistemas tecnológicos, as funções que antes eram limitadas ao uso de escrituras em livros, enciclopédias e glossários específicos, passaram a utilizar sistemas automatizados para diversas tarefas, e um considerável aumento na quantidade de dados produzidos. A partir desse momento, os SGT passaram a ser constituídos por mecanismos complexos de compilação, estruturação e busca. Porém, devido ao sistema computacional ainda recente no início de suas atividades, os SGT possuíam limitações simples, como por exemplo, um banco terminológico poderia ser compilado em que a língua

fonte fosse a língua inglesa (EN), e a língua alvo a língua portuguesa (PT), onde as buscas poderiam ser realizadas na direção EN > PT, porém na direção inversa (PT > EN) não.

Os sistemas utilizados desde este período até metade da década de 90 eram exclusivos para grandes companhias como Termium e Eurodictautom. O fator principal para tal eram os custos extremamente elevados e a necessidade de uma boa infra-estrutura. No final da década de 90, com a disseminação do próprio computador para os usuários domésticos, vários programas criados para realizar gerenciamento de terminologia surgiram com o objetivo de entrar neste mercado e atingir pequenas organizações e tradutores que trabalhavam de forma autônoma. Surge assim o que conhecemos atualmente como Sistemas de Gerenciamento de Terminologia, totalmente eletrônicos, com grande eficiência e com seus custos reduzidos.

Definição, Vantagens e Desvantagens

Sistemas de Gerenciamento Terminológicos são ferramentas capazes de realizar a manipulação de bancos terminológicos e recursos linguísticos preparados para fins específicos (Christian Galinski & Gerhard Budin, 1993). Para isso faz-se necessário coletar e/ou extrair termos de um corpus, formar os bancos terminológicos, gerenciar esses bancos terminológicos e fornecer um mecanismo suficientemente capaz de realizar buscas e apresentar os dados, de forma estruturada, para o usuário final.

Para que possamos definir mais detalhadamente o que são os sistemas de GT precisamos primeiro definir o que entendemos por terminologia no âmbito de gerenciamento terminológico. Para POINTER (1997) em seu relatório intitulado “Proposals for an Operational Infrastructure for Terminology in Europe” [Propostas para uma Infraestrutura Operacional da Terminologia na Europa], terminologia (ou como chamado no plural: recursos terminológicos) é definida como: “um conjunto estruturado de conceitos e suas designações (símbolos gráficos, termos, unidades fraseológicas, etc.) em uma área específica” (http://www.computing.surrey.ac.uk/ai/pointer/report/section1.html#1). Logo, sistemas responsáveis por gerenciar esses termos, símbolos gráficos e etc. são denominados como Sistemas de Gerenciamento Terminológico.

No relatório realizado por POINTER (ibid.), duas definições complementam e reforçam a utilização da palavra ‘estruturado’ quando no uso de um conjunto estruturado para o manuseio de conceitos e designações. Para essas duas definições temos: (i) trabalho terminológico, ou seja, o trabalho realizado na criação ou documentação de recursos terminológicos e (ii) atividades terminológicas, isto é, um termo mais amplo que inclui além do trabalho terminológico, o ensino, o desenvolvimento de ferramentas e as medidas organizacionais e administrativas que envolvem o sistema terminológico.

Em uma análise realizada em SGT na Europa (Ahmad et al., 1995, p. 4), gerenciamento terminológico é definido por uma coleção de termos, os quais podem ser vistos simultaneamente como o gerenciamento de um artefato para informar sobre a natureza das línguas, o qual tem a função de promover a ciência, a artes, o comércio, o esporte nas áreas de desenvolvimento humano. Já Sistemas de Gerenciamento de Terminologia podem ser definidos como:

> Sistemas de Gerenciamento de Terminologia são a parte essencial da infra-estrutura da terminologia em que tais sistemas têm um aspecto funcional robusto, que é a produção e disseminação da terminologia, e

> têm um aspecto metodológico igualmente robusto, baseado, por um lado, em semântica e pragmática e, por outro lado, na filosofia da ciência e na ciência da biblioteca e busca de informações[1] (ibid., nossa tradução).

A designação de Sistemas de Gerenciamento de Terminologia se deu pela necessidade de organizar esses dados para uma busca mais precisa e apurada. Desta forma, gerenciamento desses bancos terminológicos é uma metodologia que usa sistemas de alta complexidade para a função da organização e estruturação desses dados. Esta metodologia empregada para a manipulação de terminologia é definida como:

> "Gerenciamento de Terminologia", por si só um neologismo, foi cunhado para enfatizar a necessidade de uma metodologia para coletar, validar, organizar, armazenar, atualizar, trocar e buscar termos individuais ou conjuntos de termos para uma dada disciplina. Esta metodologia é colocada em operação através do uso de sistemas de gerenciamento de informação por meio de computador chamados sistemas de gerenciamento de terminologia (SGT)[2] (ibid, p. 3, nossa tradução).

Em suma, estes são sistemas que tornam possível a manipulação de dados constituídos de um sistema lexical específico. Os sistemas de GT podem lidar tanto com terminologias monolíngues (por ex. constituído de termos somente em um língua, entrada e significados), bilíngues (constituídos de termos com a equivalência do mesmo em um devido par linguístico), como também multilíngues (constituídos de termos com equivalências em vários conjuntos linguísticos).

Dentre as vantagens que os sistemas de GT podem oferecer em relação a outras formas de se criar e gerenciar bancos terminológicos, Austermühl (2001) destaca as seguintes:

- são apropriados para rotinas específicas da tradução (e.g. procurando termos a partir de processadores de texto, importando informações do banco de dados para o processador de texto);
- concentram-se na funções relevantes à tradução;
- realizam pesquisas rápidas e flexíveis;
- fornecem uma comunicação automatizada entre bancos de dados e processadores de texto (p. 106).

Já com relação às desvantagens dos sistemas do GT, Austermühl (ibid.) aponta as seguintes:

- uso limitado (sem endereço ou gerenciamento de ordem);
- altos preços (p. 107).

É importante salientar que com os novos avanços dos sistemas de GT, essas desvantagens podem ser desconsideradas. Além disso, o fato que sistemas de gerenciamento de dados são

---

[1] Terminology management systems are an essential part of a terminology infrastructure in that such systems have a strong utilitarian aspect, that is production and dissemination of terminology, and have an equally strong methodological aspect, grounded in semantics and pragmatics on the one hand and on the other in philosophy of science and in library science and information retrieval.

[2] 'Terminology management', itself a neologism, was coined to emphasise the need for a methodology to collect, validate, organise, store, update, exchange and retrieve individual terms or sets of terms for a given discipline. This methodology is put into operation through the use of computer based information management systems called terminology management systems (TMS).

projetados para tradutores como o grupo alvo principal, suas vantagens sobre quaisquer outras formas de gerenciamento terminológico superam quaisquer eventuais desvantagens.

Funcionamento dos Sistemas de GT

O processo de funcionamento de um sistema de GT tem como base a estruturação de uma terminologia específica. Existem vários sistemas de GT disponíveis para utilização, apesar de cada um ter seu próprio sistema de funcionamento, as funções básicas se coincidem. Além da estruturação dos termos em um sistema de gerenciamento terminológico, funções como a extração terminológica e reconhecimento automático de termos, como também de buscas são os principais elementos que devem ser observados.

Para Bowker (2002), um Sistema de Gerenciamento de Terminologia tem como características fundamentais duas funções: o armazenamento e o processo de busca. Discutiremos a seguir, com base nessa autora, essas funções essenciais em um SGT.

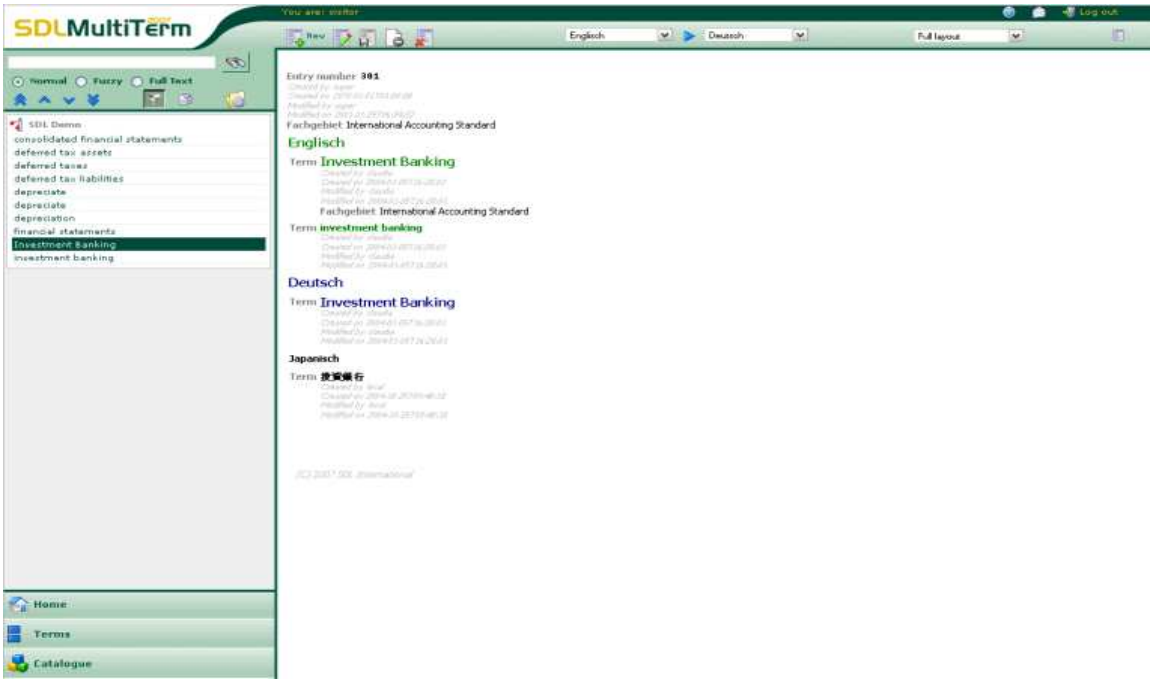
Armazenamento

De acordo com Bowker (ibid., 78), o armazenamento é uma das funções essenciais em um Sistema de Gerenciamento de Terminologia, pois é onde toda a informação ficará guardada e conterá também a meta informação sobre o banco terminológico. SGT mais antigos contavam com um sistema de armazenamento limitado em suas funções, sendo necessário, por exemplo, criar dois bancos terminológicos idênticos, apenas invertendo a posição de termo fonte e termo alvo para que pudesse ser utilizado em uma tradução em que exigiria mais do que uma direção (ex.: EN > PT, como PT > EN). Sistemas de GT atuais são capazes de armazenar bancos terminológicos de tal maneira que o usuário possa trabalhar em ambas as direções da tradução, tornando-os mais produtivos.

Além disso, com um sistema mais sofisticado, o volume de informações que podem constituir um banco terminológico aumentou, trazendo ganhos na estruturação do banco terminológico e proporcionando ao tradutor uma maior autonomia durante a atividade tradutória. Pela limitação encontrada nos sistemas de GT mais antigos, era apenas possível utilizar dois campos principais para a construção do banco terminológico, o termo na língua fonte e seu equivalente na língua alvo. Com a possibilidade desse aumento no volume de informações presentes na estrutura dos bancos terminológicos, a categorização mais detalhada dos termos constituintes pode ser realizada. Informações sobre contexto de utilização, campo disciplinar, definição, informações referenciais e, até mesmo, comentários podem ser adicionados aos bancos terminológicos. Figura 1 abaixo exemplifica está característica dos sistemas de GT mais recentes.

Com a possibilidade de adicionar informações extras sobre os termos em um banco terminológico, os sistemas de GT possibilitam ainda uma estruturação de dados livres para os termos que constituem tal banco terminológico. Certos termos que possuem uma grande quantidade de informações contextuais, referenciais e pertencentes a diversos campos disciplinares e outros, podem conter informações em uma estrutura livre, enquanto outros que não possuem esse grande volume de informações trazem apenas as informações básicas para a tradução dos mesmos.

Estes sistemas podem ainda utilizar-se de diferentes configurações visuais, para que os usuários identifiquem de forma mais fácil durante a pesquisa. Isso inclui, por exemplo, a utilização de cores para cada campo de informação, utilização de diferentes fontes e formatos. Outro recurso que se faz importante também, devido ao grande número de SGT disponíveis, é o formato de arquivo utilizado para a importação e exportação de dados entre os mesmos. O formato mais utilizado é o de texto simples (.txt), mas alguns sistemas utilizam formatos que utilizam uma estruturação mais avançada, como XML (eXtensible Markup Language) e CSV (Comma Separated Values). Esta troca de informações entre os diversos Sistemas de Gerenciamento de Terminologia proporciona uma maior interação entre os usuários desses sistemas e um aumento de produtividade para os mesmos. A estruturação de um banco terminológico pode ser visualizada na figura a seguir, SDL MultiTerm, executado em um plataforma de Internet.

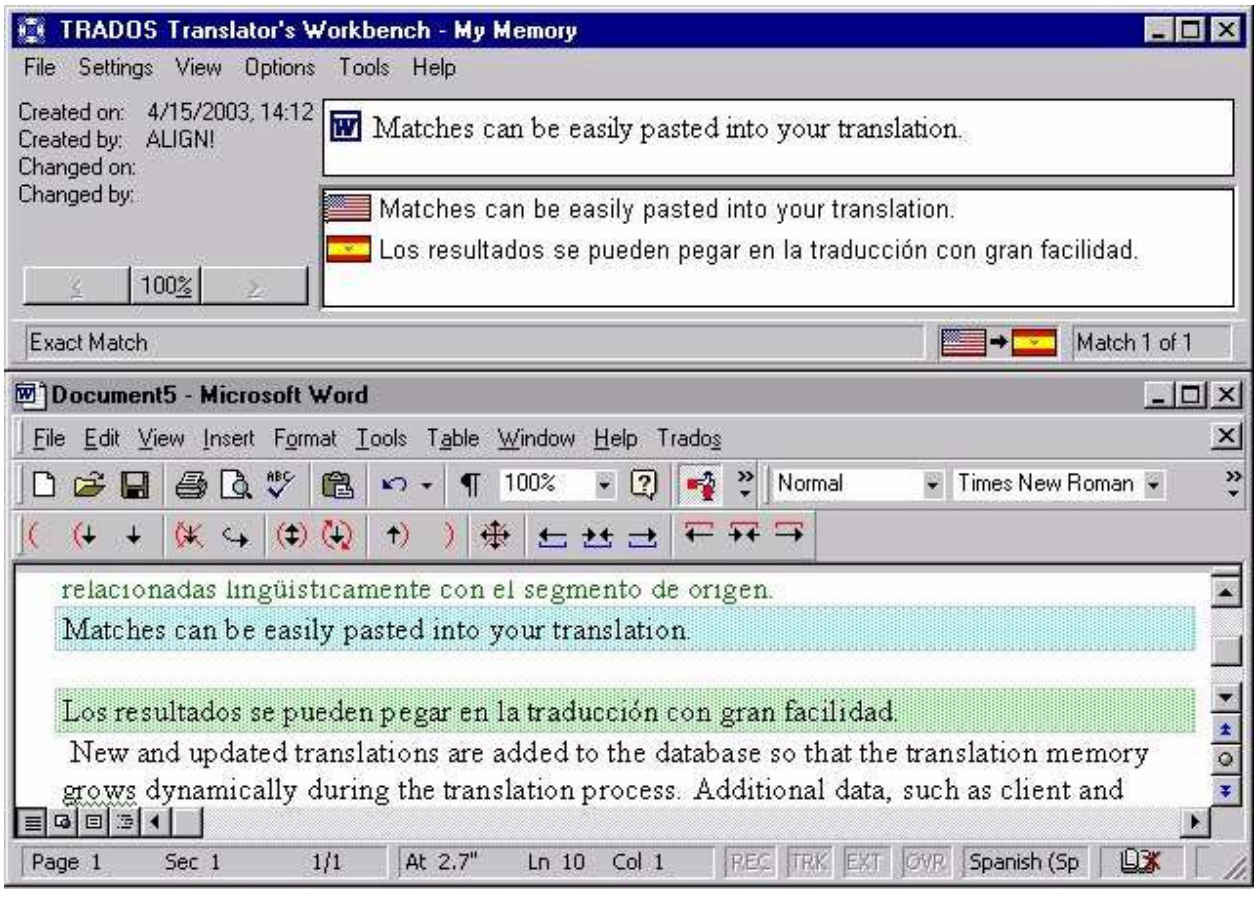


O Processo de Busca

O processo de buscas em um Sistema de Gerenciamento Terminológico é uma função essencial. Após ter feito a estruturação do banco terminológico e seu armazenamento, este processo tem como funcionalidade realizar buscas (pesquisas) no banco terminológico e apresentar ao usuário resultados eficientes desta busca. O método mais comum utilizado para buscas é a procura direta por termos, ou seja, uma “combinação exata” (exact match) para o termo solicitado. No entanto, com o progresso do sistema computacional, buscas utilizando caracteres especiais que possibilitam um método de busca mais livre podem ser incorporadas. Um exemplo para este método é a utilização do caractere coringa (*), com este tipo de operador pode se conseguir resultados mais amplos de um termo, como em sinal* = sinal, sinalizador, sinalizadores, sinalizante, sinalizantes, sinalizava, sinalizam, etc.

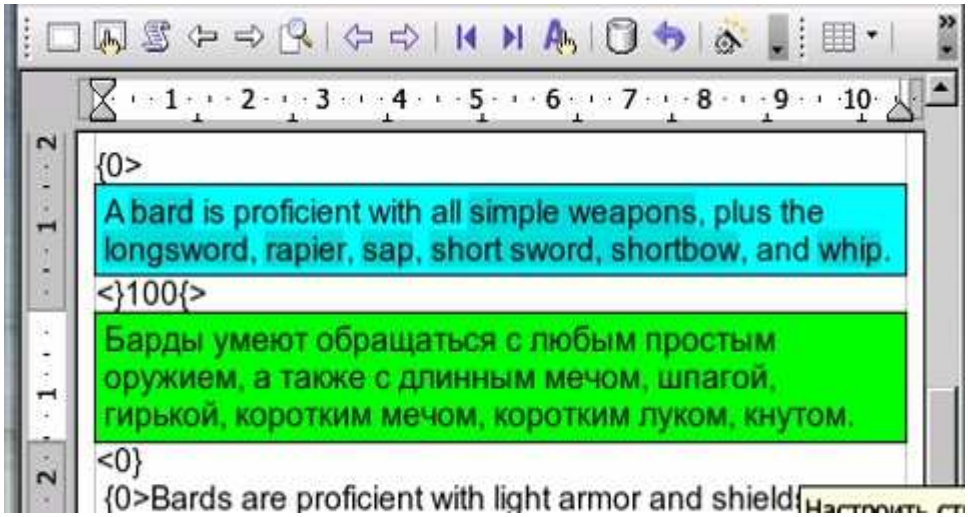
Outro método que tem sido bastante utilizado nos sistemas de GT atuais é o método de "combinação difusa" (Ver UNIDADE I, Seção **Funcionamento dos sistemas de MT)** que permite ao usuário realizar buscas a partir do termo solicitado e encontrar no banco de dados diversas combinações para as diferentes partículas do termo. Bowker aponta que, quando utilizado um caractere coringa ou um "combinação difusa" para realizar uma busca, podem ser encontrados diversos equivalentes potenciais para um mesmo termo (para equivalentes potenciais, ver Krings (1986), em Estudos da Tradução I).

A utilização de Sistema de Gerenciamento de Terminologia integrado como Sistemas de Memória de Tradução tem se tornado útil para a tradução, pois, assim, estes dois sistemas podem operar juntos para garantir rendimento e qualidade maior em uma atividade tradutória. Diversos sistemas, como por exemplo, Trados, Déjà Vu possuem uma integração com um sistema externo de gerenciamento terminológico. A integração dos SMT Trados e o SGT MultiTerm pode ser visto na figura a seguir.

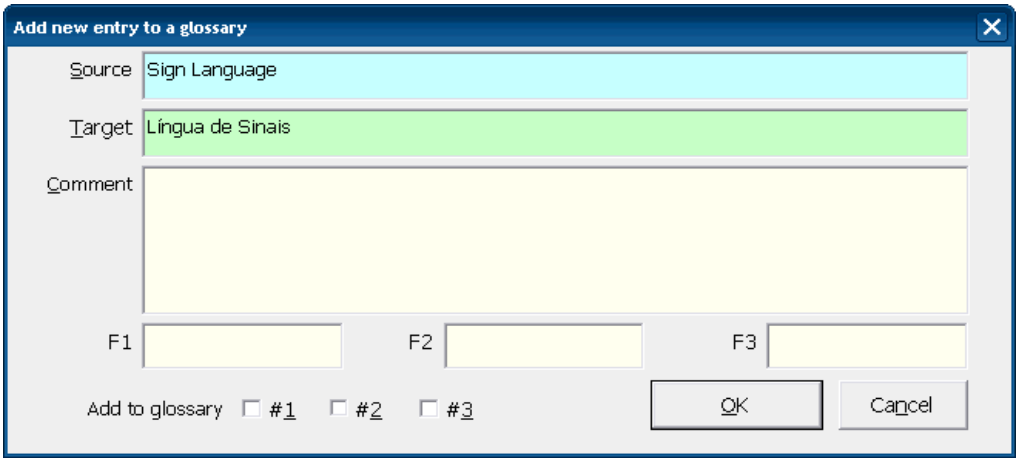


Outros como Wordfast, Heartsome e Anaphraseus possuem sistema de gerenciamento terminológico baseado em seus glossários internos. Estes sistema atuando junto na atividade tradutória agilizam as buscas e, com uma função crucial, operam no reconhecimento automático de terminologia em tempo real. Isso possibilita ao tradutor uma uniformização terminológica durante sua tradução e garante um nível maior na qualidade final de seu

produto. Um exemplo de reconhecimento automático de terminologia utilizando o Anaphraseus pode ser visto na figura abaixo, nos termos realçados.



Na figura seguinte podemos visualizar a entrada de um termo no glossário ativo do Sistema de Memória de Tradução que, como citado acima, também possui um sistema de gerenciamento ativo de terminologia.



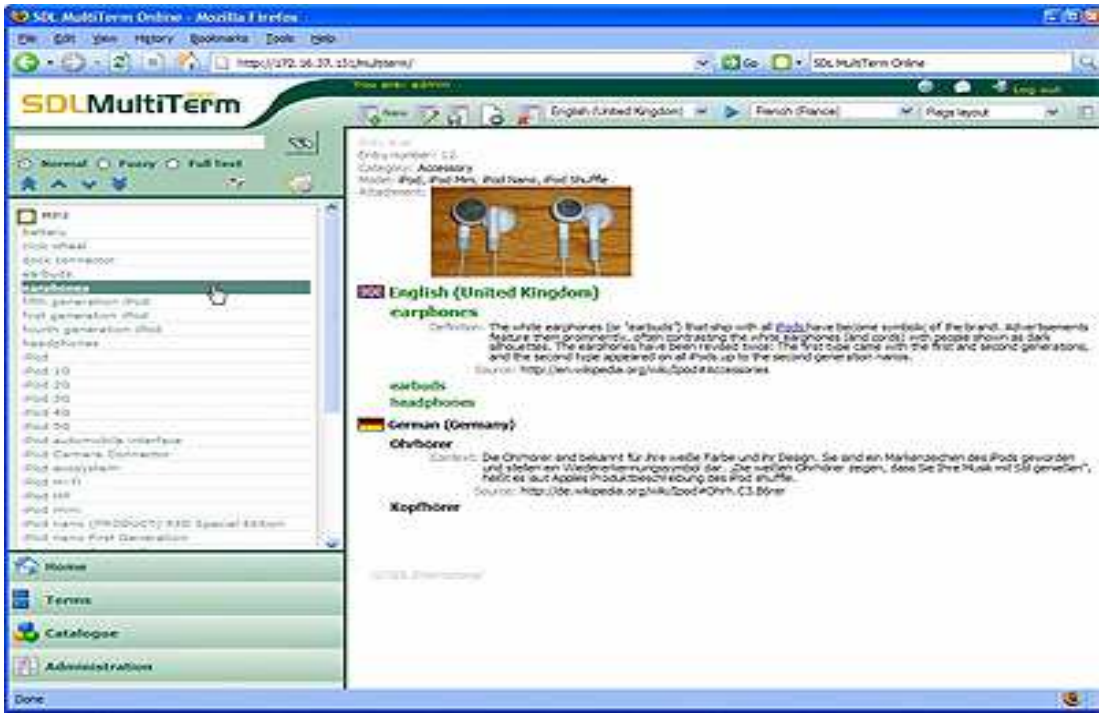
Outras funções como, por exemplo, a extração automática de termos também constitui sistemas mais avançado de gerenciamento terminológico. Com este método de extração automática podem ser criados bancos terminológicos de maneira automatizada, feito por meio da abordagem estatística e considerando agrupamentos de palavras que possuem frequências altas em um determinado corpus e tornam-se candidatos a termos.

Em Frank Austermühl (2001), podemos encontrar diferentes maneiras de construir e gerenciar bancos terminológicos, tanto com a utilização de sistemas avançados, específicos para esta finalidade, como utilizando planilhas e documentos de texto.

Sistemas de Gerenciamento Terminológico e Línguas de Sinais

Muitos bancos terminológicos estão disponíveis na Internet. Muitos deles possuem estruturas complexas, contento informações adicionais como contexto, imagens e até mesmo vídeos, o que é de grande auxílio para o tradutor/intérprete de língua de sinais. Alguns exemplos podem ser encontrados nos seguintes endereços: http://www.acessobrasil.org.br/libras/, http://commtechlab.msu.edu/sites/aslweb/browser.htm, http://wings.avkids.com/Book/Signing/, http://www.aslpro.com/.

Para a construção de um banco terminológico próprio, usuários que utilizam línguas de sinais podem utilizar sistemas mais completos como o MultiTerm, que possibilita uma estrutura de dados livre e conta com suporte a utilização de imagens, como pode ser visualizado na figura a seguir.



Com este tipo de recurso é possível construir e gerenciar bancos terminológicos com autonomia, principalmente pelo seu processo de busca, que aumenta a qualidade e produtividade da tradução e, até mesmo, da interpretação, pois pode ser acessado antes de se realizar a interpretação para, então, solucionar problemas relacionados à terminologia de cada área específica e também sanar dúvidas sobre a utilização de termos mais adequados para tal situação.

A utilização de Sistemas de Gerenciamento de Terminologia em conjunto com Sistemas de Memória de Tradução no ambiente de línguas de sinais é praticamente inexistente, devido ao fator que as línguas de sinais estão tendo sua entrada neste processo, e o desenvolvimento da língua de sinais escrita depende ainda de outros programas que sejam capazes de tratar de caracteres especiais para trabalhar em um ambiente totalmente eletrônico.

Programas disponíveis

Atualmente, uma lista enorme de Sistemas de Gerenciamento Terminológico está disponível no mercado. A características de cada um diferem apenas em funções adicionais específicas, mas, na base funcional, todos possuem características semelhantes e essenciais.

MultiTerm – Disponível em: http://www.sdl.com/en/products/terminology-management/MultiTerm.asp, é um programa desenvolvido por uma das maiores empresas que atua no mercado de tradução e localização, a SDL, também responsável pelo Trados. Um dos programais mais utilizados, devidos a sua interface intuitiva e ao alto grau de complexidade em suas funções, possibilitando a criação de bancos de dados profissionais e com altíssima qualidade. Possui ainda uma versão de demonstração onde é possível utilizá-la em uma versão baseada na Internet disponível em: http://www.MultiTerm.com/MultiTerm/.

Déjà Vu – Disponível em: http://www.atril.com/, um programa desenvolvido por uma das empresas pioneiras no mercado de Sistemas de Memória de Tradução. Possui funções semelhantes aos outros SGT, porém somente é possível utilizá-lo quando instalado em uma estação de trabalho.

TermBases – Disponível em: http://www.termbases.eu, é um sistema desenvolvido para a plataforma Internet. Com o TermBases é possível criar e gerenciar bancos terminológicos totalmente on-line, o que melhora no compartilhamento e na distribuição dos bancos terminológicos entre os usuários.

Dentre uma lista imensa de SGT que pode ser encontrada, para cada projeto uma especificidade de funções são necessárias, neste caso devem ser avaliadas todas as características para a escolha de um SGT que seja adequado para um projeto de tradução.

Pontos Principais

- Programas de gerenciamento terminológico existem de uma forma ou de outra desde a década de 60, mas programas atuais caracterizam-se por uma série de melhoramentos, inclusive armazenamento e opções de recuperação mais eficazes e flexíveis;
- Eles também podem armazenar mais informações e possuem uma estrutura de verbetes livres que permite o usuário definir e formatar seus próprios campos de dados;
- Características de recuperação incluem caracteres coringa, combinação difusa, reconhecimento ativo de terminologia, pré-tradução e extração de termos;
- Incorporado em diversos sistemas de TA, os sistemas de GT trazem uma série de benefícios quando trabalhando com reconhecimento ativo de terminologia durante a tarefa tradutória, e também pode auxiliar no processo de revisão, quando realizado por um terceiro, que não o próprio tradutor;
- O tempo necessário para se construir um banco terminológico criterioso é bastante grande, o que, em alguns casos, resulta na construção de bancos terminológicos com estruturas mais simples para a manipulação dos termos;

- A reutilização (reciclagem) dos bancos terminológicos para diversos trabalhos futuros, quando inseridos em uma mesma área de conhecimento, podem ser atualizados e melhorados a cada novo projeto de tradução, aumentando a qualidade e, consequentemente, a produtividade por parte do tradutor.

Referências

AUSTERMÜHL, Frank. Electronic Tools for Translators. Manchester: St. Jerome Publishing, 2001.

BOWKER, Lynne. Computer-Aided Translation Technology. A practical introduction. Ottawa: University of Ottawa Press, 2002.

K. Ahmad, W. Martin, M. Hoelter, M. Rogers. Aspects of Terminology Infrastructure in Europe: Volume 3 - Specialist Terms in General Language Dictionaries. POINTER Report (Available as a University of Surrey report), 1995.

KRINGS, Hans P. 'Translation problems, and Translation Strategies of Advanced German Learners of French (L2)'. IN: Juliane House and Shoshana Blum-Kulka (Orgs.). Interlingual and Intercultural Communication – Discourse and Cognition in Translation and Second Language Acquisition Studies. Tübingen: gnv- Gunter Narr Verlag, 1986.

POINTER - Proposals for an operational Infrastructure for terminology in Europe. Disponível em: http://www.computing.surry.ac.uk/ai/pointer/report/intro.html, 1996.

Wright, Sue Ellen & Leland D.Wright. 'Terminology Management for Technical Translation'. in Sue Ellen Wright & Gerhard Budin, The Handbook of Terminology Management, Vol.1. Ps. 147-159. Amsterdam & Philadelphia: John Benjamins Publishing Company, 1997.

Wüster, Eugen. Die allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften (The Theory of General Terminology – An Intersection of Linguistics, Logic, Ontology, Information Science, and the Technical Sciences), 1972.

UNIDADE IV

Corpora Eletrônicos e Tradução

Esta unidade explora a utilização de corpora eletrônicos como ferramentas de auxílio aos tradutores. A unidade está dividida em cinco partes principais: (i) histórico – onde apresentamos uma breve contextualização histórica dos estudos da tradução em corpora (ii) definição – onde oferecemos uma definição de trabalho do que vem a ser um corpus eletrônico para o estudo da tradução; (iii) tipos – onde rapidamente sugerimos uma tipologia para a classificação de corpora baseada em Baker (1995); (iv) compilação – onde brevemente delineamos o processo básico de criação de um corpus eletrônico; e (v) aplicabilidade – onde descrevemos alguns corpora online utilizados no ensino de algumas línguas de sinais.

Histórico

Segundo Sara LAVIOSA, em um a palestra proferida na edição 2003 da *International Corpus-Based Conference* [Conferência Internacional Baseada em Corpus] em Pretoria, África do Sul, o alvorecer dos *Estudos da Tradução em Corpora* (ETC) ocorreu entre os anos de 1993 e 1995. Foi nesse período que Professor Mona Baker da Universidade de Manchester publicou dois artigos que seriam considerados os textos seminais deste "novo paradigma" de pesquisa em tradução.

No primeiro artigo publicado em 1993, "Corpus Linguistics and Translation Studies: Implications and Applications" [Linguística de Corpus e Estudos da Tradução: Implicações e Aplicações], Baker argumenta que "a disponibilidade de grandes corpora tanto de textos originais e textos traduzidos, juntamente com o desenvolvimento de uma metodologia baseada em corpus, permitirá aos acadêmicos de tradução revelar a natureza do texto traduzido enquanto um evento comunicativo mediado[3]" (p. 243, nossa tradução).

Já em seu artigo de 1995, "Corpora in Translation Studies: An overview and some suggestions for future research" [Corpora em Estudos da Tradução: panorama e algumas sugestões para pesquisa futura], Baker, além de oferecer uma tipologia de corpora para a tradução, delineia alguns dos pontos a serem considerados ao se criar uma metodologia baseada em corpus para o estudo e o ensino de tradução.

Segundo KENNY (1998), o trabalho de Baker (1993 e 1995) foi instrumental não somente ao incorporar os métodos e ferramentas da Linguística de Corpus aos Estudos Descritivos da

---

[3] "The availability of large corpora of both original and translated text, together with the development of a corpus based methodology will enable translation scholars to uncover the nature of translated text as a mediated communicative event" (BAKER, 1993, p. 243).

Tradução, mas também por destacar os desafios específicos que a tradução apresenta para os estudos e aplicação prática de corpora (p. 50).

Definição

Tradicionalmente, a palavra corpus (plural corpora) significa um "corpo" ou coleção de escritos, textos, material oral, etc. (COD, 1995). Entretanto no contexto dos Estudos da Tradução, a definição de corpus possui conotações mais específicas: formato digital, textos completos, auto-configuráveis e representativos.

Formato Digital – hoje em dia, para que os textos possam ser armazenados e processados por ferramentas computacionais, é necessário que eles estejam em formato digital. Textos digitalizados permitem uma maior manipulação dos dados e, consequentemente, permitem a investigação de fenômenos que antes permaneciam indetectáveis a olho nu.

Textos Completos – no passado conjuntos de frases, sentenças e excertos tirados de um texto eram considerados suficientes para se construir um corpus. Atualmente, entretanto, opta-se em criar corpora com textos completos para que se possa também levar em consideração outros níveis linguísticos de significação que vão além do nível de sentença tais como, coesão e coerência.

Auto-Configuráveis – um corpus deve ser construído de tal forma que o usuário possa re-configurar os textos do corpus de acordo com suas necessidades. Isso faz com que um corpus possa ser utilizado de várias formas e por vários usuários, o que por sua vez faz do corpus uma ferramenta inesgotável de dados e aplicações práticas.

Representativos – os textos incluídos em um corpus devem tentar representar o máximo possível o fenômeno tradutório a ser investigado ou as aplicações práticas a serem utilizadas através do corpus em questão. A representatividade como um conceito estatístico é difícil de ser atingida, mas cabe ao criador do corpus tentar diminuir o máximo possível quaisquer distorções.

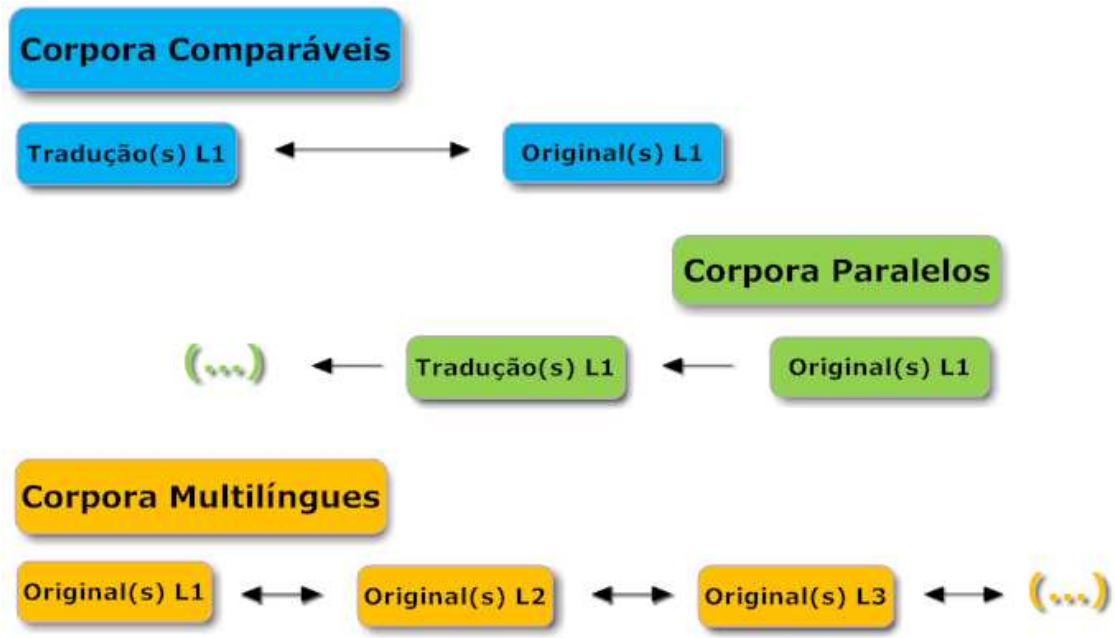
Neste sentido, um corpus pode ser interpretado como **uma coleção auto-configurável de textos completos digitalizados, analisáveis automática ou semi-automaticamente e, coletados a fim de serem representativos ao máximo do fenômeno tradutório sendo examinado** (c.f. Baker, 1995).

Tipos de Corpora

Faz mais de uma década que Baker (1995) propôs sua renomada tipologia de corpora para pesquisa e ensino de tradução. Ao discuti-la, a autora sugere três tipos principais de corpora "em antecipação ao surgimento da atividade[4]" (p. 230, nossa tradução) nesta área específica, a

[4] "in anticipation of the surge of activity" (p. 230)

saber, corpus comparável, corpus multilíngue e corpus paralelo. Esses três tipos de corpora são ilustrados pela figura abaixo.



Corpora Comparáveis – "consistem em duas compilações separadas de textos na mesma língua: um corpus consiste de textos originais na língua em questão e o outro consiste de traduções naquela língua a partir de uma dada língua fonte ou línguas[5]" (p. 244, nossa tradução);

Corpora Paralelos – consistem de "textos originais da língua fonte A e suas versões traduzidas na língua B[6]" (p. 230, nossa tradução);

Corpora Multilíngues – são "conjuntos de dois ou mais corpora monolíngues em línguas diferentes, construídos ou pelas mesmas, ou diferentes instituições, tendo como base critérios de desenho semelhantes[7]" (p. 232, nossa tradução).

Segundo FERNANDES (2006), a classificação tripartite de Baker (1995) pode ser reformulada utilizando-se apenas duas categorias principais: comparável e paralelo. O autor argumenta que o termo multilíngue não possui nenhuma característica contrastiva que possa distingui-lo dos outros dois tipos principais de corpora. Além disso, essa classificação parece não ter sido muito utilizada na área, já que o termo corpus comparável multilíngue tem

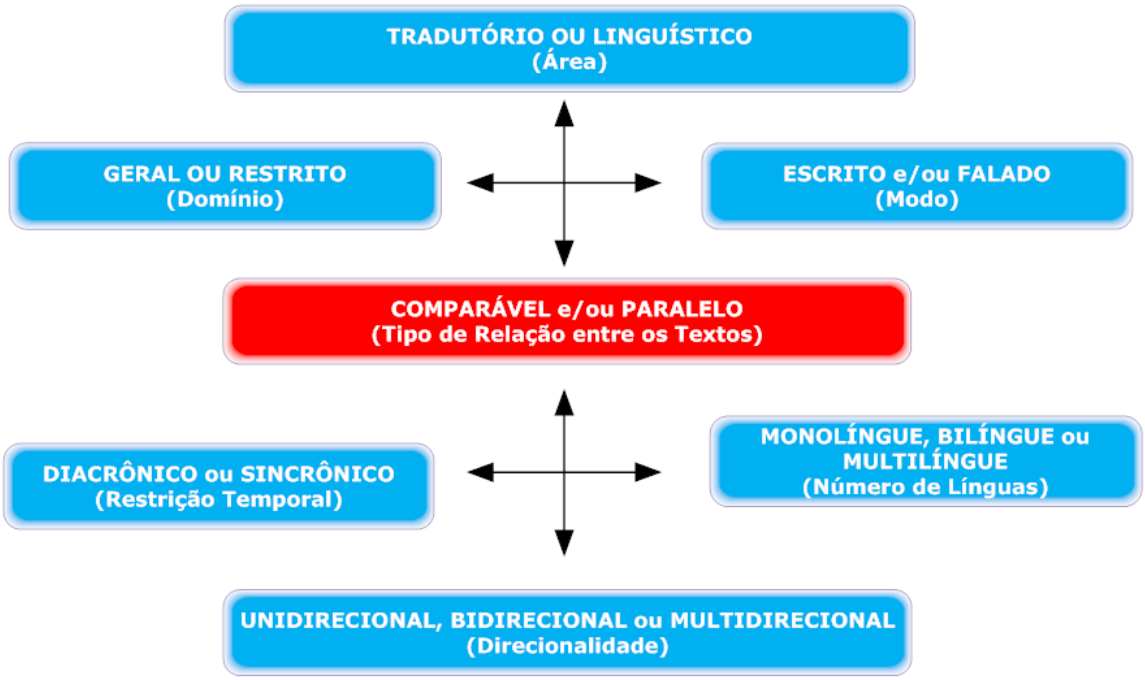
---

[5] "consist of two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages" (p. 234).

[6] "original, source language-texts in language A and their translated versions in language B" (p. 230).

[7] "sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria" (p. 232).

sempre sido utilizado como substituto do termo corpus multilíngue (ver Teubert, 1996 e Kenny, 2001). Vale a pena observar também que no primeiro livro introdutório sobre ETC, intitulado *Introducing Corpora in Translation Studies* [Introduzindo Corpora aos Estudos da Tradução] por Maeve OLOHAN (2004), o autor centra a atenção nos corpora comparável e paralelo, o que, por sua vez, pode indicar uma leve mudança de perspectiva na maneira com que os tipos de corpora são classificados. A figura a seguir mostra a classificação de corpora proposta por FERNANDES (2006), no que se referem os ET.



Tipo de Relação entre os Textos: Comparável ou Paralelo?

Para FERNANDES (2006), seria muito mais vantajoso centrar a atenção nos termos 'comparável' e 'paralelo' a partir da perspectiva de suas características contrastivas. Estas características têm a ver com o tipo de relação que existe entre os textos que integram o corpus (cf. TEUBERT, 1996). Em um corpus comparável, por exemplo, os textos são selecionados tendo como base uma relação textual entre eles (i.e. os textos são selecionados de acordo com suas semelhanças em termos de assunto, tipo de texto, função comunicativa, etc.). Em um corpus paralelo, por outro lado, os textos são agrupados tendo como base uma relação tradutória (i.e. os textos são selecionados de acordo com algum tipo de relação de tradução entre eles).

Se voltarmos à classificação tripartite de Baker (1995) (ver acima), é possível observar que o termo 'multilíngue' não possui nenhuma característica contrastiva que o faça diferente dos outros dois tipos de corpora. O termo multilíngue parece adquirir uma característica contrastiva somente quando comparado a outros corpora em termos de número de línguas (ver abaixo). Neste sentido, o que Baker (ibid.) apresenta como corpus multilíngue poderia ser classificado, segundo esta nova perspectiva, como um corpus comparável linguístico multilíngue. Linguístico, pois este tipo de corpora não está fundamentalmente preocupado com o estudo da tradução (ver abaixo); multilíngue, devido ao número de línguas envolvidas, e comparável

porque os textos integrantes deste tipo de corpora são reunidos tendo como base a semelhança textual entre eles.

Área: Linguístico ou Tradutório?

Este segundo critério, proposto por FERNANDES (2006), está relacionado à distinção entre estudos baseados em corpora desenhados para o estudo de línguas e aqueles construídos com vistas a investigar produtos e processos da tradução. O autor, então, sugere os termos "linguístico" e "tradutório, respectivamente, para distinguir esses dois tipos de corpora. Apesar dos ETC estar basicamente preocupado com corpora tradutórios, muitos estudiosos interessados na educação do tradutor (ver Schäffner, 1998; Zanettin, 1998; Stewart, 2000; e mais recentemente, Zanettin, Bernardini and Stewart, 2003) também utilizam corpora linguísticos como ferramentas para melhorar e desenvolver a competência linguística e tradutória de tradutores em formação.

Domínio: Geral ou Restrito?

O termo domínio se refere à área de pesquisa linguística sob a qual o corpus centra sua atenção. Com relação ao domínio, existem basicamente dois tipos de corpora: geral e restrito (Baker, 1995, p. 229). Como o próprio nome sugere, um corpus de domínio geral possui um escopo mais amplo, por ser construído para estudar a linguagem do material traduzido como um todo. Por outro lado, um corpus de domínio restrito investiga a tradução da linguagem de gêneros e tipos de textos específicos.

Modo: Escrito e/ou Falado?

Modo tem a ver com a maneira que os conteúdos originais de um texto são apresentados. Por exemplo, um texto transcrito a partir de um fonte de vídeo ou áudio é considerado "falado" e um texto escaneado a partir de um livro e convertido em formato eletrônico é considerado "escrito". Segundo Atkins et al. (1992), quando o modo de apresentação não for especificado, ele será "escrito" por padrão.

Restrição Temporal: Diacrônico ou Sincrônico?

Com relação às restrições de tempo, um corpus pode ser categorizado como sincrônico – quando ele centra sua atenção em um objeto de estudo em um período temporal específico, ou diacrônico – quando ele se preocupa com o desenvolvimento histórico deste objeto de estudo através do tempo (Atkins et al., 1992, p. 6).

Numero de Línguas: Monolíngue, Bilíngue, Trilíngue ou Multilíngue?

No que diz respeito ao número de línguas, um corpus pode ser classificado como monolíngue, bilíngue, trilíngue ou multilíngue quando mais de três línguas estão envolvidas. Outro aspecto relacionado ao número de línguas sendo representado no corpus tem a ver com as variedades linguísticas de uma mesma língua. Se um corpus, por exemplo, for descrito como bilíngue e o par linguístico envolvido for português e inglês, é importante especificar a variedade linguística

sendo coberta pelo corpus (e.g. português europeu X português brasileiro ou inglês britânico X inglês americano).

Direcionalidade: Unidirecional, Bidirecional ou Multidirecional?

Zanettin (2000) considera a direcionalidade a direção tradutória dos textos que integram o corpus. Por exemplo, em um corpus formado por textos originalmente escritos em L1 e suas respectivas traduções em L2, a direção das traduções ocorre em apenas uma direção, portanto, é classificado como unidirecional. Agora, se um corpus é formado de textos originalmente escritos em L1 e suas traduções em L2, mais originais em L2 e suas respectivas traduções em L1, temos um corpus bidirecional. Corpora multidirecionais são também possíveis, principalmente, quando mais de duas línguas estão envolvidas e a direção das traduções não está centrada na L1, mas na interação de todas as línguas que integram o corpus (p. 106).

Uma última questão que merece ser discutida tem a ver com a combinação de corpora, dependendo dos objetivos da pesquisa, um corpus pode ser combinado com outros corpora a fim de atingir aqueles objetivos específicos. Os Usuários do *Translation English Corpus* (TEC), por exemplo, têm que contar com o *British National Corpus* (BNC) para terem seu corpus comparável, o que aponta ao fato de que uma maior uniformização em relação a codificação dos textos se faz necessária para que mais e mais corpora possam ser combinados e seus usos divulgados por todo o globo.

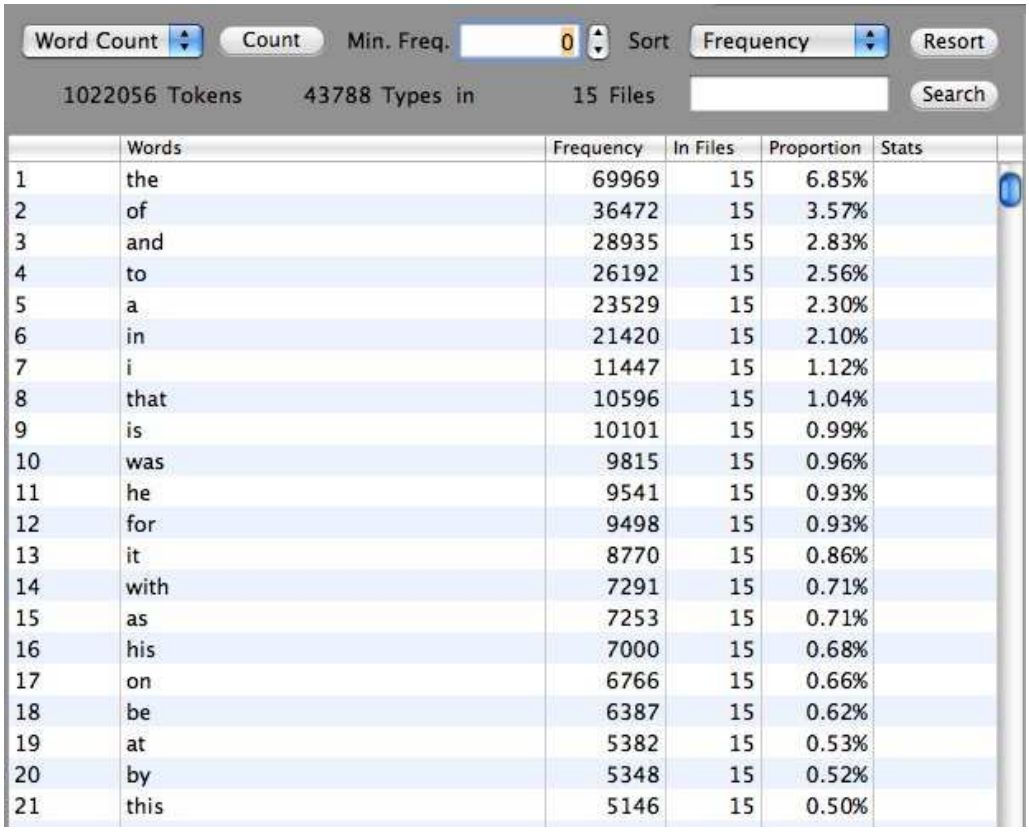
Técnicas de Processamento de um Corpus

De acordo com Kenny (2001), um corpus *per se* tem pouca utilização prática se não houver técnicas para pesquisar, classificar e catalogar uma grande quantidade de dados que possam por ele ser fornecido (p. 33). Nesta seção, centraremos nossa atenção em técnicas que podem ser utilizadas com textos em seus estados naturais (isto é, textos sem/ou com apenas marcação mínima) e em análises lexicais, mostrando como estas técnicas podem ser utilizadas para manipular dados em um corpus paralelo bilíngue. As técnicas básicas adotadas no presente estudo são: listas de palavras e concordâncias, porém serão discutidas também questões relacionadas a palavras-chave.

Listas de Palavras

A técnica mais básica para exibir informações sobre os elementos linguísticos em um corpus é gerada por meio de listas de palavras (Kennedy, 1998, p. 244). As listas de palavras permitem ao pesquisador obter informações estatísticas sobre o número de tipos (palavras diferentes) e ocorrências (número total de palavras) para textos individuais em um corpus, como também para um corpus por completo. A relação entre os tipos de palavras para com as suas ocorrências no corpus, neste caso, exibe a amplitude e diversidade de vocabulário utilizada por um escritor ou um tradutor representado naquele corpus. Para Baker (2000), uma relação alta de tipos/ocorrências pode significar que um escritor/tradutor utiliza um conjunto amplo de palavras, enquanto uma relação baixa de tipos/ocorrências pode significar que um escritor/tradutor utiliza um conjunto de palavras mais restrito (p. 250). Segue abaixo um

exemplo de Lista de Palavras extraída do programa CasualConc, disponível no sítio http://casualconc.googlepages.com/gettingstarted-wordcount .



	Words	Frequency	In Files	Proportion	Stats
1	the	69969	15	6.85%	
2	of	36472	15	3.57%	
3	and	28935	15	2.83%	
4	to	26192	15	2.56%	
5	a	23529	15	2.30%	
6	in	21420	15	2.10%	
7	i	11447	15	1.12%	
8	that	10596	15	1.04%	
9	is	10101	15	0.99%	
10	was	9815	15	0.96%	
11	he	9541	15	0.93%	
12	for	9498	15	0.93%	
13	it	8770	15	0.86%	
14	with	7291	15	0.71%	
15	as	7253	15	0.71%	
16	his	7000	15	0.68%	
17	on	6766	15	0.66%	
18	be	6387	15	0.62%	
19	at	5382	15	0.53%	
20	by	5348	15	0.52%	
21	this	5146	15	0.50%	

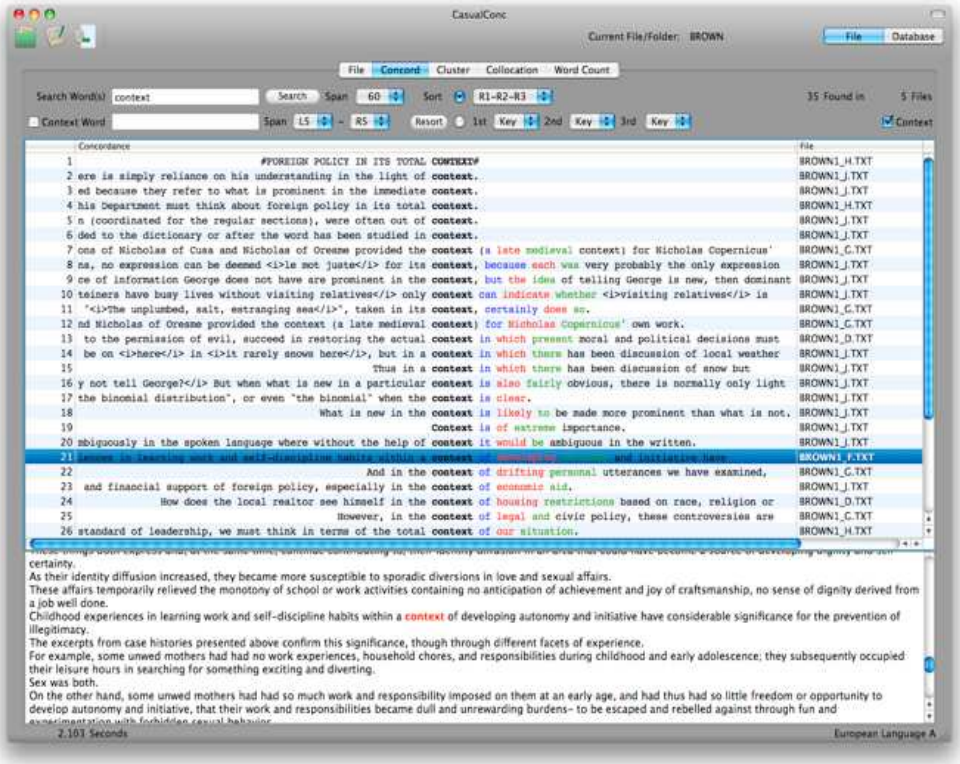
Outras técnicas quantitativas fornecidas por meio de listas de palavras incluem calcular a média de palavras e a extensão de sentenças; número de parágrafos e suas respectivas extensões; e, também, a quantidade de espaço necessária para o armazenamento de um caractere utilizado pelo corpus, como um todo e, também, individualmente, cada um dos textos. A vantagem dessas técnicas é que elas permitem ao analista humano obter uma visão geral quantitativa da maneira que os textos em um dado corpus estão estruturados, em termos de informações estatísticas que tais técnicas proporcionam (isto é, número de tipos, ocorrências, parágrafos e etc.).

Entretanto, é válido citar que as técnicas quantitativas descritas acima não são tão simples e, tampouco, livres de problemas. Há certo número de problemas práticos que os analistas humanos devem levar em consideração. No caso da relação entre tipos/ocorrências, por exemplo, Kenny (2001) aponta que este tipo de relação é extremamente sensitivo para a extensão do texto, o mais provável é que palavras gramaticais irão estar naquele texto, resultando assim em uma relação baixa de tipos/ocorrências (p. 34). E, a fim de superar esse problema, as relações entre tipos/ocorrências estão, normalmente, padronizadas para permitir comparações entre textos de diferentes extensões. Esta padronização é normalmente obtida pelo cálculo da relação entre partes do texto corrido (diz 1.000 ocorrências), e então feita uma média no final tudo isso.

É importante citar que não é nossa intenção discutir os problemas relacionados para todas as técnicas utilizadas em listas de palavras, visto que já tenham sido feitas em outros estudos (para uma revisão desses problemas, ver Kenny, 2001, pp. 34-35). O que parece importante enfatizar é que, para realizar essa contagem, o computador tem como base palavras ortográficas, e, portanto, não contam com a desambiguação semântica de palavras homográficas.

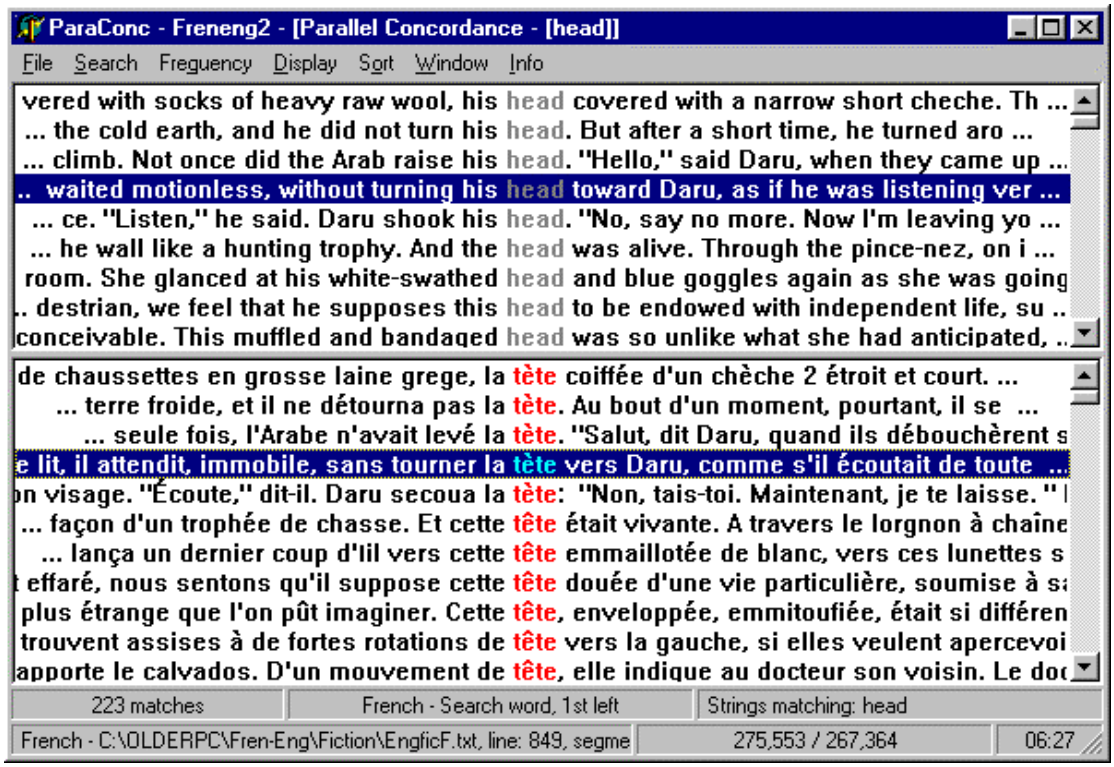
Concordâncias

A técnica mais básica para o processamento de concordâncias é a listagem de todas as ocorrências (tokens) de um tipo (type) específico em um corpus. De acordo com Kennedy (1998), um tipo é geralmente denominado uma palavra-chave, porém às vezes pode se referir a um item de pesquisa/alvo ou, mais comumente, como uma palavra denominada “nódulo” (p. 251). O formato mais comum para concordâncias é o de Palavras-Chave no Contexto (KWIC – Key Word in Context), onde o programa produz uma lista de exemplos de um nódulo, exibindo o contexto no qual este nódulo está se encontra. A figura abaixo mostra uma tela de concordância monolíngue do programa CasualConc com o nódulo “context” (Disponível em http://casualconc.googlepages.com/ConcResult.png **)**.



Algumas ferramentas de concordância disponíveis no mercado podem até mesmo oferecer pesquisas mais flexíveis, por permitir a utilização de caracteres coringa. Os caracteres coringa são caracteres que podem preencher o lugar de outros caracteres. O caractere asterisco (*) representa um caractere coringa, significando um caractere que toma o lugar do caractere zero ou de outros caracteres desconhecidos. Isso pode ser útil em pesquisas de palavras cujo analista humano precisa classificar variantes de uma palavra. Por exemplo, na pesquisa de combinações do termo “Jo*n”, nomes como “Jon”, “John”, “Joan” e “Johnson” podem ser obtidos. O caractere ponto de interrogação (?), que, por outro lado, combina com cada caractere único em uma série de caracteres. Por exemplo, na pesquisa de combinações do nódulo “Jo?n”, nomes como “Jon”, “John”, “Joan” são obtidos, porém não “Johnson”.

Todas as características descritas até agora são relacionadas especificamente para concordâncias monolíngues, mas há também concordâncias bilíngues. Concordâncias bilíngues podem suportar textos em duas diferentes línguas ao mesmo tempo, enquanto mantém todas as capacidades de um concordanciador monolíngue.



A figura acima mostra a tela do concordanciador bilíngue ParaConc desenvolvido por Michael Barlow (Disponível no sítio: http://athel.com/product_reviews.php?products_id=30).

Compilação

Segundo FERNANDES (2004), podemos dividir a atividade de compilação de um corpus em três estágios principais: (i) desenho do corpus, onde são discutidas as questões teóricas gerais associadas com o planejamento do corpus; (ii) construção do corpus, onde são descritas as decisões técnicas feitas durante a compilação do corpus; e (iii) processamento do corpus, onde são especificados os equipamentos, os programas e o conjunto de ferramentas computacionais utilizados para o processamento do corpus.

Desenho

Muitos pesquisadores por todo o mundo caíram, recentemente, sob os encantos de corpora computadorizados. Pesquisadores sob esse encanto mantêm, frequentemente, uma falsa ideia inicial de que tudo o que precisam para fazer um trabalho baseado em corpus é um computador pessoal, um escaner de mesa com tecnologia de reconhecimento óptico de

caracteres – OCR (Optical Character Recognition), um programa padrão de processamento de corpus, e um grande número de livros. No entanto, quando esse encanto termina, eles descobrem que as coisas não são tão simples quanto, inicialmente, haviam imaginado. Um trabalho baseado em corpus envolve muito planejamento, o estabelecimento de critérios explícitos e rigorosos na seleção de equipamentos, programas e textos. É esse planejamento cuidadoso que possibilita um corpus fornecer descrições precisas e confiáveis, garantindo que ele possa ser utilizado ou referenciado por outros pesquisadores (Kennedy, 1998, p. 70). Além disso, o desenho ideal de um corpus depende muito do objetivo para qual se pretende utilizá-lo e, também, com as questões associadas ao tipo de corpus, a representatividade, os direitos autorais e a seleção dos textos (Sinclair, 1991; Atkins et al., 1992; Baker, 1995; Kenny, 2001).

Construção

Neste segundo estágio de compilação de um corpus é exigida muita paciência e atenção do pesquisador pelo fato da natureza do trabalho manual monótono e repetitivo a ser executado. Além disso, apesar de todo cuidado e trabalho árduo, erros nas versões eletrônicas dos textos são inevitáveis, devido à tecnologia de escaneamento que tem muito que avançar ainda. Entretanto, o pesquisador também deve estar preparado para compreender que o processo de compilação de um corpus é demorado. Em primeiro lugar, o compilador terá que converter os textos fontes e os textos alvos em formato eletrônico. Em seguida, procedimentos de revisão e edição devem ser adotados para a correção dos textos em formato eletrônico. Em seguida, as convenções que informam a codificação de algumas características textuais relevantes devem ser inseridas nos textos em formato eletrônico. Por fim, os estágios para alinhar os textos fontes com os textos alvos, no caso do corpus paralelo, devem ser levados em consideração.

Processamento

Uma nova dimensão para descrição de tradução e para várias outras aplicações desenvolvidas tem sido iniciada por meio de metodologias baseadas em corpus, permitindo a análise automática de textos. Este grau de análise automática contribui para o legado de ferramentas de pesquisa do investigador, pelo desenvolvimento de programas capazes de identificar, classificar, extrair e exibir uma quantidade enorme de dados em vários formatos (Kennedy, 1998, p. 204).

Para a construção de um corpus, na escolha dos equipamentos, deve ser levado em conta aspectos relacionados ao métodos, principalmente, de captura e edição dos textos que compreendem o corpus. Estas especificações podem ser encontradas nos manuais dos periféricos a serem utilizados, como por exemplo, para a captura por meio de escaneamento e/ou reconhecimento de voz, geralmente disponibilizados junto com estes aparelhos ou no sitio do fabricante. Já para os programas de processamento, as configurações mínimas exigidas irão depender do programa a ser utilizado para a execução dessas tarefas, porém em nenhum caso esses tipos de programas exigem uma máquina com uma configuração que um computador pessoal não possa suportar, porém é importante citar que esses programas devem ser, no mínimo, utilizados respeitando suas configurações mínimas, para que seja obtido um resultado satisfatório.

Com relação aos programas, podemos citar alguns utilizados pela maioria da comunidade de pesquisadores e que oferecem recursos capazes de suprir as necessidades do pesquisador, dentre eles estão o *WordSmith Tools* (Versão 3.0, 4.0 e, atualmente, 5.0) e o *Multiconcord* (Versão 1.53), disponíveis comercialmente para o processamento de corpus. Os programas WordSmith Tools e Multiconcord podem processar um corpus que não tenha sido etiquetado ou analisado gramaticalmente, contendo apenas mínimas anotações para indicar estruturas tais como capítulos, parágrafos e sentenças. Além disso, eles também são capazes de executar pesquisas complexas que incluam etiquetas, caracteres coringa (?/*) e/ou/ sem operadores e sequencias interrompidas. Esses dois programas, juntamente com as ferramentas fornecidas por eles para processamento de corpus, ferramentas para utilização em ambiente Windows, são descritos e explicados abaixo.

WordSmith Tools (Versão 3.0, 4.0 e 5.0)

WordSmith Tools (Versão 4.0) – uma poderosa suíte de programas integrados para análises lexicais – foi desenvolvida por Mike Scott (2004) e é distribuída pela Oxford University Press no sítio http://www.lexically.net/wordsmith/. As ferramentas de análise fornecidas pelo WordSmith Tools utilizadas para processar o corpus onde encontram-se suas ferramentas de listagem de palavras (WordList) e de concordância monolíngue (Concord). A ferramenta de Lista de Palavras gera estatísticas descritivas básicas que incluem informações tais como número de arquivos envolvidos, tamanho dos arquivos (em bytes), número de ocorrências e tipos; extensão de palavras (em letras); relação tipos/ocorrências; número de sentenças e parágrafos; extensão de sentenças e parágrafos (em palavras) para textos individuais e para o corpus como um todo. Essas informações estatísticas básicas produzidas pela ferramenta WordList pode ser então utilizada para mostrar fatos interessantes sobre as escolhas lexicais nos textos fontes e textos alvos, bem como um quadro geral da maneira em que as palavras se comportam nesse modelo específico de investigação. A ferramenta Concord auxilia a isolar alguns dos itens de pesquisa ou “nódulos” nos textos fontes analisados, para que eles possam ser analisados subsequentemente com a ferramenta de Multiconcord. A ferramenta de concordâncias exibe todas as ocorrências de um nódulo específico em um corpus de um dado item de pesquisa em uma única coluna. A localização de colocados, identificação de agrupamentos de palavras e itens de pesquisa selecionados manualmente puderam ser feitos por meio da ferramenta Visualizador de Textos.

Multiconcord (Versão 1.53)

O programa Multiconcord (Versão 1.53) é um concordanciador paralelo desenvolvido por David Woolls sob o amparo do Projeto Língua (Woollss, 1998), e é disponibilizado pela CFL Software Development, podendo ser encontrado no sítio http://web.bham.ac.uk/johnstf/lingua.htm. O programa permite ao pesquisador fazer buscas por uma palavra ou uma expressão, do mesmo modo que fazem outros programas de concordância. O resultado da busca é exibido em duas colunas ao invés de uma apenas, tornando possível visualizar como o textos fontes e os textos alvos codificam certos elementos linguísticos, e então detectar os possíveis procedimentos empregados por tradutores quando em contrapartida com esses elementos específicos. Como mencionado anteriormente, o Multiconcord é fornecido com uma ferramenta de marcação, o MinMark (Versão 1.1), que insere uma marcação mínima em formato SGML, exigido pelo

Multiconcord para executar buscas bilíngues (ver *Codificação do Texto* acima). Embora o Multiconcord exija uma marcação mínima para estar apto para processar os textos do corpus, a inserção de etiquetas, outras que não aquelas fornecidas pelo MinMark podem interferir no alinhamento necessário, reduzindo assim ,rigorosamente, a confiabilidade do programa (KENNY, 2002, p. 1240. Por essa razão, um conjunto de textos separados sem etiquetas deve ser utilizado como uma forma de evitar tais interferências. O programa Multiconcord é excelente para a exploração rápida de textos, especificamente para fenômenos com bases lexicais, pois permitem também a utilização os caracteres coringa (*) no início, meio e/ou final de uma palavra ou uma expressão. Além disso, o programa Multiconcord é um programa bastante flexível que não precisa ser específico para a língua dos textos a serem investigados.

ParaConc (Versão 1.0)

O programa ParaConc (Versão 1.0) apresenta-se como uma ferramenta de fácil utilização para executar tarefas de buscas em concordâncias bilíngues. Utiliza-se de uma interface simples e intuitiva, principalmente voltada para o alinhamento de corpus bilíngue, contado com ferramentas que podem trabalhar com textos já etiquetados ou não, o que proporciona ao pesquisador eficiência e rapidez tanto na preparação do corpus de estudo, quanto em suas buscas dentro do corpus. O sistema de resultados conta também com um leiaute baseado na saída de dados no formato de palavra-chave no contexto (KWIC) e também em resultados estatísticos relevantes à dimensão do corpus. Devido às diversas e precisas funções disponíveis por expressões de busca, o ParaConc pode fazer buscas em corpus não somente com base em funções simples, mas também um legado de combinações de buscas as quais são de extrema importância para alcançar o objetivo de pesquisa de um corpus de estudo. Desenvolvido por Michael Barlow, o ParaConc pode ser obtido por meio de seu distribuidor, a Athel, em http://www.athel.com/para.html.

Além das ferramentas mencionadas acima, podemos também citar outras as quais são ferramentas de código aberto, ou seja, que podem ser utilizadas gratuitamente, sem necessidade de uma licença ou de limitações, geralmente ocorridas em versões de demonstração destes programas. Dentre elas podemos citar o KWIC Concordance (Versão 4.7), que possui as funções de um concordanciador bilíngue, leiaute de resultados em formato de palavras-chave em contexto (KWIC) e a possibilidade de operação de um corpus codificado em diversos formatos, incluindo o formato BNC (British National Corpus), um dos formatos mais conhecidos no processamento de corpus. Desenvolvido por Satoru Tsukamoto na Nihon University, o KWIC Concordance é pode ser utilizado sem a necessidade de comprar uma licença e é disponibilizado no sítio http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html.
Descrevemos nesta seção algumas ferramentas fundamentais para a pesquisa em busca em um corpus de estudo, citamos também que existem além dessas ferramentas, outras que podem trabalhar com corpus com interface para Internet e outros sistemas que proporcionam, até mesmo, buscas mais complexas, porém baseados em sistemas Linux.

Aplicabilidade

Dentre os corpora on-line disponíveis para o estudo de língua de sinais podemos destacar três: o *British Sign Language Corpus* [Corpus da Língua de Sinais Britânica]; o ATIS (Air Travel Information System) Sign Language Corpus [Corpus de Língua de Sinais e Sistemas de

Informação de Viagens Aéreas] e o ASL Corpus [Corpus da Língua de Sinais Norte Americana].

The British Sign Language Corpus (BSL) – é parte de um projeto financiado pela Conselho da Pesquisa Social e Econômica da Grã-Betanha e gerenciado pelo Centro de Pesquisa sobre a Língua e Cognição Surda (DCAL) da University College London, mas também inclui outras universidade tais como a Universidade de Bangor (País de Gales), a Universidade Heriot-Watt (Escócia), a Queens University Belfast (Irlanda do Norte) e a Universidade de Bristol (Inglaterra).

O objetivo do projeto é criar um corpus de vídeos clipes mostrando pessoas surdas utilizando a BSL que será disponibilizado on-line. Além disso, o projeto pretende desenvolver pesquisa no que diz respeito à gramática e o vocabulário da BSL e a variação da BSL através do país e como a mesma está mudando.

O BSL Corpus, que se encontra ainda em construção, pode ser acessado no sítio eletrônico http://www.bslcorpusproject.org/ .

The ATIS Sign Language Corpus – é um corpus adequado para a análise de língua de sinais e sistemas estatísticos de tradução automática (ver UNIDADE II). O corpus é baseado em um conjunto de dados sobre sistemas de informações sobre viagens aéreas (Hempphill et al. 1990). Ele contém frases e sentenças transcritas em inglês e extraídas de publicações para a reserva de voos e viagens. Deste conjunto de dados, 595 sentenças foram escolhidas como base.

O corpus foi traduzido com a ajuda de falantes nativos e se encontra disponível em cinco línguas: inglês, alemão, língua de sinais irlandesa (ISL), língua de sinais alemã (DGS) e língua de sinais sul-africana (SASL).

Segundo os compiladores, este corpus é especialmente interessante para sistemas de tradução automática, pois ele está limitado a apenas um domínio. Além disso, o mesmo permite moldar métodos para se lidar com características específicas da língua de sinais tais como o posicionamento de objetos no espaço de sinalização, conforme o uso extensivo de referências espaciais relacionadas a aeroportos e outras localizações.

Mais informações sobre o corpus podem ser obtidas no sítio eletrônico: http://www-i6.informatik.rwth-aachen.de/publications/download/537/Bungeroth-LREC-2008.pdf .

The ASL Corpus – este corpus faz parte de um projeto que pesquisa o reconhecimento baseado em computador dos sinais da ASL. Um dos objetivos é o desenvolvimento de uma interface de pesquisa lexical como parte de um dicionário multimídia da língua de sinais norte americana. Embora, dicionários impressos sobre a ASL existem, eles geralmente estão organizados de acordo com a tradução inglesa mais próxima do sinal da ASL, pois não há forma escrita para a ASL.

O corpus proposto permite que o sinalizante selecione um vídeo clipe correspondente a um sinal desconhecido ou produzir um sinal em frente de uma câmera, para pesquisa. O sistema encontra

uma melhor combinação (s) a partir do seu inventário com milhares de sinais da ASL. Conhecimento sobre as limitações linguísticas da produção de sinais é utilizado para melhorar o reconhecimento.

Mais informações sobre o corpus podem ser obtidas no sítio: http://www.bu.edu/asllrp/ .

Pontos Principais

- A palavra corpus no contexto dos ET pode ser interpretada como uma coleção auto-configurável de textos completos digitalizados, analisáveis automática ou semi-automaticamente e coletados a fim de serem representativos ao máximo do fenômeno tradutório sendo examinado;
- Há um número de diferentes tipos de corpora, mas basicamente eles podem ser classificados em corpora paralelos ou compráveis;
- Técnicas para o processamento de um corpus permitem que o usuário acesse, manipule e exiba as informações contidas em um corpus de várias formas úteis;
- As técnicas fundamentais para o processamento de corpora são listas de palavras e concordanciadores;
- Listas de frequência oferecidos pela maioria das ferramentas de análise de corpora permitem que o usuário descubram quantas palavras diferentes existem em um corpus e com que frequência elas aparecem;
- Concordanciadores recuperam todas as ocorrências de um padrão de pesquisa específico (nódulo) em seu contexto imediato e mostra essas ocorrências em formato de fácil leitura (i.e. KWIC);
- É muito importante observar que as ferramentas de análise de corpora não interpretam os dados – é de responsabilidade do tradutor analisar as informações encontradas em um corpus;
- O uso de ferramentas baseados em corpora para o ensino/aprendizagem de língua de sinais está se tornando cada vez mais comum em várias partes do globo. Exemplo disso são os corpora: BSL, ATIS e ASL.

Referências

ATKINS, S., Clear, J., & Ostler, N. Corpus Design Criteria. Literary and Linguistic Computing, 7(1), 1992, 1-16.

Baker, M. “Corpora in Translation Studies. An Overview and Suggestions for Future Research”. Target, 7(2), 1995, 223-243.

Baker, M. "Corpus Linguistics and Translation Studies. Implications and Applications", in: Baker et al., 1993, 233-250.

BIBER, D. “Representativeness in Corpus Design”. Literary and Linguistic Computing, 8(4), 1993.

BOWKER, L. Towards a Corpus-based Approach to Terminography. Terminology, 3(1): 27-52, 1996.

CONCISE OXFORD DICTIONARY [CD-ROM]. Oxford: Oxford University Press, 1996.

FERNANDES, L. Brazilian Practices of Translating Children's Fantasy Literature: A Corpus-based Study. Unpublished PhD Thesis, Universidade Federal de Santa Catarina, 2004.

FERNANDES, L. "Corpora in Translation Studies: Revisiting Baker's Typology". Revista Fragmentos, vol. 30, pp. 87-112, 2006 .

KENNEDY, G. An Introduction to Corpus Linguistics. London/New York: Longman, 1998.

KENNY, D. Corpora in Translation Studies. In Mona Baker, Routledge Encyclopedia of Translation Studies. London/New York: Routledge, 1998, pp. 50-53

KENNY, D. Lexis and Creativity in Translation. A Corpus-based Study. Manchester, UK: St Jerome, 2001.

MATTHEWS, P. (1997). The Concise Oxford Dictionary of Linguistics. Oxford: Oxford University Press.

McENERY, T., & Wilson, A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996.

MUNDAY, J. A Computer Assisted Approach to the Analysis of Shifts. Meta, 43(4), 1998, 543-556.

OLOHAN, M. Introducing Corpora in Translation Studies. London/New York: Routledge, 2004.

PEARSON, J. Terms in Context. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.

SCHÄFFNER, C. (1998). Parallel Texts in Translation. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), Unity in Diversity? Current Trends in Translation Studies. Manchester, UK: St. Jerome.

SHUTTLEWORTH, M., & Cowie, M. Dictionary of Translation Studies. Manchester, UK: St Jerome, 1997.

SINCLAIR, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.

VINAY, J.-P., & Darbelnet, J. Comparative Stylistics of French and English: A Methodology for Translation (J. C. Sager & M.-J. Hamel, Trans.). Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

ZANETTIN, F., Bernardini, S. & Stewart, D. Corpora in Translator Education. Manchester, UK: St. Jerome Publishing, 2003.

ZANETTIN, F. Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In M. Olohan (Ed.), Intercultural Faultlines. Research Models in Translation Studies I Textual and Cognitive Aspects. Manchester, UK: St Jerome Publishing, 2000.

Bibliografia Completa

ARROJO, R. Oficina de Tradução: A teoria na prática. 3º edição. São Paulo: Editora Ática, 1997.

ATKINS, S., Clear, J., & Ostler, N. Corpus Design Criteria. Literary and Linguistic Computing, 7(1), 1992, 1-16.

AUSTERMÜHL, Frank. Electronic Tools for Translators. Manchester, UK: St. Jerome Publishing, 2001.

Baker, M. “Corpora in Translation Studies. An Overview and Suggestions for Future Research”. Target, 7(2), 1995, 223-243.

Baker, M. "Corpus Linguistics and Translation Studies. Implications and Applications", in: Baker et al., 1993, 233-250.

BAR-HILLEL, Y. Automatic Translation of Languages, 1960. Disponível em: http://www.mt-archive.info/Bar-Hillel-1960.pdf. Acessado em: 10/07/2008.

BARTHOLOMEI, Lautenai. Wordfast: Utilização e Avaliação em um Projeto de Tradução. Monografia de Especialização em Língua Inglesa: Ênfase em Tradução. Chapecó, SC: UNOCHAPECÓ, 2008.

BIBER, D. “Representativeness in Corpus Design”. Literary and Linguistic Computing, 8(4), 1993.

BOWKER, L. Towards a Corpus-based Approach to Terminography. Terminology, 3(1): 27-52, 1996.

BOWKER, Lynne. Computer-Aided Translation Technology. A practical introduction. Ottawa: University of Ottawa Press, 2002.

BOWKER, Lynne. Computer-Aided Translation Technology. A practical introduction. Ottawa: University of Ottawa Press, 2002.

CONCISE OXFORD DICTIONARY [CD-ROM]. Oxford: Oxford University Press, 1996.

European Association for Machine Translation (EAMT). Disponível em: http://www.eamt.org/mt.html. Acessado em: 10/07/2008.

FERNANDES, L. “Corpora in Translation Studies: Revisiting Baker’s Typology”. Revista Fragmentos, vol. 30, pp. 87-112, 2006 .

FERNANDES, L. Brazilian Practices of Translating Children’s Fantasy Literature: A Corpus-based Study. Unpublished PhD Thesis, Universidade Federal de Santa Catarina, 2004.

HEYN, Matthias. Translation Memories: Insights and Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.). Unity in Diversity? Current Trends in Translation Studies. Manchester, UK: St. Jerome Publishing, 1998.

HUTCHINS, W.J. History of MT in a nutshell. A two-page sketch, from the beginnings to the present, 2001.

HUTCHINS, W.J; SOMERS, H.L. An introduction to machine translation. London: Academic Press. 1992.

K. Ahmad, W. Martin, M. Hoelter, M. Rogers. Aspects of Terminology Infrastructure in Europe: Volume 3 - Specialist Terms in General Language Dictionaries. POINTER Report (Available as a University of Surrey report), 1995.

KENNEDY, G. An Introduction to Corpus Linguistics. London/New York: Longman, 1998.

KENNY, D. Corpora in Translation Studies. In Mona Baker, Routledge Encyclopedia of Translation Studies. London/New York: Routledge, 1998, pp. 50-53

KENNY, D. Lexis and Creativity in Translation. A Corpus-based Study. Manchester, UK: St Jerome, 2001.

KRINGS, Hans P. 'Translation problems, and Translation Strategies of Advanced German Learners of French (L2)'. IN: Juliane House and Shoshana Blum-Kulka (Orgs.). Interlingual and Intercultural Communication – discourse and Cognition in Translation and Second Language Acquisition Studies.Tübingen: gnv- Gunter Narr Verlag, 1986.

MAKOTO, N. A framework of a mechanical translation between Japanese and English by analogy principle. In ELITHORN, A; Banerji, R. Artificial and Human Intelligence. Elsevier Science Publishers, 1984.

MATEUS, M. H. M. Tradução automática: um pouco de história. In Engenharia da Linguagem. Maria Helena M. Mateus e António Horta Branco (Orgs.). Lisboa, Edições Colibri, 1995, pp. 115-120.

MATTHEWS, P. (1997). The Concise Oxford Dictionary of Linguistics. Oxford: Oxford University Press.

McENERY, T., & Wilson, A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996.

MELBY, Alan e WARNER, Terry C. The Possibility of Language: A Discussion of the Nature of Language with Implications for Human and Machine Translation. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

MUNDAY, J. A Computer Assisted Approach to the Analysis of Shifts. Meta, 43(4), 1998, 543-556.

OLOHAN, M. Introducing Corpora in Translation Studies. London/New York: Routledge, 2004.

PEARSON, J. Terms in Context. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.

Pierce, J. R; CARROLL, J.B; et al. Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

POINTER - Proposals for an operational Infrastructure for terminology in Europe. Disponível em: http://www.computing.surry.ac.uk/ai/pointer/report/intro.html, 1996.

SANTOS, D. “Tradução automática”. Material de ensino na Escola de Verão da Linguateca, 2006. Disponível em: www.linguateca.pt/escolaverao2006/TA/TraducaoEscolaVerao.pdf. Acessado em: 10/07/2008.

SCHÄFFNER, C. (1998). Parallel Texts in Translation. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), Unity in Diversity? Current Trends in Translation Studies. Manchester, UK: St. Jerome.

SHUTTLEWORTH, M., & Cowie, M. Dictionary of Translation Studies. Manchester, UK: St Jerome, 1997.

SINCLAIR, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.

SOUZA, Vinícius Costa. Sign WebMessage: um ambiente para comunicação via web baseado na escrita de Libras. Trabalho de conclusão - Unisinos. São Leopoldo, 2002.

VAUQUOIS, B. A survey of formal grammars and algorithms for recognition and transformation in machine translation, IFIP Congress-68 (Edinburgh), 1968, pp. 254-260.

VINAY, J.-P., & Darbelnet, J. Comparative Stylistics of French and English: A Methodology for Translation (J. C. Sager & M.-J. Hamel, Trans.). Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

Wright, Sue Ellen & Leland D.Wright. 'Terminology Management for Technical Translation'. in Sue Ellen Wright & Gerhard Budin, The Handbook of Terminology Management, Vol.1. Ps. 147-159. Amsterdam & Philadelphia: John Benjamins Publishing Company, 1997.

Wüster, Eugen. Die allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften (The Theory of General Terminology – An Intersection of Linguistics, Logic, Ontology, Information Science, and the Technical Sciences), 1972.

ZANETTIN, F. Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In M. Olohan (Ed.), Intercultural Faultlines. Research Models in Translation Studies I Textual and Cognitive Aspects. Manchester, UK: St Jerome Publishing, 2000.

ZANETTIN, F., Bernardini, S. & Stewart, D. Corpora in Translator Education. Manchester, UK: St. Jerome Publishing, 2003.