Os corpora eletrônicos nos estudos da tradução automática

Cleydstone Chaves dos SANTOS*

Resumo[1]: Os corpora têm assumido um papel importante em muitos campos da pesquisa no cenário acadêmico. Recentemente, no contexto da tradução, os estudos baseados em corpora ganharam grande importância como método de pesquisa (FERNANDES, 2006), revelando mais e mais a necessidade de sua investigação. Nesta perspectiva, o presente ensaio tem por objetivo revisar alguns dos principais teóricos (Baker, 1993; 1995; KENNY, 1998; BERBER SARDINHA, 2002, OLOHAN, 2004) na trajetória dos estudos baseados em corpora aplicada à pesquisa em tradução, discutindo suas contribuições no crescente campo de tradução automática (doravante TA). Neste sentido, tem-se observado um progresso gradativo na qualidade de textos traduzidos automaticamente desde a chegada dos corpora eletrônicos.
Palavras-chave: Corpora. Contribuições. Tradução Automática.

Abstract: Corpora have assumed an important role in many fields of the academic setting research. Recently, in the context of translation, corpora based studies have gained great importance as a research method (FERNANDES, 2006), revealing more and more the need for its investigation. Embarking on this perspective, this essay aims at revisiting some of the main scholars (BAKER, 1993; 1995; KENNY, 1998; BERBER SARDINHA, 2002, OLOHAN, 2004) in the trajectory of corpora based studies applied to translation research by discussing their relevance in the increasing field of machine translation. On this very sense, it has been observed a gradual advance in the quality of automatically translated texts since the arrival of electronic corpora.
Keywords: Corpora. Contributions. Machine Translation.

1. Introdução

No âmbito dos Estudos da Tradução (doravante ETs), assim como na linguística de corpus, converge a crença de que o uso de corpora *"resulta de uma coletânea de textos para várias formas de análise linguística"* (AUSTERMÜHL, 2010, p.124). Isto pode ser possível através da técnica de alinhamento com uma determinada abordagem, perspectiva e ou paradigma (FERNANDES, 2006).

Inserida nesta dimensão, Olahan (2004) traz em seu discurso ecos de todo um percurso de estudiosos na área dos ETs (BAKER, 1993; 1995; KENNY, 1998; BERBER SARDINHA, 2002) que buscaram estabelecer o trabalho com corpora como uma metodologia de pesquisa, de modo que pudesse atender a uma gama de fins específicos, tais como: a) o trabalho com textos autênticos; b) a busca e quantificação de muitos dados em curto prazo; c) a rápida

* Doutorando pela Universidade Federal de Santa Catarina - PGET e professor de Língua Inglesa da Unidade Acadêmica de Letras (UAL). E-mail: teachertone@gmail.com.

[1] Tradução automática: revisão minha.

comparação com outros textos através da técnica de alinhamento, dentre outros. Neste contexto, é também viável trazer à tona o que Baker (1995) já definia como corpora, já que foi a partir de sua obra que se teve inicio a desenfreada corrida para o estabelecimento dos corpora como uma metodologia de pesquisa nos ETs:

> Em suma, um corpus nos Estudos de Tradução em Corpora não é simplesmente uma ampla coletânea de textos escritos ou falados como as definições tradicionais postulam. Define-se mais precisamente como qualquer coletânea em aberto de textos digitalizados analisáveis automaticamente ou semi-automaticamente, e organizados de tal maneira a fim de serem representativos do fenômeno de tradução sob análise[2] (BAKER, 1995, p.225). (Tradução automática revisão minha).

Tal inquietação manifestava todos seus anseios acerca da integração dos corpora como uma ferramenta metodológica nos ETs. É, portanto, aceitável concordar com a postura de "*Baker ao estabelecer um programa de pesquisa que seria seguido por outros investigadores*" (BERBER SARDINHA, 2004, p.25), já que, desde então, vem gerando verdadeiros frutos entre os pesquisadores desta área. Além do mais, tem-se observado uma considerável influência resultante da urgência do desenvolvimento de uma metodologia baseada em corpus na pesquisa nos ETs.

Contudo, no que tange à pesquisa de corpora aplicados à Tradução Automática (TA), tem-se observado pouca atenção nos ETs no cenário nacional. Daí, portanto, a escassez de diálogos mais consistentes quanto às potencialidades e limitações da TA enquanto ferramenta de apoio à tradução humana em contexto profissional (ALFARO & DIAS, 1998).

Tomando como ponto de partida esse pensamento, este ensaio discute até que ponto pode ser relevante a pesquisa em TA baseada em corpora, levando em consideração seu caráter incipiente no crescente campo dos ETs, e como também focalizando as aplicações metodológicas dos corpora mediante o escopo dos sistemas de TA.

2. A linguística de corpus nos primeiros Estudos da Tradução Automática

[2] All in all a corpus in CTS is not simply a large body of written text or spoken material as traditional definitions have often implied. It is defined more accurately as any open-ended body of machine-readable full texts analyzable automatically or semi-automatically, and sampled in a principled way in order to be maximally representative of the translation phenomenon under examination (BAKER, 1995, p.225).

Para se compreender a dimensão dos atuais estudos da linguística de corpus aplicados aos ETs e com desdobramentos na TA, deve-se levar em conta que o paradigma cientifico das teorias linguísticas nas pesquisas referentes a esses estudos tivera seu inicio durante o próprio desenvolvimento de cada uma das teorias, como será listado a seguir.

Sedundo Guidére (2010) o "*boom"* das teorias de equivalência, nos ETs e em TA nas décadas de 50 e 60, manifestava uma tendência mais *intra-organismo*, não abrangendo, naquele momento, a complexidade do ato social da tradução. Nesse viés, teóricos partiam para uma comparação entre línguas, sem considerar em sua totalidade propósitos de natureza comunicativa, com foco apenas em seus padrões estruturais, desvinculados da construção de sentidos, de usos em contextos sociais distintos, e, portanto, a tradução era vista apenas como mero ramo da linguística aplicada para análise contrastiva.

É, portanto, em meio a esse avanço que os primeiros programas de TA começam a surgir. Embarcando nesta fase promissora, nas décadas de 30, datam-se as tentativas do russo Smirnov-Trojanskij com sua proposta de sistema automático que traduzia exemplares linguísticos em diversas línguas, cuja abordagem seguia padrões pré-estabelecidos em corpora paralelos. A década de 40, por sua vez, inova com traduções de cunho automático a partir da utilização de uma calculadora. Contudo, tal dispositivo não surtiu grandes resultados devido ao seu limitado repertório de combinações linguísticas a partir da utilização de um corpus paralelo. Entretanto, acabou impulsionando a pesquisa em TA, que, por sua vez, ainda estava mais voltada para uma tendência de cunho estruturalista, ou seja, para a construção de aspectos de natureza morfossintática do que semântica numa abordagem conhecida como abordagem direta (SOMERS, 2001, HUTCHINS, 2000).

Como resultado, a pesquisa em TA não poderia ir de encontro ao paradigma científico instaurado naquela época, acentuando-se, então, a busca pela análise contrastiva (c.f: SELINKER, 1969) na comparação de línguas. A partir, então, desse caráter visionário, embora com aparentes traços rudimentares, do uso da ferramenta de TA, insistiu-se num projeto de melhoria do seu processo tendo em vista resultados mais abrangentes.

Naquele momento, a pesquisa em TA começa a investir nos chamados sistemas baseado em regras gramaticais (ZANETTIN ET al, 2003) através de uma abordagem de corpus comparado e ou paralelo apenas. Dentre os quais, destaca-se o experimento Georgetown da IBM na década de 50. Esse sistema, por sua vez, concebia a TA em nível lexicográfico apenas, isto é, a partir de um repertório não muito significativo, contendo cerca

de não mais que 250 exemplos de usos da língua alvo e da língua de chegada (WILKS, 2009), a tradução era realizada, alcançando resultados satisfatórios dentro do seu escopo.

Em meados dessa mesma época, começa a se desdobrar a pesquisa em TA, todavia, com um escopo limitado (SOMERS, 2001). A chamada primeira geração de TA visava, até então, atender fins de natureza militar com o intuito inicial de decifrar códigos no período da guerra fria entre os EUA e a extinta União Soviética. Segundo Hutchins (2000), isto só era possível porque as primeiras versões de tradução realizadas automaticamente tinham como base a abordagem de tradução direta, que, por sua vez, era pautada num sistema de natureza lexicográfica cujo método consistia na pesquisa em dicionários armazenados na memória da máquina, sendo traduzida, assim, palavra por palavra através do rastreamento de termos correspondentes ou equivalentes entre a língua de partida e a língua de chegada.

3. Os corpora como abordagem do processo de TA

Referente à abrangência e velocidade que as ferramentas de corpora podem proporcionar, é na interação com determinados aspectos da linguagem, sejam eles: linguísticos, culturais, discursivos e sociais, que se pode propiciar ao pesquisador o acesso a "*uma coleção de usos de uma língua para observação de linguagem autêntica"* (VIANA & TAGNIN, 2010, p.19). No entanto, concernente à TA, os três últimos aspectos da linguagem ainda são um grande desafio na atualidade, sendo foco de muitas pesquisas em andamento no cenário nacional (CASELI, 2009; 2010).

Atualmente, os estudos de corpora encontram-se ancorados no próprio desenvolvimento de ferramentas tecnológicas (TORO, 2007), dado esse que vem cada vez mais crescendo. Neste contexto, o cerne da questão é essencialmente lidar com a análise de uma ampla quantidade de textos digitalizados e disponibilizados também num sistema integrado da rede mundial de computadores, resultando nas seguintes categorias: *corpora paralelos; corpora multilíngues; corpora comparáveis*.

Segundo Fernandes & Bartholamei Jr. (2004), numa perspectiva gradual no que concerne à relevância dos corpora nos ETs, Baker (1993) em seu artigo intitulado "*Corpus Linguistics and Translation Studies*", argumenta que a disponibilidade de grandes corpora seja de textos originais e ou traduzidos deveria se aliar ao desenvolvimento de uma metodologia baseada em corpus. Para os autores, a referida pesquisadora acredita que tal

metodologia permitiria aos acadêmicos dos ETs a proeza de revelar a natureza do texto traduzido enquanto um evento comunicativo mediado.

Em sua revisão à tipologia proposta por Baker (op.cit), Fernandes (2006) apresenta uma reorganização para as três categorias sugeridas pela pesquisadora (*corpus comparável, corpus paralelo e corpus multilíngüe*), resultandou em apenas duas conforme o argumento abaixo:

> Em meu ponto de vista, a classificação tri partitiva proposta por Baker pode ser reorganizada sob a categorização de dois ângulos apenas: o comparável e o paralelo. Isto se deve ao fato de que o termo multilíngüe não apresenta quaisquer características contrastivas que o distinga dos demais tipos(...) (FERNANDES, 2006, p.4)[3]

Embora a gama de vantagens até então listadas, há uma preocupação compartilhada por alguns estudiosos da área (TYMOCZKO, 1998; IAN MASON, 2001; OLAHAN, 2004) sobre generalizações vagas realizadas, por alguns pesquisadores, a partir de dados quantitativos, que acabam por desconsiderar grande parte significativa do caráter qualitativo de qualquer pesquisa bem como do papel relevante desempenhado pela intuição do tradutor.

Nesta linha de pensamento, Tymoczko (1998 apud BERBER SARDINHA, 2002) acredita que o pesquisador que lança mão dos corpora enquanto ferramenta metodológica, fazendo apenas recortes quantitativos, na interpretação dos seus dados, corre o risco de transformar um estudo, que, de certa forma, deveria apresentar um caráter descritivo, em um ato de cunho apenas prescritivo, resultando, assim, em verdadeiras formulas de natureza cientificista. Desta feita, o caráter intuitivo e interpretativo do tradutor passa a ser indiscutivelmente rejeitado.

Contudo, há um consenso na referida literatura no que diz respeito aos ETs com base em corpus. Tal concordância ocorre sempre que há uma possível combinação de uma análise de cunho quantitativo e outra de caráter qualitativo, na exploração dos chamados *fatores pragmáticos*, ou seja, fatores de ordem contextual e co-textual que necessariamente remetem a tipologias, discursos, gêneros, registros, jargões e etc.

De fato, conforme comprovado na literatura (OLAHAN, 2004), não há como separar tais abordagens, visto que toda análise quantitativa caminha para um determinado recorte das descobertas, resultando numa perspectiva qualitativa do objeto de estudo. Não muito

[3] "In my view, Baker's tripartite classification can be re-arranged under only two main categories: *comparable* and *parallel*. This is due to the fact that the term *multilingual* does not have any contrastive feature that could make it distinctive from the other two types of corpora (...)".– (Tradução automática revisão minha)

diferente, ocorre num estudo que se rotula primordialmente como qualitativo, uma vez que dados quantitativos são observados para que haja uma descrição propícia do que se procura provar, contestar, replicar dentre outros objetivos.

Acredita-se, ainda, que não é a ênfase em uma determinada abordagem que dará maior credibilidade a pesquisa a que se pretende realizar. Ao contrário, será a partir de um equilíbrio entre ambas, como ferramenta de apoio no uso de corpora enquanto metodologia, que poderá prover ao pesquisador uma visão mais acurada do que se pretende investigar. Em suma, a visão de se estabelecer um caráter metodológico ao trabalho com corpus reconstrói todo um histórico de pesquisa que pode ser considerada como visionária para sua época, reafirmando-se a cada ano através de seus fins sólidos e de sua proposta de trabalho. Portanto, parece modo ser viável refletir sobre as possíveis aplicações dos corpora nos estudos de TA.

Conforme as discussões até então observadas, compreende-se que há, de fato, na literatura em questão, uma busca continua por uma metodologia que utilize os corpora nos ETs, dado esse que tem se revelado como uma constante corroborada em vários outros estudos (KENNY, 1998; BERBER SARDINHA, 2002; OLAHAN, 2004; FERNANDES & BARTHOLAMEI JR, 2004; FERNANDES, 2006). Contudo, ainda parece haver pouco interesse quanto a suas aplicações nos estudos em TA, devido à própria resistência à natureza automática da tradução como um avanço nos ETs.

Em analogia, pode-se fazer uma comparação com o progresso da pesquisa em outras áreas do saber. Na engenharia, a utilização de ferramentas tecnológicas - como calculadoras de natureza cientifica- softwares avançados, auxilia a resolução de cálculos que tomariam muito tempo do profissional desse campo. Não muito diferente, na medicina utilizam-se equipamentos de caráter computacional para realização de exames que possibilite um diagnóstico mais eficaz do quadro de saúde de um paciente. Nesta perspectiva, questiona-se o porquê de pesquisadores em tradução resistirem à imersão de aparatos tecnológicos como ferramentas de trabalho?

Esse questionamento pode ser o ponto de partida para uma reflexão acerca do caráter cientifico da utilização de corpora como ferramenta de pesquisa. Portanto, no tocante aos Estudos de TA, os corpora podem assumir um papel fundamental, a principio, sob dois ângulos: a) primeiramente pelo diálogo entre a linguística de Corpus e os ETs via pesquisa em TA; b) e segundo, por serem os corpora a base de geração de vários tipos de TA.

Historicamente, esse último ângulo remete diretamente à gênesis da TA, uma vez que utilizando os corpora como abordagem do processo de tradução na sequenciação automática: *geração, análise e tradução* (WILKS, 2009), deu-se origem aos primeiros tipos de tradutores automáticos, dentre os quais destacam-se: a)TA baseada em exemplos e ou em analogias; b)TA baseada em memórias; c)TA baseada em casos ou regras; d) TA de base estatística e e)TA de base híbrida.

Na figura 01 abaixo, tem-se uma ilustração da TA baseada em exemplos e ou em analogias (SOMERS et al, 2010). Nela, a geração de heurísticas, ou seja, um método de investigação baseado na aproximação progressiva de um dado problema (do grego *heurískein*, achar; descobrir; encontrar), da língua fonte busca equivalentes (c.f: BAKER, 2011) com intuito de solucionar um dado problema na língua alvo. Neste modo particular de pensar, o processo tradutório, neste caso automático, passa a ser considerado como um método de resolução de um problema, uma vez que parte de um principio inserido numa lógica algorítmica de cunho matemático (c.f: SANTOS, 2011). Consequentemente, seguindo um principio de analogias, verificam-se os termos equivalentes no corpus elaborado, geralmente um corpus paralelo, a seguir fornecem-se as possíveis resoluções.

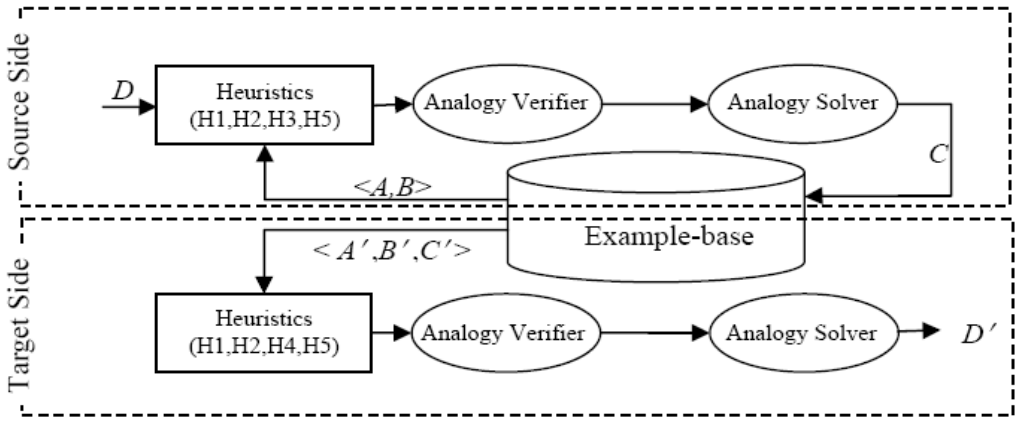

Figura 01. TA baseada em exemplos e ou em analogias

Como argumentado anteriormente, através desta ótica a tradução é vista enquanto um problema a ser resolvido, não muito diferente de alguns contextos vivenciados na tradução humana (ALFARO; DIAS, 1998), como o dilema enfrentado por muitos tradutores literários mediante a questão da recriação ou da transliteração da simbologia, da sonoridade e bem como das rimas encontradas em um determinado poema.

Embarcando na concepção acima, tem-se a visão de que os corpora nos estudos de TA apenas se constituem como coletânea de grupos de palavras e sentenças. Neste sentido, predominava apenas a busca de equivalentes entre uma língua de partida (L_2) e uma língua de chegada (L_1). Na verdade, esse era o quadro em que se inseriam nos primeiros anos da TA. Na atualidade, todavia, os programas de TA de base estatística lançam mão de corpora eletrônicos cuja busca se dá num repertório de textos diversos disponíveis on-line. Essa constatação caracteriza a visão de Baker (1995) a seguir como questionável:

> Na tradução automática, no entanto, um corpus não necessariamente consiste numa coletânea de textos; pode ser não mais que um conjunto de exemplos (Schubert, 1992:87). Uma das definições de corpus nesta área, entretanto, é a de "coletânea finita de sentenças gramaticais usadas como base para análise descritiva de uma língua". (BAKER, 1995, p.225)[4]

Embora pareça obsoleto, esse discurso pode também ser caracterizado como relevante. Isto porque historicisa o avanço dos estudos em TA na interface com os corpora enquanto critério de uma abordagem processual, ou seja, é através deles que ocorrem os processos de verificação e solução de analogias entre L_1 e L_2. Neste momento, a pesquisa em TA começava a investir nos chamados sistemas baseado em regras gramaticais (ZANETTIN et al, 2003). Neles, consideravam-se, inicialmente, apenas seis regras de cunho linguístico normativo-descritivo tendo em vista à formação de segmentos entre a língua russa e a língua inglesa (SOMERS, 1998).

A partir desse primeiro momento, a TA baseada em regras passou a recorrer a uma abordagem de corpora como base num sistema de geração e busca através do processo de análise, transferência e síntese de regras gramaticais. Tais regras contemplavam o caráter morfológico, sintático e semântico a partir da segmentação do léxico e da sentença equivalentes entre L_1 e L_2 conforme dados do site http://www.linguatec.net/products/tr/information/technology/mtranslation, como pode ser observado na figura 02 a seguir:

[4] *In machine translation, by contrast, a corpus does not necessarily consist of running texts; it may be no more than a set of examples (Schubert, 1992:87). One of the definitions of corpus in this field is therefore "the finite collection of grammatical sentences that is used as a basis for the descriptive analysis of a language"(definition given in the Glossary of Terms in Newton, 1992:223). (BAKER, 1995, p.225)* (definição extraída do Glossário de Termos em Newton, 1992:223).

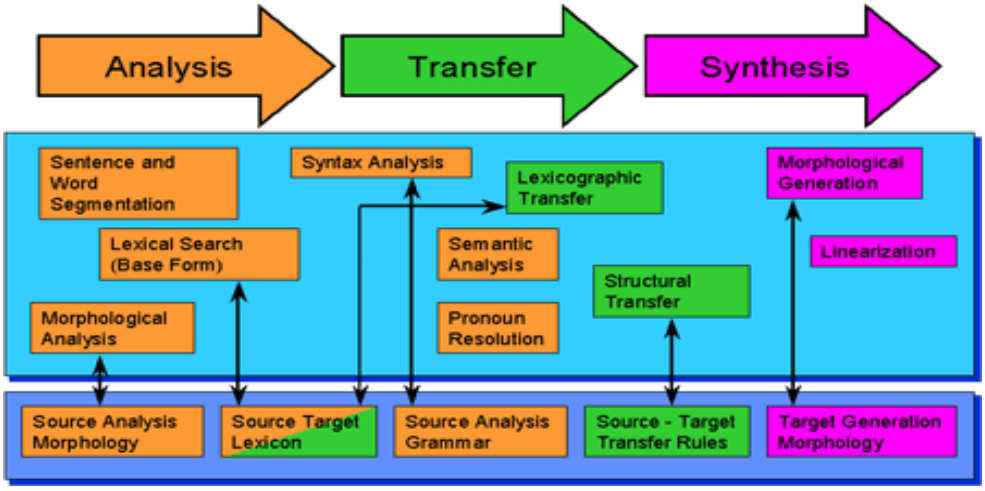

Figura 02- TA baseada em regras

Compreende-se, então, que a TA baseada em regras já permitia ao usuário uma maior confiabilidade dos seus resultados, em virtude de seus níveis de análise da natureza semântica da linguagem. Nele, possibilitava-se uma maior combinação de pares linguísticos na formação de sentenças entre L_1 e L_2 no seu processo de tradução. Nesse sentido, parte-se de uma análise de aspectos de cunho morfológico, seguida da segmentação sintática do léxico e da frase de L_2, transferindo-se os padrões equivalentes para L_1 através da resolução das diferenças nesses níveis até a síntese por meio da etapa de linearização entre os segmentos de ambas as línguas.

Porém, segundo Smith (2001) a década seguinte a esse pequeno avanço testemunharia a idade das trevas da pesquisa em TA, em virtude de um longo período de estagnação, resultante do relatório da ALPAC (Automatic Language Processing Advisory Committee - Comitê Assessor de Processamento Automático das Línguas) que:

> "(...) 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.". (BARTHOLOMEI JR.; FERNANDES, 2004, p. 17).

Em meio a essa estagnação da pesquisa em TA, a ideia de *a tradução ser uma ramificação da linguística aplicada* é retomada e ampliada por Catford (1965 apud GUIDÉRE op. cit) em "*A linguistic Theory of Translation*", que fortemente influenciado por Firth e Halliday (GUIDÉRE, op.cit), passa a considerar a teoria da tradução um ramo da linguística comparada. Daí por diante, segundo Toro (2007) observa-se uma reação em cadeia

na comunidade cientifica dos ETs a partir dos desdobramentos de diversos estudos linguísticos, que foram adquirindo uma característica mais social.

Nas décadas de 70 e 80, com as abordagens funcionalistas, surge uma tendência comunicativa da tradução. Já entre os anos 80 e 90, desenvolvem-se as abordagens discursivas agora mais voltadas para a questão do contexto, Baker (1992), Hatim e Mason (1990 apud TORO, 2007) são nomes que merecem destaque nesse período. Para Toro (op.cit), as abordagens comunicacionais introduzidas por Hatim e Mason direcionam o foco dos ETs para as dimensões que o contexto pode apresentar, sejam elas: *comunicativa, pragmática e semiótica.* Neste período, a TA também assume um viés de natureza social, começando a servir como ferramenta de tradução no setor industrial, em eventos internacionais bem como na transmissão de boletins de tempo.

Como resultado, a pesquisa em Processamento de Línguas Naturais (PLN) só encontra espaço após duas décadas através dos programas de TA de base estatística, em virtude do advento da inteligência artificial no campo computação, culminando com a chegada, desenvolvimento e progresso da internet a partir dos anos 90. A TA de base estatística, por sua vez, passa a utilizar como principio uma variada gama de recursos na busca pela adequação de suas traduções a partir de textos que compõem os corpora eletrônicos em rede, como se verifica na figura 03 abaixo do site http://www-i6.informatik.rwth-aachen.de/web/Research/machine_trans.html:

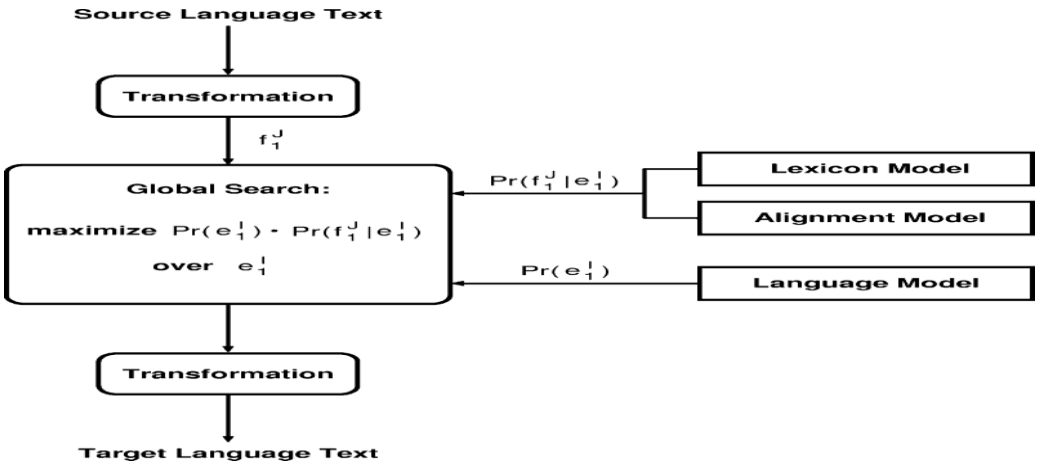

Figura 03 – A TA de base estatística

Na figura acima o processo tradutório automático de um texto numa dada língua alvo passa por uma busca global determinada por algoritmos específicos, nesse caso nos chamados

corpora eletrônicos disponíveis em rede. Aqui, ocorre também um filtro dessas buscas a partir de um processo de alinhamento lexical, que se utiliza de modelos linguísticos a fim de solucionar questões de ordem morfológica, sintática e semântica. Após essa etapa, o texto é transformado, ou seja, traduzido, como aponta o diagrama anterior.

Consequentemente, a partir dessa inovação do processo de TA, através dos corpora on-line, muitos sistemas de TA dessa natureza foram chegando ao mercado, uns disponíveis em rede gratuitamente outros pagos. No caso dos gratuitos, tem-se o Google Tradutor na figura 04 a seguir, que faz buscas de documentos traduzidos disponíveis em rede para geração de suas traduções, utilizando também de memórias de tradução geradas a partir das diversas buscas por usuários no mundo inteiro e sugestões de seus usuários:

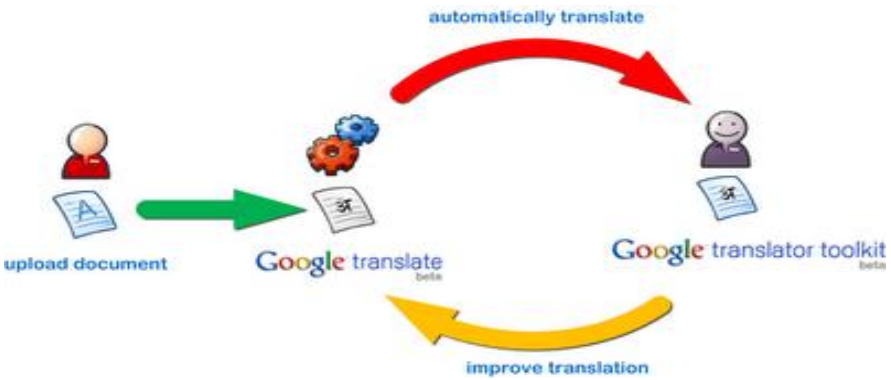

Figura 04- Sistema de busca do Google Tradutor

No caso de alguns programas pagos de TA, tem-se algumas versões do SYSTRAN, que, por sua vez, utiliza uma geração de base híbrida (JOHNSON, 2012), como se ver na figura 5 abaixo:

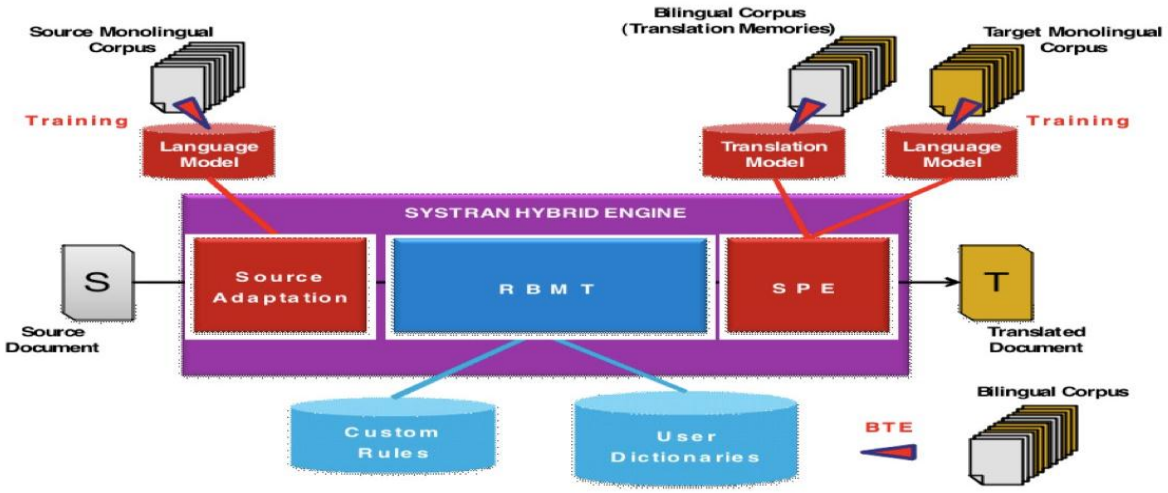

Figura 04 – A TA de base híbrida

Nesta perspectiva, há de se concordar com Tymoczko (1998) que já acreditava nos efeitos positivos do uso de corpora nos ETs. Semelhantemente, não poderia ser tão diferente naqueles resultantes da utilização da TA. Certamente, através da metodologia de corpora acima, ora servindo como base para geração de tradução ora atuando para atender fins de pesquisa, a TA de natureza híbrida utiliza uma gama de recursos para arquitetura de seu processo. Esse processo é definido por um algoritmo específico que permite partir da fase de adaptação do texto fonte em face de um corpus monolíngue, adentrando seus dados para análise em um sistema de regras. Esse sistema, por sua vez, utiliza dados dicionarizados, buscando equivalentes, a partir de dois corpora: um de memórias de tradução e outro monolíngue da língua de chegada, gerando, assim, o documento final. Entretanto, em virtude de seu caráter híbrido é possível que haja um maior intercâmbio de dados para ambos os pesquisadores e usuários desse tipo de TA (JOHNSON, 2012). Contudo, ainda limitados ao repertório linguístico armazenado nos diversos tipos de corpora que dispõe tal mecanismo de tradução.

Face ao exposto, admite-se que os corpora dão mais modernidade à área dos ETs, trazendo-a mais perto do que se espera da pesquisa contemporânea. Neste patamar, os estudos atuais de TA também podem incorporar a sua prática de investigação tais ferramentas, já que desde os primórdios, como visto anteriormente, os corpora fizeram e ainda fazem parte da arquitetura e do processo de TA.

Daí então, lançar mãos dos corpora enquanto metodologia de pesquisa aplicada aos estudos de TA possivelmente reavivaria o caráter cientifico da mesma, estabelecendo-a no

âmbito dos ETs não apenas como subárea da linguística computacional[5], mas também como ramo próprio dos ETs. Assim, mesmo embora não tenha sido amplamente contemplada por Holmes (1972), a partir de seu mapeamento do campo disciplinar dos ETs, a TA passa a ser vista como um campo de estudo de natureza interdisciplinar, possibilitando aos pesquisadores um constante diálogo com outras áreas do saber que também lançam mão da linguagem como objeto de estudo. Em suma, cabe concordar com o próprio Holmes ao alegar que se faz necessário que os ETs dediquem-se sua atenção às principais ramificações, para que possa se desenvolver como disciplina madura e estabelecida (HOLMES, 2000).

Neste sentido, embora se observe no contexto internacional um constante avanço na pesquisa em TA (SANTOS, 2011), pouco se tem levado em conta a importância que os corpora, como metodologia nas concepções de Baker (*op. cit.*), podem representar para uma guinada nestes estudos a nível nacional. Segundo Santos (2011), a pesquisa em TA no contexto nacional referente às suas aplicações e usos como ferramenta de suporte à tradução humana, ainda é muito voltada para os estudiosos da área de computação, ao passo que o próprio campo dos ETs ainda parecem se comportar um tanto reticentes a possíveis diálogos sobre os variados usos da TA enquanto ferramenta.

Além do mais, Baker (1995) propõe uma tipologia de corpora para a tradução, a qual na tentativa de se estabelecer uma metodologia, visa não só os ETs, mas como também o ensino da tradução. Contexto, esse, que poderia viabilizar a inserção de ferramentas tecnológicas na formação do tradutor como os programas de TA, por exemplo. Assim, em virtude da chamada era digital, a globalização impele uma extrema necessidade do manuseio hábil dessas ferramentas nos mais variados setores da sociedade digital entre diversos profissionais, inclusive o tradutor (AUSTERMÜHL, 2003).

Entretanto, crê-se que ainda há muito por se fazer no que diz respeito ao uso de corpora como metodologia de pesquisa nos Estudos em TA. Contudo, é fato a existência de uma conscientização, ainda que lenta e gradual, de suas aplicações nos ETs, como pode ser constado a partir da vasta gama de corpora disponíveis on-line por alguns estudiosos desta área quanto a esta questão (TAGNIN, 2011).

Embora ainda haja muita resistência, tanto por parte dos pesquisadores em tradução, quanto dos linguistas de corpus sobre a dimensão do papel que, ambas as áreas, podem

[5] À margem da tradução humana, em virtude de seu processo tradutório de cunho matemático e escopo ainda limitado à tradução de caráter mais técnico (manuais de instrução, informações de sites dentre outros).

exercer uma sobre a outra bem como nas diversas esferas sociais, é possível se observar aqui uma caminhada a um maior estreitamento entre suas diferenças, concomitantemente influenciando o avanço dos corpora nas pesquisas em TA, tendo em vista suas aplicações de caráter social (KOHEN, 2010; PYM, 2011; SANTOS, 2012), que no pensamento de Kohen (op.cit, p.20) remetem diretamente as necessidades do cotidiano das sociedades digitais:

> O uso da tradução automática pode ser dividido amplamente em três categorias: (a) assimilação, a tradução de material estrangeiro com a b-finalidade de entender o conteúdo, (b) a divulgação do texto, sua tradução para publicação em outros idiomas, e (c) a comunicação , tais como a tradução de e-mails, bate papos dentre outros. Cada uma das utilizações exige velocidade e qualidade diferentes. (KOHEN, 2010, p.20) [6].

Portanto, o pesquisador em TA ao lançar mão dos corpora como metodologia deve inicialmente considerar as categorias de aplicações supracitadas, não descartando o fato de que o escopo da TA apesar de apresentar uma gama de limitações, ele está em constante desenvolvimento (SANTOS, 2011), não podendo ser tomado como um escopo pronto e acabado, mas dinâmico e com fins práticos e crescentes nas sociedades digitais.

4. Considerações Finais

Ao longo deste ensaio buscou-se discutir questões de ordem teórica sobre a relevância dos corpora na pesquisa em TA. Percebeu-se que no decorrer no desenvolvimento da TA enquanto ramo dos ETs, a linguística de corpus servia como base de sua arquitetura no chamado processamento de línguas naturais. Aqui, os corpora inicialmente contemplavam um pequeno repertório de combinações lexicais e de sentenças entre L_1 e L_2 na busca por equivalentes de cunho morfológico, sintático e até semântico. Na atualidade, todavia, esse possível alinhamento tomou proporções gigantescas com o advento da internet e a criação da TA de natureza estatística, usando os textos disponíveis em rede como repertório de busca de suas traduções.

Face ao exposto, tendo em vista o comportamento inconstante do desenvolvimento da TA, os corpora podem ainda representar um papel de suma relevância em suas pesquisas, uma

[6] *The use of machine translation may be broken up broadly into three categories:(a) assimilation, the translation of foreign material for the purpose of understanding the content; (b) dissemination, translating text for publication in other languages; and (c) communication, such as the translation of emails, chat room discussions and so on. Each of the uses requires a different speed and quality*. Tradução automática revisão minha.

vez que também apresentam um desempenho semelhante no que diz respeito seu contínuo desenvolvimento e aplicações em diversas áreas, de modo que a utilização dos corpora como metodologia pode fornecer ao pesquisador uma visão mais abrangente e ao mesmo tempo mais acurada do objeto de estudo.

Consequentemente, lançar mão dos corpora como ferramenta e ou metodologia de pesquisa, possivelmente, possibilitará ao pesquisador e ou profissional dos ETs e TA mais habilidade em lidar com as facilidades computacionais para o processamento textual, que frequentemente, emergem no contexto das sociedades digitais.

5. Referências

ALFARO, C & M.C.P. DIAS. Tradução Automática: uma ferramenta de auxílio ao tradutor. UFSC, Cadernos de tradução, volº01, nº03, 1998.

AUSTERMÜL, Frank. *Electronic Tools for Translators*. UK, St. Jerome, 2003.

BERBER SARDINHA, Tony. Corpora eletrônicos na pesquisa em tradução. UFSC, Cadernos de Tradução, Vol. 1, No 9, 2002.

_____.(2000b) Semantic prosodies in English and Portuguese: a contrastive study. In P. Cantos Gómez & A. Sánchez Pérez (orgs.) *Cuadernosde Filología Inglesa: 9, 1. Corpus-based Research in English Language andLinguistics.* Murcia: Universidad de Murcia, pp. 93-109.

_____. Padrões lexicais e colocações do português. Apresentação no *XV ENPULI*, São Paulo, USP,1999.

BAKER, Mona. *In other words: a courserbook on translation*. Canada, Routledge, 2011.

_____. Corpora in Translation Studies: an overview for some suggestions and future research. In: *Target*. p.223-243. UMIST & Middlesex University, Amsterdan, 1995.

_____. Corpus linguistics and Translation Studies:Implications and applications. In: Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and Technology*: In Honour of John Sinclair. Amsterdam and Philadelphia: John Benjamins, 233-250

CASELI, Helena de Medeiros & NUNES, Israel Aono. Tradução automática estatística baseada em frases e fatorada: experimentos com os idiomas Português do Brasil e Inglês usando o toolkit Moses. São Paulo: USP, 2009.

CASELI, Helena de Medeiros. Portal de tradução automática: recursos e ferramentas para o português do Brasil. São Carlos, Universidade Federal de São Carlos (UFSCAR). Centro de Ciências Exatas e de Tecnologia, 2010. (Em andamento).

FERNANDES, L.P & SANTOS, Cleydstone Chaves. Da antiguidade à era informatizada: um breve percurso histórico da tradução no ensino de línguas estrangeiras. In: Leia Escola, EDUFCG, vol.11, n.02, 2011.

GERBER, Laurie. Machine translation: ingredients for productive and stable MT deployments - Part 2. USA, Association for Machine Translation in the Americas president, International Association for Machine Translation, 2009.

GUIDÉRE, Mathieu. *Introduction à la traductologie*. De Boeck, Belgique, 2010.

GOUTTE, Cyril et al. *Learning machine translation*. Massachussets, MIT press, 2009.

HATIM, Basin. Discourse analysis and translation. In: BAKER, Mona. *The routledge encyclopedia of translation studies*. Routdlege, London and New York. 2001.

HUTCHINS, W.J. and Somers, H.L. An introduction to machine translation.London, Academic Press Limited, 1992.
HUTCHINS, W, John. Machine translation: a brief history. In: KOERNER, E.F.K. and ASHER, R.E. *Concise history of the language sciences: from the Sumerians to the cognitivists*. Oxford, Oxford: Pergamon Press, 1995. Pages 431-445.
_____. Retrospect and prospect in computer-based translation: proceedings from the Singapore MT Summit, University of East Aglia, Singapore, 1999.
_____. Machine translation *In: CLASSE, Olive. Encyclopedia of literary translation into English*. London, Fitzroy Dearborn Publishers, University of East Anglia, 2000, p. 884-885.
HOLMES.J, S. The name and the nature of translation studies. In: VENUTI, Lawrence. *The translation studies reader*. Routledge, London and New York, 2000.
HUTCHINS, W. Jonh. Machine translation. In: *Encyclopedia of literary translation into English*. London: Fitzroy Dearborn Publishers, 2000. pp. 884-885.
KIM, Mira. Using systemic functional text analysis for translator education: na illustration with a focus on textual meaning. In: *The interpreter and Translator trainer*, vol. 01 p-223-246, 2007.
JOHNSON, Colin. Hybrid Systems Offer Smarter Machine Translation Among Languages. USA, Apptek Technology, McLean, Va, 2012.
KOEHN, Phillip. *Statistical machine translation*. Cambridge, Cambridge university press, 2010.
LAVIOSA-BRAITHWAITE, Sara.Universals of Translation . In: BAKER, Mona. *Routledge Encyclopedia of Translation Studies*.London & New York, 2001.
OLOHAN, Maeve. *Introducing corpora in translation Studies*. London/ New York. Routledge, 2004.
PAGANO, Adriana. Organização temática e tradução. In: PAGANO, A. MAGALHÃES,C. ALVES, F. *Competência em Tradução: Cognição e discurso*. BH, UFMG, 2005.
SANTOS, Cleydstone Chaves. Por uma estética na tradução automática: um estudo baseado em corpora eletrônicos. UFPB-CCHLA-DLEM-II ENCULT: João Pessoa, 2011.
_____. Um panorama do fluxo de recepção da tradução automática no cenário nacional. In-Traduções, Santa Catarina , Florianópolis, UFSC, vol. 05, nº01, p.167-176, 2011.
_____A tradução automática de gêneros textuais na esfera acadêmica. III SINALGE, UEPB, Campina Grande, 2012.
SELINKER, L. Language transfer [J]. In *General Linguistics.* Vol.9, 1969 (2):67-92.
SMITH, Ross. Machine translation: potential for progress. English Today 68, vol.17,nº04, Cambridge University Press, UK, 2001.
_____, GASPARI,F and NIÑO,A. Detecting inappropriate use of free online machine translation *by language students - a special case of plagiarism detection*.UK, University of Manchester, 2006.
SOMERS, H. Three perspectives on MT in the classroom. England, UMIST, 2001.
STUBBS, M. Collocations and semantic profiles: on the cause of trouble with quantitative studies. *Functions of Language*, Amsterdam, John Benjamins, 1995.
VIANA, Vandez. & TAGNIN, Stella E. O. *Corpora no ensino de línguas estrangeiras*. São Paulo, HUB, 2010.
TORO, C.G. Translation studies: an overview. In: Cadernos de Tradução. Florianópolis, nº20, vol.02, p. 09-42, 2007.

VASCONCELLOS, Mª.L. The fuzzy place of linguistics in the translations studies (TS).PhD Dissertation chapter, UFSC, 2009.
_____. & PAGANO, Adriana. Explorando interfaces: estudo da tradução, linguística sistêmico-funcional e linguística de corpus. In: PAGANO, A. MAGALHÃES,C. ALVES, F. *Competência em Tradução: Cognição e discurso*. BH, UFMG, 2005.
WILKS, Yorick. *Machine translation: its scope and limits*. U.K, Springer, 2009.
XATARA, C. A coesão lexical no conto Le Horla e em sua tradução. In: Cadernos de Tradução. Vol. 01, set, 2008.
ZANETIN, Federico et al. *Corpora in Translator Education.* UK, St. Jerome, 2003.