TRADUÇÃO AUTOMÁTICA

Ronaldo Teixeira Martins*

Resumo: O presente ensaio procura recensear e delimitar o domínio da tradução automática (TA): sua história, seus objetivos, seus métodos e seus problemas. Defende-se que a TA, a despeito de seus resultados, ainda muito precários, constitui território singular para exploração das interações entre modelos linguísticos, cognitivos e computacionais da linguagem.

Palavras-chave: Tradução automática (TA); processamento automático das línguas naturais; linguística computacional.

INTRODUÇÃO

O que se convencionou chamar "tradução automática" (TA) compreende hoje um espectro bastante diversificado de práticas de pesquisa e de projetos de desenvolvimento, muitos dos quais se afastaram definitivamente do programa original do domínio. Os sistemas originalmente projetados para a automação completa do processo tradutório foram redefinidos, progressivamente, como sistemas de tradução assistida por humanos e, de forma ainda menos ambiciosa, como tradução humana assistida por máquinas. Os desenvolvedores recuaram (estrategicamente?) da tentativa de produzir tradutores completamente automáticos e passaram a operar dentro de fronteiras bem mais definidas e com objetivos muito mais modestos. Em lugar de tentar produzir uma equivalência, na língua de chegada, para o enunciado originalmente produzido na língua de partida, os sistemas contemporâneos pautam-se,

* Doutor em Linguística pela Universidade Estadual de Campinas (Unicamp) e professor assistente da Universidade Presbiteriana Mackenzie (UPM).

especialmente, por estratégias alternativas, que se organizam em torno de quatro eixos principais, não necessariamente autoexcludentes:

1. pela estandardização – e, eventualmente, mesmo pidginização – da língua de partida, que passa a sofrer um processo de decomplexificação, com controle terminológico e simplificação sintática, de que resultaria uma versão mais regularizada (*controlled language*) da variedade original, que é naturalmente muito mais desorganizada e poluída de toda sorte de fenômenos de difícil tratamento computacional;
2. pela especialização do universo do discurso (*domain constraint*), com a seleção de uma temática, de uma forma de composição e de um estilo que, sendo caracterizados por um conjunto limitado de variações, constituiria apenas parte (*sublanguage*) da língua original, cujos componentes deveriam ser exaustivamente descritos, e para cujo reconhecimento e processamento a ferramenta de tradução deveria ser treinada;
3. pelo abandono sistemático da expectativa de produzir resultados definitivos, preteridos por traduções cruas (*raw translation*), que resgatam apenas os movimentos mais mecânicos – e menos inteligentes – do processo de tradução, e que serviriam de ponto de partida para o trabalho do especialista humano, que seria liberado das tarefas mais repetitivas;
4. pela redução da TA a uma tarefa de natureza puramente indicativa, conformada na tradução grosseira (*rough translation*), mero mecanismo de triagem, que sinalizaria, para o leigo, se o texto atende ou não a um determinado interesse de busca e, portanto, se deveria ser encaminhado ou não para tradução humana.

No que se segue, apresentam-se, em linhas gerais, os principais acidentes de percurso que conduziram a essa reorganização das práticas que se tornaram constitutivas da TA.

Breve histórico[1]

A TA tem seu registro oficial de nascimento normalmente situado em 1949, embora a possibilidade de automação da tradução humana viesse sendo discutida desde o século XVII[2]. O célebre *memorandum* do matemático Warren Weaver, parceiro de Claude Shannon no desenvolvimento da teoria da informação, é considerado o marco inicial da investigação no domínio, menos talvez por seus méritos técnicos do que pela força política, já que Weaver ocupava, à época, a direção da Divisão de Ciências Naturais da Fundação Rockefeller e gozava de boa reputação junto às agências americanas de fomento à pesquisa e ao desenvolvimento.

1 Reportam-se, aqui, especialmente, os trabalhos de Hutchins (1986, 1994, 1997, 1998, 1999), Mateus (1995), Hutchins e Sommers (1992) e Slocum (1985).

2 Francis Lodwick, que publicou, em 1647, *A Common Writing*, e, em 1652, *The Groundwork or Foundation Laid (or so Intended) for the Framing of a New Perfect Language and a Universal or Common Writing*, precursor da semântica de casos, é considerado o primeiro autor a publicar um tratado sobre uma língua fundada sobre uma escrita universal para fins de matematização da tradução (cf. ECO, 1994).

O memorando de Weaver, intitulado *Translation*, trazia quatro propostas. A primeira delas dizia respeito ao problema da desambiguização do sentido lexical, que, segundo o autor, poderia ser feita pelo processamento do cotexto imediato. A segunda proposta remetia ao teorema provado por McCulloch e Pitt (1943), para os quais um robô construído com iterações regenerativas de caráter formal seria capaz de deduzir qualquer conclusão legítima de um conjunto finito de premissas. Weaver postulava que, na medida em que a língua escrita fosse uma expressão de caráter lógico, o problema da tradução poderia ser formalmente equacionado. A terceira proposta concernia à possível aplicação de métodos criptográficos para tradução. As técnicas de criptografia, de que Weaver era profundo conhecedor, haviam tido grande desenvolvimento durante a Segunda Guerra Mundial, e o autor estava particularmente impressionado com o sucesso da criptografia baseada na frequência e nas combinações de letras. Weaver referia-se particularmente às invariâncias estatísticas observáveis entre as línguas naturais, considerando a língua tão somente como um sistema formal de codificação de informações, não muito diferente, portanto, dos outros códigos com os quais já havia trabalhado. Em sua quarta e última proposta, Weaver afirmava que, assim como deveria haver atributos lógicos comuns a todas as línguas, deveria haver universais linguísticos, que deveriam ser buscados para que se pudessem desenvolver sistemas de TA.

O texto de Weaver, de natureza programática, demarcou o território da TA, e passou a operar como referência a partir da qual se constituiu todo o domínio[3]. Em 1950, apenas alguns meses depois da divulgação do memorando, Erwin Reifler, sinologista vinculado à Universidade de Washington, que já vinha investigando invariâncias semânticas entre o inglês e o chinês, defendeu o uso de sistemas de tradução palavra-por-palavra. Para aprimorar os potencialmente precários resultados produzidos por sistemas dessa natureza, Reifler propunha a pré-edição do texto a ser traduzido e a pós-edição do texto já traduzido, e sugeria a hiper-regularização da língua de partida, ideia precursora do conceito de línguas controladas.

No mesmo ano, Abraham Kaplan, que trabalhava na Rand Corporation, explorava as invariâncias estatísticas, indicadas por Weaver, e chegava à conclusão de que não haveria necessidade de mais de duas palavras à esquerda ou à direita para que a desambiguização pudesse ser levada a cabo. No ano seguinte, Victor Oswald e Stuart Fletcher, da Universidade da Califórnia em Los Angeles, debruçavam-se sobre métodos de análise sintática automática para o alemão. Ainda em 1951, o lógico Yehoshua Bar-Hillel, originalmente vinculado à Universidade Hebraica de Jerusalém, seria indicado para um cargo de dois anos no Research Laboratory for Electronics, do MIT, especialmente para investigar as possibilidades da TA e planejar a pesquisa futura no domínio, tornando-se o primeiro pesquisador com dedicação integral ao tema.

Como produto dessas atividades, realizou-se, no MIT, em 1952, o primeiro congresso sobre problemas de TA, que contou com dezoito participantes. Para a preparação desse encontro, Bar-Hillel produziu o primeiro relatório sobre o estado da arte da TA, produto de suas visitas às (poucas) instituições em que projetos dessa natureza vinham sendo desenvolvidos.

3 Interessantemente, como aponta Hutchins (1997), a contribuição de Weaver ficaria restrita ao memorando e ao prefácio de um livro, publicado em 1955.

No relatório, Bar-Hillel reconhecia que os sistemas de TA poderiam ser desenhados para cumprir diferentes expectativas: tradução de alta qualidade nos domínios da ciência, das finanças e da diplomacia, por exemplo; e tradução rápida, talvez de baixa exatidão, para a varredura de textos em jornais, revistas e panfletos. Bar-Hillel discutia o precário equilíbrio entre exatidão e velocidade, e a pertinência da interação homem-máquina nos sistemas de automação. Reconhecia que o principal obstáculo para a TA seriam as ambiguidades semânticas, mas enfatizava que a tarefa seria possível, desde que as expectativas não fossem muito elevadas. Na linha inaugurada por Reifler, propunha um modelo híbrido de TA, com pré-edição e, especialmente, pós-edição. Em seu laudo, antecipava também sistemas de TA com vocabulário e sintaxe restritos, e afirmava que sistemas dessa natureza seriam mais realistas. Manifestava certo ceticismo sobre o alcance das abordagens estatísticas (para a desambiguização léxica e sintática), e chegaria a desenvolver, poucos meses mais tarde, um sistema de análise próprio, a gramática categorial, baseada no trabalho de Rudolf Carnap e Kazimierz Ajdukiewicz. Em sua concepção, seriam necessárias gramáticas universais, ou pelo menos gerais, para que se pudessem produzir sistemas de TA que não traduzissem apenas entre pares predeterminados de línguas.

O relatório de Bar-Hillel é particularmente emblemático porque antecipa muitos dos desdobramentos posteriores da tecnologia de desenvolvimento de sistemas de TA. Além da perspectiva de uso de sublínguas e línguas controladas, Bar-Hillel explora teoricamente a possibilidade de sistemas que incorporam a análise gramatical como uma de suas componentes. Antevê, por exemplo, a estrutura dos sistemas de tradução indireta baseados em transferência, já que – segundo ele – os sistemas de tradução deveriam ser constituídos por três elementos: análise morfológica, análise sintática e transformação sintática. O pesquisador acreditava que as dificuldades associadas à primeira componente seriam relativas à construção de dicionários e à resolução dos problemas de ambiguidade categorial, que poderiam ser diminuídos, mas não eliminados, pelo recurso a pré-editores, como os propostos por Reifler. O segundo estágio requereria o desenvolvimento daquilo que era por ele referido como "sintaxe operacional", e que poderia estar amparado na construção de "gramáticas de transferência", em que a gramática de uma língua seria definida a partir das categorias apropriadas a uma outra. Mencionava-se o trabalho, ainda não publicado, de Zellig Harris.

Durante a conferência, várias outras propostas foram discutidas e analisadas. Particular relevo alcançaram as tentativas de minimização das ambiguidades e da complexidade sintática dos textos de partida, seja por meio da pré-edição, seja por meio do treinamento dos usuários do sistema, seja por meio da utilização de línguas controladas. O esforço irrecusável de pidgnização das línguas de partida denunciava – já naquele momento – que a principal dificuldade no desenvolvimento de sistemas de TA concernia à análise mais do que à geração. O controle terminológico e a simplificação das construções sintáticas propostos para o inglês, embora muitas vezes pudessem incorrer em propostas bastante heterodoxas (como a de uma nova ortografia para as línguas, que explicitasse a classe gramatical a que pertenciam as palavras), representavam, em estado embrionário, muitas das perspectivas que seriam posteriormente (e ainda hoje) perseguidas. Esse é o caso, por exemplo, das propostas de análise

interativa (Reifler); da utilização de uma língua-pivô, idealmente o próprio inglês, em sistemas multilíngües (formulada por Leon Dostert); e a de construção de dicionários terminológicos (então chamados "microglossários"), estabelecidos a partir da frequência de ocorrência das palavras em um domínio específico (proposta por Victor Oswald).

A partir desse impulso inicial, a exploração no domínio avançou rapidamente. Muitos outros pesquisadores aderiram à causa, ampliou-se o financiamento à pesquisa no campo, espalhou-se o espectro de temas e problemas, publicaram-se as primeiras matérias em jornais de larga circulação, e reportaram-se várias experiências de simulação (manual) de sistemas de tradução. A primeira delas – do russo para o inglês –, realizada ainda em 1952, teria produzido o seguinte resultado:

> *On/Onto/At Fig.12 traced/mapped-out/drawn parabola according-to/along/in-accord-with which move thrown/deserted with/from velocity 10m/sec. under/below angle to/toward vertical line into/in/at 15§, 30§, 45§, 60§* (PERRY 1952 apud HUTCHINS, 1997).

A inteligibilidade do texto "traduzido" – e, consequentemente, a viabilidade da proposta – teria sido confirmada pela capacidade de falantes monoglotas conseguirem produzir, para a tradução tosca fornecida acima, a seguinte correspondente em inglês:

> *On Fig. 12 a parabola is drawn according to which a body moves, thrown with the velocity of 10m/sec and making angles of 15§, 30§, 45§, 60§ with the vertical line* (PERRY 1952 apud HUTCHINS, 1997).

A primeira demonstração real de um sistema de TA – que alcançaria as primeiras páginas de alguns dos principais jornais norte-americanos – ocorreria em 1954, na sede da IBM em Nova York, a partir de um sistema desenvolvido por Paul L. Garvin, da Universidade de Georgetown, e Peter Sheridan, da própria IBM, sob a supervisão de Leon Dostert. Tratava-se de um sistema de pequena escala, limitado a um vocabulário de 250 palavras e seis regras gramaticais, para traduzir sentenças do russo para o inglês.

A euforia provocada por esse primeiro experimento, que teria estimulado expectativas que não se cumpriram e que não poderiam ser cumpridas em curto prazo, é frequentemente associada à ruptura que viria afetar posteriormente todo o domínio. No entanto, o experimento contribuiu para que a TA se afirmasse como campo de investigação autônomo, para que se disseminasse entre a comunidade de pesquisadores norte-americanos, e para que se tornasse um campo razoavelmente popular, passando a frequentar, com alguma regularidade, a grande imprensa[4].

Ainda em 1954, publicou-se o primeiro número da primeira revista especializada (*Mechanical Translation*), dirigida por William N. Locke e Victor H. Yngve, e,

4 É interessante observar que as "máquinas de traduzir" foram também prejudicadas por esse processo de popularização, na medida em que passaram a integrar o anedotário da imprensa, por meio de referências, quase sempre apócrifas, aos resultados produzidos pelos sistemas. É célebre, por exemplo, a geração de "*The whisky is strong, but the meat is rotten*", supostamente derivada da retrotradução, de volta para o inglês, da sentença bíblica "*The spirit is willing, but the flesh is weak*", traduzida para o russo. O mesmo vale para "*Invisible idiot*", que teria sido produzida a partir de "*Out of sight, out of mind*". Ainda que os provérbios venham de fato oferecer problemas para a TA, os dois casos citados, segundo Hutchins (1995), nunca foram realmente produzidos por nenhum sistema de tradução conhecido.

no ano seguinte, editou-se a primeira coletânea sobre o assunto (*Machine Translation of Languages: Fourteen Essays*), organizada por Locke e Andrew D. Booth. Também em 1955 ocorreu a publicação de resultados desenvolvidos na então União Soviética e, em 1956, realizou-se, novamente no MIT, o I Congresso Internacional, de que participaram pesquisadores ingleses, canadenses e russos.

Na década de 1956 a 1966, o movimento espalhou-se e aprofundou-se consideravelmente. Em um congresso realizado em Moscou, em 1957, foram apresentadas 56 comunicações sobre a TA para línguas tão variadas quanto o alemão, o inglês, o magiar, o chinês e o francês, além obviamente do russo. Em 1961, em Teddington, na Inglaterra, um novo congresso internacional sobre Tradução Automática das Línguas e Análise Linguística Aplicada reuniu 170 delegados vindos de quinze países. Criou-se a Association por l'Étude et le Développement de la Traduction Automatique et de la Linguistique Apliquée (Atala), e publicaram-se vários outros trabalhos sobre o tema. A TA transformava-se no principal campo da computação não numérica e um "*multimillion dollar affair*", como a ela se referiria, mais tarde, Bar-Hillel.

A disseminação do campo trouxe à luz um conjunto progressivamente maior de divergências teóricas e metodológicas, que convergiram, no entanto, para a exploração sistemática da linguística, especialmente no sentido de torná-la uma ciência exata, com a utilização de métodos da matemática. Segundo o Relatório Alpac (1966), essa teria sido a principal contribuição teórica da TA, embora, na verdade, muito do desenvolvimento de abordagens formais em sintaxe e em semântica não estivesse realmente relacionado ao domínio. Em todo caso, desde cedo se tornou patente que a concepção de tradução como decriptação não poderia ser mantida. Embora pudessem ser efetivamente concebidas como sistemas formais de codificação de informações, as línguas naturais, especialmente em razão da ambiguidade, que se revelava nos mais variados níveis de análise, não podiam ser comparadas a sistemas artificiais em que a relação entre significante e significado era estabelecida de forma biunívoca. Buscaram-se outras invariâncias interlinguísticas (semânticas, lógicas), propuseram-se outros modelos de tradução (baseados no mapeamento sintático, por exemplo), restringiu-se o processo de automação (com a previsão de sistemas interativos), desenvolveram-se estudos de lexicografia e de sintaxe formal, construíram-se dicionários semasiológicos, mas não se pôde furtar ao óbvio: que a empresa era mais difícil do que a princípio tinha parecido, e que o grau de complexidade da linguagem humana talvez ultrapassasse o limite do razoável (ou do possível).

A primeira crítica sistemática às iniciativas então empreendidas veio de Bar-Hillel (1960), entusiasta de primeira hora, a quem fora encomendado, em 1958, pelo US Office of Naval Research, um novo relatório sobre o estado da arte da TA. Bar-Hillel condenava o que passou a chamar *Fully-Automatic High-Quality Translation* (FAHQT), objetivo da maior parte dos projetos então em desenvolvimento, que considerava um equívoco e um desperdício de dinheiro. Para Bar-Hillel, uma tradução completamente automática e de alta qualidade somente seria possível se fosse incorporada ao modelo uma enciclopédia que contivesse todo o conhecimento humano disponível, o que seria evidentemente impossível. A pesquisa em TA deveria retornar, portanto, ao desenvolvimento de sistemas parciais, que pudessem auxiliar o tradutor humano, e que pudessem ser gradualmente refinados com operações de revisão (pós-edição) automática.

O argumento estava baseado na discussão de uma sentença do inglês ("*The box was in the pen*") cuja ambiguidade semântica, provocada pela ambiguidade lexical de "*pen*", somente poderia ser resolvida por recurso ao conhecimento de mundo do falante[5]. O argumento provaria que a completa resolução das ambiguidades linguísticas requereria a representação de um conhecimento que, em última análise, não seria (apenas) linguístico. Na medida em que esse conhecimento seria irrepresentável para a máquina, a FAHQT não poderia ser atingida nem mesmo em um remoto futuro[6].

Bar-Hillel criticava também as abordagens que, a partir da análise estatística de grandes *corpora*, negligenciavam o conhecimento linguístico, que seria, segundo ele, muito mais econômico. E condenava os sistemas de tradução baseados no desenvolvimento de uma interlíngua universal ou independente de outras línguas, na medida em que não haveria nenhum argumento razoável para acreditar que a tradução de um enunciado para uma interlíngua lógica seria mais simples do que a tradução para uma outra língua natural.

As críticas de Bar-Hillel representam um ponto de inflexão na trajetória da investigação em TA, especialmente nos Estados Unidos. A despeito dos esforços e dos investimentos, os resultados práticos eram pífios, e de muito pouco serviam. Estabelecia-se, pouco a pouco, o consenso de que os recursos disponíveis, fossem linguísticos (como dicionários e gramáticas), fossem computacionais (como memória e processadores), eram não apenas insuficientes, mas inadequados para prover ao tipo de demanda criado pelo processamento automático das línguas naturais.

Em 1966, o Relatório Alpac, do Comitê Assessor de Processamento Automático das Línguas Naturais, da Academia de Ciências dos Estados Unidos, encarregado da análise dos resultados dos (muitos) programas subsidiados pelo governo norte-americano, provocou profundo impacto na comunidade de pes-

5 A forma do inglês "*pen*", como substantivo, possui pelo menos dois valores primitivos: 1) "*an instrument made of plastic or metal used for writing with ink*"; e 2) "*a small piece of land surrounded by a fence in which farm animals are kept*". A esses valores se associam, em cada caso, inúmeras acepções derivadas. No contexto "*The box was in the pen*", Bar-Hillel se referia, principalmente, ao conceito de "*playpen*", ou seja, "*a small enclosure, usually portable, in which a young child can play safely alone without constant supervision*".

6 Um poderoso argumento a favor da crítica de Bar-Hillel concerne ao fato de que, mais de quarenta anos depois de seu relatório, a tradução de "*The box was in the pen*" não foi ainda equacionada pelos sistemas de TA disponíveis. Considere-se, a título de exemplo, os resultados da tradução do inglês para o português indicados abaixo:

Sistema	**Resultado**
Amikai	A caixa estava na caneta.
BabelFish	A caixa estava na pena.
FreeTranslation	A caixa estava na caneta
Google	A caixa estava na pena.
Intertran	A caixa era na caneta
Linguatec e-translation server	A caixa estava na caneta.
Systran	A caixa estava na pena.
T-Mail	A caixa estava na pena.
WorldLingo	A caixa estava na pena.

Embora muitas das ferramentas descritas façam uso dos mesmos sistemas de TA, em diferentes versões e com diferentes funcionalidades, é forçoso reconhecer que, pelo menos nesses casos, e em que pese a simplicidade morfossintática de "*The box was in the pen*", nenhuma das traduções propostas poderia ser aceita como válida por um falante do português, excetuado, evidentemente, o contexto fabuloso, mas absolutamente improvável, em que caixas podem estar dentro de penas e canetas.

quisadores e desenvolvedores em TA. Todas as iniciativas até então consumadas foram vigorosamente criticadas, e denunciados o seu espontaneísmo, a sua precariedade teórica e a falta de conhecimento e tecnologia necessários para a execução das propostas. Especialmente: criticava-se a real utilidade da TA.

O relatório era relativamente sucinto, mas vinha complementado por vários anexos. Analisava as reais necessidades de tradução dos órgãos que vinham financiando os projetos de desenvolvimento de sistemas de TA, para concluir que a demanda não era tão expressiva a ponto de justificar os investimentos, da ordem de vinte milhões de dólares, que haviam sido empregados nos dez anos anteriores. Além disso, observava que não haveria, àquela época, nenhum sistema de TA que dispensasse a pós-edição (revisão) humana dos resultados; e a revisão humana seria não apenas mais demorada, mas mais cara (*sic*) do que a tradução humana convencional[7]. O documento terminaria por sugerir que os esforços fossem concentrados no desenvolvimento de ferramentas para tradutores (o que era então referido como "*machine-aided translation*").

Como resultado, o investimento público no setor refluiu consideravelmente por pelo menos vinte anos. Segundo Slocum (1985), em 1973 havia apenas três projetos de TA subsidiados pelo governo dos Estados Unidos. Em 1975, nenhum projeto teria sido contemplado. As iniciativas, pelo menos na América do Norte, ficaram bastante circunscritas a experiências isoladas (como o desenvolvimento do sistema Systran, por Peter Toma). Na Europa, o declínio do campo teria sido menos expressivo.

No início dos anos 1980, com o desenvolvimento tecnológico, o domínio da TA voltou a receber maior atenção dos pesquisadores, especialmente na Europa e no Japão, mas sob novas bases e com outra acepção. A primeira fase da TA havia provado que a automação do processo somente se tornaria viável se a complexidade da tarefa pudesse ser expressivamente reduzida. Para que pudesse sobreviver como objeto de investigação, a TA precisava redefinir seu escopo.

Escopo

O reiterado fracasso das tentativas fez que a ideia de uma tradução completamente automática de qualidade fosse definitivamente abandonada. Propuseram-se, como alternativas, a redução do grau de complexidade dos textos a serem traduzidos, a redução do grau de automação do processo, ou mesmo a completa transformação da tarefa.

No primeiro caso, propôs-se que as ferramentas de TA operassem não sobre quaisquer textos em língua natural, mas sobre textos específicos, que fizessem uso controlado da linguagem. Por "uso controlado" deve-se entender aqui uma de duas noções: 1. o controle derivado da seleção (natural) de textos de determinada forma e sobre determinado conteúdo; ou 2. o controle artificial derivado da imposição de uma forma padronizada para os textos a serem traduzidos. Em ambos os casos, restringe-se o escopo da TA, que passaria a operar apenas

7 A esse propósito, convém reportar o depoimento de um dos tradutores ouvidos pelo Comitê, anexado ao relatório: "*I found that I spent at least as much time in editing as if I had carried out the entire translation from the start. Even at that, I doubt if the edited translation reads as smoothly as one which I would have started from scratch. I drew the conclusion that the machine today translates from a foreign language to a form of broken English somewhat comparable to pidgin English. But it then remains for the reader to learn this patois in order to understand what the Russian actually wrote. Learning Russian would not be much more difficult*" (ALPAC, 1966).

sobre um subconjunto (higienizado) da língua natural, normalmente chamado "sublíngua" ou "língua controlada".

O conceito de sublíngua aposta na ideia de que a especialização da forma e do conteúdo dos textos tratados é necessária e suficiente para a eliminação da ambiguidade. Está amparado na hipótese de que textos parentes (por afinidade temática, formal ou funcional) comportam uma série de invariâncias (de vocabulário, de estruturas sintáticas) que facilitariam o processo de análise. Nesse caso, quanto mais próximos e fixos (enlatados, padronizados) os textos, mais eficazes as traduções. É o que tem sido observado, com sucesso, no sistema Météo (CHEVALIER et al., 1978), que traduz boletins meteorológicos canadenses do inglês para o francês. Como a estrutura e o vocabulário dos textos são (naturalmente) muito restritos e repetitivos, o sucesso da automação do processo de tradução amplia-se consideravelmente.

É forçoso observar, no entanto, que essa delimitação implica o desenvolvimento de sistemas excessivamente especializados, de utilidade bastante localizada. Esse é o caso, aliás, do próprio Météo, cujas tentativas de extensão (para o domínio dos manuais de aviação, por exemplo) foram todas malsucedidas. Além disso, a especialização da forma e do conteúdo dos textos tratados não tem podido funcionar como panaceia para muitos dos problemas encontrados, já que o próprio traçado das fronteiras temáticas e formais tem estado frequentemente em discussão. Os gêneros textuais, por exemplo, têm revelado maior variabilidade interna do que o esperado: fracassaram todas as tentativas, na linguística do texto, de construção de "gramáticas textuais", ou seja, de formalização de conjuntos finitos de regras que estabelecessem, de forma inequívoca, os princípios de formação de todos e apenas dos textos pertencentes a um determinado gênero (como o jornalístico, por exemplo) ou a uma determinada tipologia (como a narrativa)[8].

Não se afirmará, evidentemente, que cada gênero textual não tenha suas próprias preferências vocabulares e sintáticas, cuja identificação poderia evitar muitos dos casos de ambiguidade presentes na língua do dia a dia. Mas têm sido reportados experimentos em que, apesar da restrição da forma e do conteúdo dos textos, a ambiguidade não vem sendo expressivamente reduzida. O Projeto Verbmobil (KAY et al., 1991), por exemplo, apesar de restringir o universo do discurso para o de uma conversa entre dois interlocutores sobre o lugar e o momento de um próximo encontro, tem deparado com inúmeras ambiguidades residuais, que apenas o contexto de enunciação poderia resolver.

A proposição de uma sublíngua artificial traz a vantagem de evitar a busca de invariâncias textuais e o desenvolvimento de sistemas excessivamente especializados, mas passa a exigir a pré-edição, por um especialista, do texto de partida. O controle artificial é provocado por uma de duas formas: pode ser de-

8 A respeito do fracasso programático da linguística textual em emular, no nível do texto, os mesmos procedimentos utilizados pela gramática gerativo-transformacional, ver Koch e Travaglia (1990, p. 57-58): "Com a evolução dos estudos percebeu-se [...] que não existe a seqüência lingüística incoerente em si e, portanto, não existe o não-texto. Se todos os textos são em princípio aceitáveis, não é possível uma gramática com regras que distinguem entre textos e não-textos. Por isso, passou-se à construção de uma Teoria do Texto ou Lingüística do Texto, que é constituída de princípios e/ou modelos cujo objetivo não é predizer a boa ou má-formação dos textos, mas permitir representar os processos e mecanismos de tratamento dos dados textuais que os usuários põem em ação quando buscam compreender e interpretar uma seqüência lingüística, estabelecendo o seu sentido e, portanto, calculando sua coerência".

rivado do estabelecimento de regras rígidas de redação dos textos, cuja forma passaria a ser induzida e padronizada; ou pode consistir na sinalização, nos textos, por meio de marcações, etiquetas e outras formas de anotação, de informações de natureza metalinguística, que pudessem reduzir seu nível de ambiguidade. No primeiro caso, os textos a serem traduzidos deveriam submeter-se a alguns protocolos de redação, como a restrição do vocabulário, a hiper-regularização sintática, o preenchimento de todas as elipses ou a explicitação de todas as relações anafóricas, por exemplo[9]. No segundo caso, o texto deveria ser previamente submetido a ferramentas de anotação (*taggers*), que explicitassem as categorias gramaticais envolvidas. Nos dois casos, restringe-se novamente o escopo de atuação da ferramenta, mas em outro sentido. Trata-se agora da tentativa, na linha inaugurada por Reifler em 1950, de combinar esforço humano e mecânico de forma a reduzir os custos do processo de tradução – o que viria a ser conhecido como TA auxiliada por humanos (*Human-aided machine translation*, ou HAMT). As técnicas de HAMT investiriam ainda na automação dos processos de tradução, mas prevendo a intervenção humana na edição do texto de partida (pré-edição), na tradução do texto de partida para o texto de chegada (interação) ou na edição do texto de chegada (pós-edição). Estaria abandonada a ideia de um processo de tradução completamente automático.

Por fim, uma última alternativa para a redução da complexidade do problema consistiria em alterar significativamente a definição do que seria "TA". Oferecem-se, nesse caso, dois caminhos: o de passar a conceber a TA como suporte ao tradutor humano, ou como processo de geração paralela de textos originais, em línguas diferentes, a partir de uma mesma representação da informação.

A primeira opção vem sendo chamada de tradução humana auxiliada por máquina (em inglês: *Machine-aided human translation*, ou MAHT, em oposição a HAMT). A opção por técnicas de MAHT é bastante pessimista: o processo de automação da tradução estaria agora circunscrito ao desenvolvimento de ferramentas de apoio ao tradutor humano, como dicionários bilíngues, corretores ortográficos e revisores gramaticais.

A segunda opção procura reduzir o domínio da tradução ao da comunicação multilíngue, admitindo que a ferramenta deveria tomar, como ponto de partida, não um texto em língua natural, mas um conjunto de informações a respeito da realidade, representado pelo falante de uma língua natural específica, por meio de sistemas de representação (de natureza dialogada, por exemplo) em que os dados seriam registrados de forma não necessariamente linguística. Admite-se que o objetivo da tradução seria recuperar o conteúdo informativo registrado em um determinado texto de partida, o que poderia ser facilitado se fosse adotado, em lugar da forma linguística (normalmente ambígua e indeterminada), um outro tipo de estruturação (e representação) dos dados, como o preenchimento, por exemplo, de campos com valores predeterminados em um formulário padronizado.

9 O Model English, inglês regularizado, elaborado por Stuard C. Dodd, assim como o Basic English, espécie de inglês simplificado, constituído de apenas 850 palavras, para comunicação internacional, proposto, na década de 1930, por Charles Ogden, teriam sido, nos primeiros anos da TA, alternativas (teóricas) frequentes entre os precursores do domínio que anteviam a possibilidade de línguas controladas.

Objetivos

Reconhece-se, de maneira geral, que a grande dificuldade dos sistemas de TA é justamente o processo de análise e interpretação dos enunciados em língua natural. Diferentemente do que ocorre em outras ferramentas computacionais, a TA é particularmente sensível à representação do conteúdo semântico das sentenças e dela profundamente dependente. Como diria Santos (1995, p. 128):

> *A operação de fazer a transição de uma língua para outra – consistindo afinal na tradução de itens lexicais da língua de partida para itens lexicais da língua de chegada – é a parte mais trivial de todo o processo, e o ônus da tarefa de traduzir (pelo menos se for encarada do ponto de vista computacional) recai sobre as competências monolíngües envolvidas.*

Acreditou-se, em princípio, que a especificação dos dicionários e das gramáticas das línguas naturais seria suficiente para o equacionamento da estrutura semântica das sentenças. Mas foi observado, desde cedo, que há muito mais do que simplesmente trocas lexicais e sintáticas no processo pelo qual um tradutor humano processa uma sentença. É o que demonstrou Bar-Hillel, por exemplo, no já referido relatório publicado em 1960, a propósito da sentença do inglês: "*The box was in the pen*".

Em defesa da máquina, pode-se dizer que (a) ela poderia prover todas as traduções possíveis para a sentença, de forma que o leitor poderia optar pela mais apropriada, o que, embora venha a constituir, em algum grau, certo constrangimento (e certo incômodo desnecessário), não provaria a inutilidade da TA; e (b) sentenças como a indicada não seriam exatamente frequentes e não representariam, do ponto de vista estatístico, problema para o funcionamento global de uma ferramenta de TA. Sobre esses dois pontos caberia dizer que não há evidência empírica ou científica a seu favor, mas antes na direção contrária; e que ambos constituem antes impressões do senso comum do que experiências efetivamente comprováveis de sucesso no processo de construção de ferramentas automáticas. Se não, vejamos.

O primeiro argumento está diretamente relacionado à ideia de que qualquer tradução é melhor do que nenhuma tradução. No entanto, a possibilidade de a máquina prover todas as traduções possíveis para as sentenças de entrada está associada a vários problemas. O mais grave talvez seja o fato de que não se traduzem, normalmente, sentenças apenas, mas textos inteiros. Na medida em que a mesma sentença de entrada pode corresponder a várias sentenças de saída, instala-se o risco da explosão combinatória, diretamente proporcional ao número e à complexidade das sentenças envolvidas. Outro problema concerne à cooperatividade do usuário: embora saibamos que o usuário está geralmente disposto a aceitar, em larga medida, falhas da ferramenta, ele o faz apenas quando percebe que, feitas as contas, o processo é simplificado. Nenhuma garantia nesse sentido pode ser dada por um sistema que exija a intervenção do usuário a cada caso de ambiguidade. Por fim, esse primeiro argumento está também relacionado a duas visões de tradução que caberia discutir: a chamada "tradução crua" (*raw translation*) e a chamada "tradução grosseira" ou "tradução rudimentar" (*rough translation*).

Tradução rudimentar ou grosseira seria aquela utilizada tão somente como instrumento para que o usuário possa tomar a decisão de requisitar ou não

uma tradução humana para o texto. Em última instância, seria apenas a indicação de uma palavra-chave, tema ou ideia geral, derivada de uma varredura do texto, sempre superficial e bastante imprecisa. Serviria, por exemplo, como instrumento para sistemas de busca de informações, mas jamais como objeto de publicação ou difusão. A tradução não tem aqui nenhum compromisso com a qualidade, e os resultados seriam tão problemáticos que um tradutor humano, por considerá-los de muito pouca validade, seguramente julgaria menos dispendioso (e mais rápido) retraduzir todo o texto a partir do nada do que tentar corrigir os problemas verificados. Esse tipo de tradução, bastante robusta, completamente automática, rápida e de baixo custo, não requereria (para a sua produção) nenhum tradutor ou revisor profissional, e poderia ser operada por um usuário comum, na medida em que não exigiria nenhuma espécie de pré-edição do texto de entrada (talvez apenas a confirmação das propostas do sistema) ou de pós-edição do texto traduzido (limitada, quando muito, a operações de formatação). Em compensação, seus resultados são de validade bastante controvertida. A maior parte dos sistemas de TA atualmente franqueados ao público (como Systran, Google Translator, Microsoft Translator etc.) incidiria nesse caso.

Tradução crua, ao contrário, seria aquela planejada para a pós-edição, ou seja, seria a produção de resultados propositadamente parciais (e imperfeitos) para que pudessem servir de ponto de partida para a correção (ou tradução) humana, feita por revisores (ou tradutores) especializados. Teria como mérito reduzir o tempo de tradução de um texto, na medida em que resolveria problemas básicos para o tradutor humano, deixando-o livre para se ocupar apenas de casos mais espinhosos (ou de difícil resolução por parte da máquina). À ferramenta caberia produzir, portanto, o primeiro rascunho (jamais o texto final), a ser trabalhado pelo especialista, para o qual seria observado um expressivo ganho de produtividade (em torno de 40% a 50% do tempo dedicado a cada lauda, segundo os registros reportados em Boitet (1995a)). Os sistemas dessa natureza envolveriam conhecimento mais especializado e interação mais intensa com o usuário, a quem também caberia, com frequência, a pré-edição do texto de partida. Em razão do treinamento exigido, dificilmente seriam úteis ao grande público. Quase todos os grandes sistemas corporativos (Duet, da Sharp; Hicat, da Hitachi; Atlas-II, da Fujitsu; Metal, da Siemens; AS-Transac, da Toshiba; Pivot, da NEC; etc.) funcionam dessa maneira.

A produção de traduções cruas e a de traduções grosseiras constituiriam, em última análise, duas estratégias de MAHT, ou seja, de tradução humana auxiliada por máquinas, já que o papel do tradutor humano, se considerado todo o processo, longe de subsidiário, seria fundamental. O resultado, em ambos os casos, fica normalmente muito aquém das expectativas de um usuário não familiarizado com a complexidade da tarefa. Com efeito, ambas as abordagens estão muito distantes da ideia de TA tal como concebida originalmente.

O segundo argumento a favor da máquina envolve uma impressão equivocada da linguagem. A ambiguidade não é um fenômeno periférico e marginal (e por isso rarefeito) nos enunciados em língua natural. Ela é constitutiva da própria linguagem, na medida em que todos os enunciados sofrem de vagueza e de indeterminação, se isolados os índices contextuais (relativos ao contexto extratextual) e co-textuais (relativos ao contexto intratextual) que provocam, com frequência, a ilusão de que os enunciados seriam exatos e precisos.

Há, evidentemente, marcas mais ostensivas dessa ambiguidade. É o caso da ambiguidade categorial e da ambiguidade léxica, por exemplo. Existem na língua portuguesa, efetivamente, formas que podem indicar mais de uma classe gramatical, ou que, mesmo indicando a mesma classe gramatical, podem comportar acepções diferentes. O dicionário Aurélio registra, por exemplo, doze classificações gramaticais diferentes para a forma "que", que não é exatamente incomum nos textos em língua portuguesa; e raras são as entradas, no mesmo dicionário, que comportam apenas uma acepção. As formas homônimas (como "banco") representam apenas a parte mais radical dessa possibilidade de variação. Mesmo formas aparentemente não ambíguas (como "abacaxi") comportam muitos sentidos diferentes ("a planta como um todo", "apenas a parte comestível da fruta"), especialmente se incorporadas as variedades regionais e sociais da língua (em que "abacaxi" pode figurar, por exemplo, ora como "coisa complicada", ou "pessoa desagradável", ou até mesmo "dançarino desajeitado").

A par da ambiguidade em nível lexical, a ambiguidade sintática é também facilmente perceptível. E, como no caso anterior, não é exatamente rara na língua portuguesa. A análise dos casos mais conhecidos talvez faça parecer que o fenômeno seja localizado a algumas construções específicas, como a adjunção ao verbo. No exemplo, muito explorado, de "A menina viu o menino com o telescópio", não se poderá precisar, fora de contexto, se o sintagma "com o telescópio" constitui um modificador do verbo "viu" ou do seu objeto, "o menino". Mas é preciso perceber que mesmo sentenças de estruturação supostamente mais clara contêm ambiguidades estruturais. A exatidão de "A menina chegou atrasada" desaparece, por exemplo, ao considerarmos que há pelo menos duas possibilidades diferentes de enquadramento do sintagma "atrasada": como adjunto a "menina" (com o qual, aliás, concorda), ou como adjunto a "chegou" (a quem parece modificar: "chegar atrasadamente"). A escolha por uma entre essas possibilidades, embora talvez irrelevante para o português, torna-se estratégica no processo de tradução para línguas que vão optar por meios bastante diferenciados de representação dos mesmos fenômenos.

De resto, a ambiguidade em língua natural se espalha em inúmeras outras direções, como na recuperação das relações anafóricas e no preenchimento das elipses, por exemplo. Mas a ambiguidade mais insidiosa é justamente aquela que não se revela na superfície do texto, e se reveste de suposta univocidade, a nos convencer de que a sentença admite uma única e exclusiva possibilidade de interpretação. Essa monovalência esconde o fato de que a univocidade dos enunciados linguísticos deriva muito mais de fatores contextuais (extralinguísticos, portanto, e em princípio irrepresentáveis para a máquina) do que de fatores propriamente linguísticos. Estaríamos de tal forma hipnotizados pelo contexto, ou nele inseridos, que seríamos incapazes de perceber o quanto há de ambiguidade nos enunciados cotidianos. A ambiguidade, em muitos casos, não se percebe senão retrospectivamente, pela necessidade de adaptação do enunciado a uma língua (ou a uma situação) em que outros níveis de análise acabam por se revelar necessários.

Método

O campo da TA não pode ser delimitado de forma única, e constitui antes uma dispersão, que varia conforme todo um conjunto de pressupostos sobre o

que seja a linguagem humana, sua natureza, sua estrutura, sobre o papel do conhecimento linguístico e do conhecimento de mundo na interpretação dos enunciados, e – talvez especialmente – sobre o que seja tradução. Os sistemas de TA variam ao sabor das premissas que os orientam, e há inúmeras formas de agrupá-los.

Os sistemas podem ser classificados, por exemplo, em relação ao número de línguas envolvidas: há sistemas de tradução bilíngues ou multilíngues. Esses sistemas podem ser, por sua vez, unidirecionais ou bidirecionais, na medida em que permitem (ou não) que a língua de chegada possa ser também a língua de partida. Outra perspectiva a partir da qual os sistemas podem ser analisados diz respeito ao paradigma utilizado. Há sistemas de tradução baseados em regras (sistemas simbólicos) e há sistemas de tradução baseados em casos (sistemas subsimbólicos). Há sistemas de tradução que utilizam apenas recursos linguísticos (dicionários e gramáticas) e há sistemas de tradução que incorporam também outras formas de conhecimento (bases de conhecimento, ontologias, *corpora* etc.). Os sistemas podem ser classificados também de acordo com o papel do usuário: há sistemas interativos (que requerem a ajuda ou a intervenção do usuário) e há sistemas não interativos. Por fim, todos os sistemas podem ser classificados segundo as estratégias utilizadas: a tradução direta ou indireta. O objetivo desta seção será recuperar algumas dessas variações e estratégias.

Dorr et al. (2000), por exemplo, salientam a existência de pelo menos três abordagens (que os autores chamam de "paradigmas") predominantes: a tradução baseada exclusivamente em conhecimento linguístico, ou seja, em dicionários e gramáticas (*Language-Based Machine Translation* – LBMT); a tradução baseada em conhecimento, ou seja, em dicionários, gramáticas e, adicionalmente, enciclopédias e bases de conhecimento (*Knowledge-Based Machine Translation* – KBMT); e a tradução baseada em exemplos, ou seja, em dicionários, gramáticas e *corpora (Example-Based Machine Translation* – EBMT*)*. Os dois primeiros casos constituiriam, especialmente, modelos de tradução baseada em regras, ou na elicitação do conhecimento linguístico inato do falante; o último seria particularmente amparado em análises e dados estatísticos.

O primeiro modelo, em razão do custo relativamente mais baixo se comparado aos demais, seria mais adequado para sistemas mais genéricos e mais robustos, mas produziria resultados menos satisfatórios e mais sujeitos a erro. Os dois outros, por envolverem o desenvolvimento de recursos mais dispendiosos (enciclopédias e *corpora* convenientemente anotados, separados por domínio do conhecimento), produziriam resultados mais exatos, mas seriam indicados apenas para sistemas mais especializados, de domínio restrito. Ambos os paradigmas pretendem enriquecer o ponto de partida do processo de tradução, aparelhando a máquina com conhecimento adicional para que as ambiguidades e a indeterminação das línguas naturais possam ser reduzidas. Em ambos os casos, percebe-se que o processo de tradução, feito anteriormente no sentido de baixo para cima (*bottom-up*), ou seja, utilizando apenas o conhecimento linguístico contido na própria sentença, começa progressivamente a comportar variações que procuram localizar, em primeiro plano, informações de natureza macroestrutural (como o universo do discurso, por exemplo), para, e apenas então, processar integralmente o material linguístico. São estratégias, portanto, que pretendem representar (ou emular) o conhecimento mobilizado pelo leitor durante o processo de tradução.

É forçoso reconhecer, contudo, que as estratégias, tanto no caso de KBMT quanto no caso de EBMT, encontram-se ainda em estágio experimental. Oferecem-se vários problemas para ambas as abordagens, muitos dos quais aparentemente insolúveis. Os sistemas de EBMT, por exemplo, estão amparados em juízos de similaridade (entre sentenças já traduzidas e sentenças a serem traduzidas) de natureza bastante controvertida. A similaridade estatística entre enunciados linguísticos muitas vezes não corresponde à identidade de significado. Não se pode afirmar, categoricamente, que uma mesma estrutura sintático-semântica, utilizada em contextos diferentes, seja portadora do mesmo significado. Provam-no todas as formas de uso figurativo da linguagem (caso da metáfora, por exemplo). Além disso, a versatilidade dos sistemas de EBMT está geralmente reduzida à amplitude do conjunto de traduções prévias, cuja variabilidade sacrifica, por sua vez, a exatidão do sistema, na medida em que passa a incluir, para uma mesma estrutura, possibilidades diversas. Ou seja, os sistemas de tradução baseados em exemplos têm sido prisioneiros de uma lógica perversa: quanto mais limitado o *corpus* de sentenças previamente traduzidas, piores os resultados; quanto mais rico, mais inexatas as respostas.

Algumas dessas mesmas ciladas valem para os sistemas de KBMT. É o caso, por exemplo, da construção de bases de conhecimento, enciclopédias e ontologias que venham a representar o conhecimento que o homem tem do mundo. O conhecimento humano tem parecido, em muitos instantes, não formalizável. Nem sempre é discreto, preciso e, especialmente, nunca é estático. Sua organização é variável, e o repertório de conceitos e de relações entre conceitos é dependente não apenas da cultura, mas da experiência muitas vezes pessoal e intransferível dos interlocutores. A comunicação parece basear-se antes em um jogo de inferências de regras pouco conhecidas.

Também a categorização da realidade, fundamental para a organização de ontologias, é problema complexo, com desdobramentos na filosofia, nas ciências sociais e na psicologia, e não é de todo incontroverso que a realidade possa ser de fato encapsulada nas unidades descritivas que vêm sendo propostas. Os modelos de categorização sugeridos pela psicologia, por exemplo, são bastante variados: ora admite-se que categorias são integradas por todos os elementos que respeitam um conjunto finito e discreto de condições necessárias e suficientes de pertencimento categorial (o chamado modelo clássico, de inspiração aristotélica); ora propõe-se que seriam formadas por todos os elementos para os quais se pudesse estabelecer uma relação de parentesco (semelhança) com uma instância considerada prototípica (o modelo dos protótipos); ora pulveriza-se essa mesma instância prototípica, admitindo-se a existência de várias instâncias exemplares (o modelo dos exemplares); ora afirma-se que os processos de filiação a uma determinada categoria são externos à própria categoria e governados pelo contexto. Na medida em que as bases de conhecimento fazem, obrigatoriamente, escolhas entre as inúmeras teorias disponíveis, sacrificam inevitavelmente o alcance de suas propostas, e passam a estar confinadas a uma representação antes parcial do conhecimento humano.

Transfere-se, portanto, o problema da ambiguidade linguística, criando-se o problema da ambiguidade contextual. Se se revela efetivamente plausível que a tradução não seja uma atividade estritamente linguística, mas linguístico-cognitiva, não se revela igualmente óbvio em que medida a introdução de categorias cognitivas (ou de exemplos prévios) poderia resolver a ambiguidade e a indeter-

minação das formas linguísticas. Os modelos propostos têm ainda um sabor experimental, e se vêm efetivamente aprimorando os resultados dos sistemas anteriores, fazem-no em uma escala ainda relativamente modesta.

A par dos recursos utilizados (se apenas dicionários ou gramáticas, ou se também bases de conhecimento ou *corpora*), os sistemas de TA diferem entre si também em relação às estratégias utilizadas. São duas, nesse caso, as possibilidades: a tradução direta (os chamados sistemas de primeira geração) ou indireta (sistemas de segunda geração). A tradução indireta admite ainda duas variações: as abordagens de transferência (sintática ou semântica), e aquelas realizadas por meio de uma língua pivô, intermediária, também chamada "interlíngua".

A tradução direta prevê, em linhas gerais, que a língua de chegada seja considerada o próprio instrumento de análise da língua de partida. Ou seja, não haveria, em princípio, nenhum estágio intermediário entre uma e outra. O vocabulário da sentença de entrada seria automaticamente vertido para a língua de chegada por meio de um dicionário bilíngue, com a ajuda, talvez, de algum processamento morfológico. Uma vez geradas as equivalências lexicais na língua de chegada, haveria algum reordenamento (bastante superficial e localizado) dos itens lexicais, para produzir resultados mais aceitáveis (como a posposição do adjetivo, por exemplo, no caso das traduções do inglês para o português). Não haveria propriamente processamento sintático das sentenças originais da língua de partida, nem qualquer tipo de processamento semântico. Os sistemas de tradução direta constituiriam, pois, sistemas de tradução palavra-por-palavra, com a possibilidade de alguma pós-edição, automática, dos resultados. A Figura 1 ilustra a arquitetura geral de um sistema desta natureza.

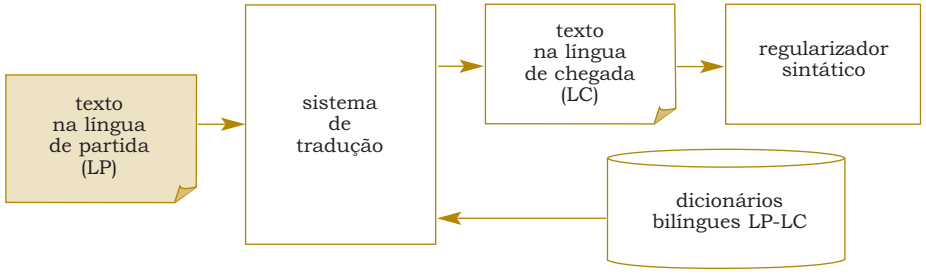

Figura 1 – Arquitetura geral de um sistema de tradução direta.

Esse modelo de tradução já provou há muito não ser adequado, não apenas porque há uma relação de muitos para muitos entre os conjuntos de palavras que integram as línguas naturais, mas também porque são frequentes as expressões idiomáticas e de sentido formulaico (isto é, não composicional) que não podem ser traduzidas a partir de suas unidades constituintes. Provérbios, por exemplo, perdem completamente o sentido quando traduzidos a partir das palavras que os compõem. É forçoso reconhecer, no entanto, que sistemas dessa natureza podem passar a incorporar dicionários específicos de expressões idiomáticas (e mesmo de sentenças inteiras recorrentes), ou podem ser enriquecidos para lidar com a

falta de correspondência estrita entre o vocabulário das duas línguas, criando-se, por exemplo, regras de desambiguização categorial amparadas em alguma sensibilidade ao cotexto mínimo (à esquerda e à direita).

O grande obstáculo ao sucesso dos sistemas de tradução direta consiste no grau de distanciamento das estruturas entre as línguas a serem traduzidas. O sistema está amparado na hipótese, hoje considerada ingênua, do isomorfismo sintático entre as línguas naturais, seja na sua versão forte (o mapeamento sintático é completamente dispensável), seja na sua versão mais fraca (o mapeamento sintático pode ser localizado em algumas construções bastante específicas, envolvendo quase sempre itens lexicais contíguos). Contra essa possibilidade, acumulam-se hoje evidências de toda sorte. Observa-se que mesmo línguas historicamente muito próximas – como o português e o espanhol, por exemplo – envolvem processos de gramaticalização muito diferentes, e que essas diferenças de estruturação sintática não são acidentais ou excepcionais, mas extremamente frequentes, e não podem ser negligenciadas no processo de tradução, sob o risco de serem produzidos resultados inúteis. Traduzir uma construção bastante corriqueira do português, como "Gosto de Pedro", para a correspondente em espanhol, "Me gusta Pedro", a par dos problemas relacionados à ambiguidade de "gosto" (que pode ser substantivo ou verbo, sem que o cotexto possa resolvê-lo), envolveria inverter completamente as relações sintáticas estabelecidas pelo verbo, fazendo do sujeito o objeto da oração, e do objeto, o sujeito (ainda que posposto). Sem essa inversão, dramática para a máquina, os resultados seriam agramaticais, e quiçá mesmo ininteligíveis para o usuário final monolíngue, desavisado das diferenças entre as duas línguas.

Em defesa, no entanto, dos sistemas de tradução direta, é preciso dizer que surgiram no início da história da TA, quando eram ainda utilizadas calculadoras numéricas para o processamento das informações. A ingenuidade teórica foi, em muitos casos, derivada de restrições de natureza prática, em um momento em que não havia ainda tecnologia e recursos disponíveis para um processamento mais refinado das línguas naturais.

A tradução indireta prevê o desenvolvimento de uma forma de representação intermediária entre a língua de partida e a língua de chegada. Essa forma de representação pode ser dependente das línguas envolvidas, no sentido de constituir uma interface específica (unidirecional ou bidirecional), ou pode ser independente tanto da língua de partida quanto da língua de chegada, procurando organizar-se como uma outra língua, artificial, autônoma, neutra, porém mais adequada ao processamento automático (porque livre de ambiguidade, por exemplo). No primeiro caso, fala-se em tradução indireta baseada em transferência; no segundo, em tradução indireta baseada em interlíngua.

A tradução por transferência envolve o desenvolvimento de uma metalinguagem entre a língua de partida e a língua de chegada. Essa metalíngua pode possuir caráter ora sintático, ora semântico, ora sintático-semântico, e consiste, quase sempre, em uma tabela de correspondências entre as duas línguas envolvidas. Na medida em que essa metalíngua seria diretamente decalcada da língua de partida e da língua de chegada, prevendo apenas suas possibilidades de combinação, não consistiria, ainda, uma interlíngua, mas tão somente um filtro necessário para o equacionamento das diferenças (principalmente estruturais) entre as duas línguas. A Figura 2 apresenta a arquitetura geral de um sistema de tradução por transferência.

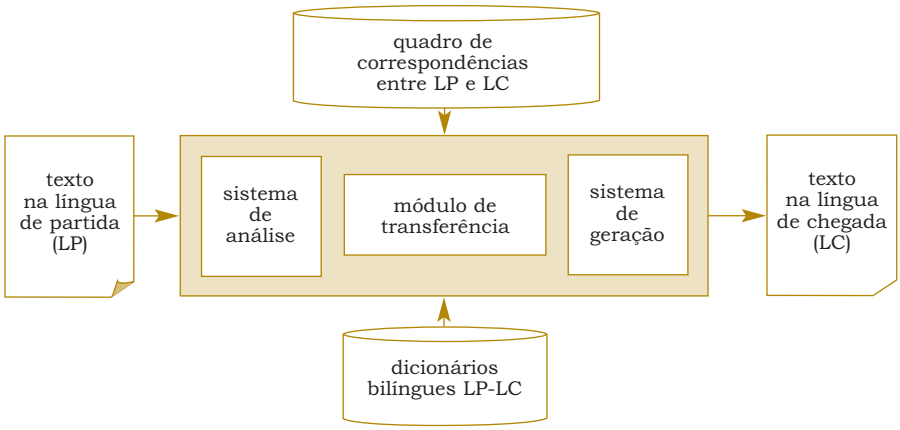

Figura 2 – Arquitetura geral de um sistema de tradução por transferência.

A abordagem baseada em interlíngua consiste, em última análise, no aprofundamento da abordagem baseada em transferência. A metalíngua de análise ganharia, nesse caso, o estatuto de componente autônoma, em princípio desligada seja da língua de partida, seja da língua de chegada. Ofereceria um sistema de representação de natureza não ambígua, para o qual seriam vertidas as informações disponíveis na sentença de entrada, e da qual seriam geradas as informações a serem incluídas nas sentenças de saída. Diferentemente do que ocorre em um sistema de transferência, essa representação seria (completamente) abstrata, no sentido de modularizar o processo de tradução, desmembrando-o em duas etapas marcadamente independentes: a projeção ou representação do texto de partida na língua intermediária seria feita independentemente da língua de chegada; e a geração do texto de saída, na língua de chegada, a partir dessa representação intermediária, seria executada à revelia das informações sobre o processo de análise. A Figura 3 apresenta a arquitetura geral de um sistema baseado em interlíngua.

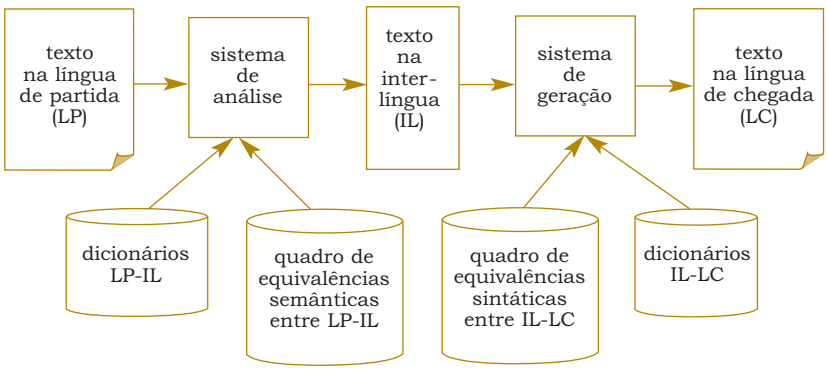

Figura 3 – Arquitetura geral de um sistema de tradução por interlíngua.

Para que pudesse integrar sistemas multilíngues, essa interlíngua – também chamada língua-pivô – deveria ser genérica (e plástica) o suficiente para acomodar diferenças, não apenas de conteúdo, mas de forma de representação, que seriam privativas de cada uma das línguas envolvidas. Reside nesse ponto a grande crítica à abordagem interlingual: em última análise, ela proporia a formalização de um sistema adâmico, pré-babélico, que pudesse conter todas as línguas existentes, das quais constituiria uma parte universal (tal como uma gramática subjacente a todas e a cada uma das línguas naturais). Seria, portanto, a língua perfeita, que compreenderia todas as demais (ou que estaria compreendida em todas elas). Os críticos da abordagem interlingual querem crer que essa é uma perspectiva exageradamente ingênua, na medida em que a existência de princípios universais (de uma gramática universal, enfim) é controvertida, e envolve um estágio de conhecimento da estrutura linguística a que não se teve ainda acesso.

Do ponto de vista teórico, a abordagem interlingual seria mais adequada para sistemas multilíngues, na medida em que a efetividade da modularização permitiria o desenvolvimento de sistemas independentes de análise e geração e, especialmente, reduziria o custo de incorporação de novas línguas ao sistema. A abordagem interlingual prevê a necessidade de um número de sistemas equivalente apenas ao dobro do número de línguas envolvidas (2n), dado que seriam necessários apenas os módulos da tradução da língua de partida para a interlíngua, e da interlíngua para a língua de chegada. Nas abordagens de transferência, para a produção dos mesmos resultados seriam necessários n(n-1) sistemas.

A par da vantagem operacional, a abordagem interlingual admite que os desenvolvedores tenham apenas o conhecimento de sua própria língua materna e da interlíngua proposta. Como a interlíngua, por seu caráter universal, conservaria várias das propriedades da língua do desenvolvedor, e dado que seria necessariamente não ambígua, é de esperar que as relações entre língua natural e interlíngua sejam mais simples do que as existentes entre duas línguas naturais, cada uma das quais organizada a partir de seus próprios princípios de imprecisão e vagueza. Por fim, e como vantagem adicional, a abordagem interlingual permitiria ainda a geração, de volta para a língua de partida, do texto já projetado para a língua intermediária. Seria essa uma excelente estratégia de validação do processo de representação e verificação dos resultados. Apenas os sistemas de transferência de natureza bidirecional, e ainda assim de forma bastante imprecisa (porque os resultados estariam inevitavelmente contaminados pela imprecisão da língua de chegada), poderiam realizar o mesmo movimento.

A apesar de todas essas vantagens, no entanto, a tradução baseada em interlíngua vem sendo preterida pela abordagem por transferência, mesmo em sistemas multilíngues. Não apenas porque não se pôde ainda chegar a uma interlíngua que contivesse, efetivamente, princípios mais gerais (universais), participantes de todas as outras línguas, ou de pelo menos um subconjunto expressivo de línguas (como as línguas neolatinas, por exemplo), mas pela complexidade do processo de projeção da língua de partida para a interlíngua. Apesar de teoricamente menos vantajosa, a abordagem por transferência tem provado que o desenvolvimento de interfaces específicas entre a língua de partida e a língua de chegada, embora exija formação bilíngue por parte do desenvolvedor, é menos complexo (e consequente menos oneroso, mais rápido e mais factível) do que os módulos de projeção para a interlíngua, envolvidos com sistemas de representação de natureza muito abstrata. Da mesma forma,

tem sido observado que esses mesmos módulos de transferência podem ser otimizados, de forma a serem reaproveitados, em alguma medida, por novas línguas a serem incorporadas ao sistema.

Perspectivas

A TA constitui hoje, no início do século XXI, um domínio de encruzilhada. Nenhuma outra aplicação envolve de forma tão cabal os desafios da interação entre modelos matemáticos, linguísticos e cognitivos. Na medida em que mobiliza competências e habilidades linguísticas de natureza antes geral – cujos recortes metodológicos e reducionismo têm significado, não a viabilização do processo, mas a má qualidade dos resultados –, a tradução parece requerer, como condição *sine qua non*, a formalização de um modelo global da linguagem e do conhecimento, de que a Linguística e a Inteligência Artificial estão ainda muito distantes.

Oferecem-se, então, dois caminhos básicos para a TA: o primeiro, a que aqui referirei como "idealista", aposta, em última instância, na possibilidade de mimetização, pela máquina, dos procedimentos e juízos efetivamente verificáveis para os homens. Acompanharia, em linhas gerais, a ideia de que seria importante persistir investigando o comportamento humano, para que se possam produzir modelos mais fiéis a serem imitados. É uma perspectiva que reforça, em muitos sentidos, o caráter derivado (aplicado) do PLN, a depender de uma ciência básica, a (psico)linguística.

Com efeito, a combinação de estratégias de formalização e representação do conhecimento linguístico e do conhecimento de mundo representa, sem sombra de dúvida, o grande desafio atual do domínio da TA. Existe mesmo quem proponha que a TA deva aguardar o desenvolvimento desses recursos linguísticos e conceituais antes de enveredar pela proposição de sistemas, que significariam, no atual estágio, apenas desperdício de tempo, energia e dinheiro. Ou seja, em razão da precariedade de informações a respeito da estrutura e do funcionamento da linguagem e do cérebro humano, não restaria alternativa senão investir, primeiramente, no desenvolvimento das ciências cognitivas (entre elas a linguística, a psicologia e a neurologia) para, e apenas então, chegar-se a um modelo minimamente capaz de emular a tradução humana.

À posição idealista contrapõem-se, como alternativa, as abordagens que aqui chamarei "utilitaristas", orientadas para a aplicação, mais preocupadas com a operacionalidade das estratégias (muitas vezes *ad hoc*) de formalização da linguagem. Já não se trata, aqui, da imitação e da representação do que o homem sabe ou julga saber acerca da língua, mas da (tentativa de) criação de um novo paradigma de trabalho. No primeiro caso, o papel da linguística é claro: são válidos para a máquina essencialmente os mesmos dispositivos e conceitos traçados para o homem, e restaria apenas implementá-los. No segundo caso, porém, esse papel é bem mais confuso: os dispositivos linguísticos só são válidos na medida em que formam o conjunto de condições de testagem de outros dispositivos e conceitos (de origem não linguística) capazes de produzir os mesmos resultados.

As diferenças entre as duas posições se espalham em muitos sentidos: os idealistas acompanham, em regra, modelos simbólicos, enquanto os utilitaristas oscilam entre modelos simbólicos e estatísticos. A principal diferença, porém, conforma o conjunto de premissas que servem de ponto de partida a cada abordagem. Os utilitaristas partem, com frequência, da observação de que aviões não batem asas. Acreditam que, assim como a emulação de um aspecto especí-

fico do comportamento das aves dispensou a imitação da fisiologia do voo e pôde ser alcançada por meio de estratégias alternativas, estaríamos também dispensados da obrigatoriedade de modelos diretamente baseados no comportamento humano para a simulação da habilidade de tradução. Afirma-se mesmo o contrário: o desenvolvimento de estratégias automáticas de tradução (embora imperfeitas e restritas) pode permitir que se amplie o conhecimento da habilidade (ou dos requisitos) do homem para a tradução. Longe de conduzir ao imobilismo, pois o reconhecimento das limitações da TA tem alimentado os desenvolvedores, certos de que apenas a experimentação poderá produzir, nesse caso, o conhecimento necessário para a implementação (futura) de sistemas bem-sucedidos. O desenvolvimento de sistemas de TA revelar-se-ia, pois, lugar extraordinário para a experimentação de teorias linguísticas e cognitivas, e para o acúmulo processual e contínuo de conhecimento sobre a linguagem, o pensamento e seu funcionamento

Referências

ALPAC. *Languages and machines*: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, DC: National Academy of Sciences; National Research Council, 1966. 124 p.

AUROUX, S. *A filosofia da linguagem*. Campinas: Editora da Unicamp, 1998.

BAR-HILLEL, Y. The present status of automatic translation of languages. *Advances in Computers*, v. 1, n. 1, p. 91-163, 1960.

BOITET, C. (Human-Aided) Machine translation: a better future? In: COLE, R. A. et al. (Ed.). *Survey of the state of the art in human language technology*. Oregon: NSF/CEC/CSLU; Oregon Graduate Institute, November, 1995a. Disponível em: <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>.

_____. Machine-aided human translation. In: COLE, R. A. et al. (Ed.). *Survey of the state of the art in human language technology*. Oregon: NSF/CEC/CSLU; Oregon Graduate Institute, November, 199b5. Disponível em: <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>.

CHEVALIER, M. et al. TAUM-METEO: Description du Systéme. Université de Montréal, 1978.

DORR, B. J. et al. A survey of current paradigms in machine translation. In: ZELKOWITZ, M. (Ed). *Advances in Computers*. London: Academic Press, 2000. p. 1-68.

ECO, U. *La recherche de la langue parfaite dans la culture européene*. Paris: Seuil, 1994.

FURUSE, O. and IIDA, H. Cooperation between transfer and analysis in example-based framework. In: PROCEEDINGS OF THE 14TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. Nantes, France, 1992.

HUTCHINS, J. *Machine translation:* past, present, future. Chichester: Ellis Horwood, 1986.

_____. The Georgetown-IBM demonstration, 7th January 1954. *MT News International*, n. 8, p. 15-18, May 1994.

HUTCHINS, J. The whisky was invisible, or persistent myths of MT. *MT News International*, n. 11, p. 17-18, June 1995.

_____. ALPAC: the (in)famous report. *MT News International*, n. 14, p. 9-12, June 1996.

_____. From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology. *Machine Translation*, n. 12, p. 195-252, 1997.

_____. Bar-Hillel's Survey, 1951. *Language Today*, n. 8, p. 22-23, May 1998.

_____. Warren Weaver memorandum: 50th anniversary of machine translation. *MT News International*, n. 22, p. 5-6, July 1999.

HUTCHINS, J.; SOMERS, H. L. *An introduction to machine translation*. San Diego: Academic Press, 1992.

KAY, M. Machine translation: the disappointing past and present. In: COLE, R. A. et al. (Ed.) *Survey of the state of the art in human language technology*. Oregon: NSF/CEC/CSLU; Oregon Graduate Institute, November 1995. Disponível em: <http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html>.

KAY, M. et al. *Verbmobil:* a translation system for face-to-face dialog. Technical Report. Stanford: Stanford University, 1991.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coerência textual*. São Paulo: Contexto, 1990.

MATEUS, M. H. TA: um pouco de história. In MATEUS, M. H.; BRANCO, A. H. (Org.). *Engenharia da linguagem*. Lisboa: Colibri, 1995.

NIRENBURG, S. (Ed.). *Machine translation* – Theoretical and methodological issues. Cambridge: Cambridge University Press, 1987.

_____. (Ed). *Progress in machine translation*. Amsterdam: IOS Press, 1993.

OLIVEIRA JUNIOR, O. N. et al. A critical analysis of the performance of English-Portuguese-English MT Systems. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA ESCRITA E FALADA, 5., 2000. *Anais...*

SANTOS, P. TA. In: MATEUS, M. H.; BRANCO, A. H. (Org.). *Engenharia da linguagem*. Lisboa: Colibri, 1995.

SLOCUM, J. A survey of machine translation: its history, current status and future prospects. In: SLOCUM, J. (Org.). *Machine translation systems*. Cambridge: Cambridge University Press, 1985.

MARTINS, R. T. Machine translation. *Todas as Letras* (São Paulo), volume 10, n. 2, p. 148-169, 2008.

Abstract: *The present essay aims at revising and constraining the domain of machine translation (MT), its history, goals, methods and shortcomings. We claim that MT, despite the rough results, is still an outstanding framework for exploring interactions between linguistic, cognitive and computational models of human language.*

Keywords: *Machine translation (MT); natural language processing; computational linguistics.*