PRECISÃO E *COBERTURA*: DOIS REQUISITOS EM CONFLITO NA CONSTITUIÇÃO DE UMA GRAMÁTICA PARA UM SISTEMA DE TRADUÇÃO AUTOMÁTICA

ANDRÉ ELISEU
ADRIANA CARDOSO
CATARINA MAGRO
ERMELINDA GONÇALINHO
(ILTEC)

Introdução

Esta comunicação assenta numa reflexão sobre questões de natureza linguística, motivada pelo processo de desenvolvimento de um sistema automático de processamento de linguagem, o qual está necessariamente submetido a uma tensão criada pela necessidade de satisfazer requisitos, em princípio, contraditórios. No caso dos sistemas de tradução automática[1] (STA), os pólos dessa tensão são a necessidade de optimização (que se traduz em termos de robustez, eficiência e, também, de *cobertura*) e o requisito de qualidade (quanto à adequação descritiva e generalidade das representações, a que nos referimos como *precisão*) que frequentemente se apresentam como mutuamente exclusivos quando se tomam decisões de implementação; mais especificamente, o texto apresentado resulta da experiência de desenvolvimento do projecto GLEP[2]. Começaremos por situar a discussão no contexto geral dos STA e, depois de apresentar uma definição dos conceitos de *precisão* e *cobertura*, faremos uma análise comparativa de dois sistemas de tradução; por fim, e num tom mais especulativo, abordaremos a questão, subjacente a esta discussão, da relação entre a teoria linguística e a linguística computacional.

Ao considerar as propriedades de um STA, devem ser tidas em conta as condicionantes externas a que tal sistema deve obedecer, as suas características internas e o seu desempenho. Estes três parâmetros — que são em larga medida

interdependentes — permitem caracterizar e avaliar genericamente os sistemas de TA.

As condicionantes externas incluem os requisitos estabelecidos anteriormente à etapa de concepção e desenvolvimento do STA, como o tipo de tradução que se pretende obter ou a definição do contexto de utilização do sistema (por exemplo, a opção por uma utilização industrial ou individual condiciona largamente a concepção dos sistemas, de acordo com os recursos computacionais disponíveis). Para tipificar os objectivos e contextos de utilização dos sistemas de tradução podemos utilizar a distinção entre tradução destinada à *aquisição de informação* e a destinada à *disseminação de informação*[3]. Exemplos do primeiro caso são os sistemas destinados a manipular enormes massas de dados, como os usados pela informação militar ou, mais recentemente, os sistemas disponíveis *online* para tradução de páginas na *World Wide Web*[4]. Em situações deste tipo, uma tradução de baixa qualidade pode ser suficiente, tendo em consideração os fins em vista.

Já no caso da tradução destinada à disseminação de informação —como é o caso, por exemplo, da tradução dos manuais técnicos que acompanham os produtos exportados para países de diferente língua — os critérios de qualidade da tradução são relevantes.

Dentro deste mesmo parâmetro podemos incluir a definição das ferramentas de TA, de acordo com os objectivos funcionais adoptados. Usualmente consideram-se as seguintes categorias: sistemas de Tradução Automática, em que a tradução é obtida sem intervenção humana[5], sistemas de Tradução Assistida por Computador de dois tipos: aqueles em que o computador realiza as tarefas de tradução solicitando a intervenção de um tradutor humano em determinadas circunstâncias (por exemplo, para resolver anáforas) ou aqueles em que um tradutor humano usa um sistema em que pode recorrer ao computador em determinadas situações (por exemplo, consultar uma base de dados terminológica). Finalmente, o terceiro tipo de sistema é representado pelos Bancos de Dados Terminológicos, que funcionam como uma ferramenta de apoio especializada.

Do ponto de vista das suas características internas, é habitual classificar os sistemas de TA, de acordo com as propriedades do seu motor de tradução, em sistemas directos e indirectos (cf. Arnold *et al* 1995, Slocum 1988). Os sistemas do primeiro tipo adoptam uma estratégia segundo a qual a tradução de uma frase da língua fonte para um frase da língua alvo consiste na sua transformação por substituição de palavras. Esta operação, que requer grandes dicionários bilingues, é baseada numa análise elementar das frases (a segmentação (parsing) produz apenas a representação requerida pela operação de substituição) e é eventualmente seguida pela aplicação de regras de reordenação e concordância. Os sistemas de tradução directa são sistemas rígidos que são desenhados

exclusivamente para efectuar a tradução de uma dada língua fonte para uma língua-alvo. A maioria dos sistemas de TA explorados comercialmente adopta esta arquitectura, como é o caso do SYSTRAN e do SPANAM / EGSPAM.

Ao longo dos último anos, a investigação em TA tem desenvolvido sistemas que adoptam uma arquitectura baseada na representação do conhecimento linguístico. Estes sistemas, ditos de arquitectura indirecta, incorporam informação linguística sobre as línguas de cada par de tradução sob a forma de representações formais (gramáticas, num dado sentido) bem como informação sobre as relações entre elas. Uma vez que o tratamento dado à língua fonte e à língua alvo é equivalente, estes sistemas são, em princípio, reversíveis. A tradução obtém-se pela aplicação sucessiva de uma gramática de análise e de uma gramática de síntese. Os sistemas indirectos requerem gramáticas abstractas e extensivas, já que a tradução é feita com base em representações obtidas a partir da segmentação de objectos linguísticos (pertencentes a diversos níveis de representação).

Os sistemas indirectos actuais correspondem a uma de duas variantes, designadas como sistemas de *transfer* e *interlingua*. Estas duas arquitecturas permitem estabelecer a relação de tradução segundo duas estratégias distintas: no caso dos sistemas de *transfer*, a gramática de análise e a gramática de síntese contêm representações de topo, abstractas mas particulares, que são postas em relação através de regras da componente de *transfer*, que projectam uma representação-fonte numa representação-alvo. No caso dos sistemas de *interlingua*, as representações de topo da gramática de análise e da gramática de síntese são universais no conjunto de línguas considerado e a tradução dispensa o sistema de *transfer*. Em ambas as arquitecturas, os sistemas são baseados em representações linguísticas (de diversa natureza e grau de abstracção) e a sua operação central é a produção dessas representações através de sistemas de regras, associados a dicionários.

Finalmente, o desempenho de um STA pode ser avaliado de uma forma sistemática pela aplicação de protocolos mais ou menos padronizados[6] que envolvem a observação de parâmetros como as capacidades funcionais (par de línguas incluído, tipo de texto, etc.), robustez, modularidade, etc. Neste tipo de avaliação são tidos em conta não apenas as características, os resultados e as performances de um sistema, como dados de natureza comercial e industrial. (cf. Jordan *et al* s.d., Hutchins 1997, King 1997, King & Falkedall 1990). Nesta comunicação apenas consideraremos um parâmetro de avaliação, acessível e directamente observável: a qualidade da tradução. Como critério para determinar a qualidade de uma tradução consideraremos a) o grau de correspondência entre uma frase na língua fonte e uma frase na língua alvo e b) a boa-formação da frase obtida na língua de chegada[7].

Em princípio, os sistemas de arquitectura directa assentam na computação dos valores lexicais das palavras representados em dicionários de grande dimensão, enquanto os sistemas indirectos possuem dicionários contendo apenas a informação lexical requerida para aplicação das regras. Assim, simplificando a questão, podemos considerar que os sistemas directos adoptam uma estratégia de tradução baseada na informação lexical, enquanto os sistemas indirectos se baseiam no processamento de informação estrutural. Isto leva a que os sistemas de cada tipo se distingam pela natureza dos seus módulos internos (dicionários e gramáticas, respectivamente). No entanto, as restrições a que a implementação de um sistema particular está necessariamente sujeita pode levar a adoptar soluções mistas. No caso de um sistema indirecto, por exemplo, para tratar de um caso específico, pode ser adoptada uma estratégia que assente na informação lexical e não na representação estrutural[8], como pode ser exemplificado pelo tratamento das estruturas de controlo de objecto na gramática do GLEP, às quais se aplica a regra geral que trata das estruturas ditransitivas, sendo a interpretação do elemento controlado determinada a partir das especificações lexicais do verbo[9] (cf. Eliseu org. 1998).

Precisão e cobertura

Os termos *precisão* e *cobertura* de um STA remetem para requisitos que, em princípio, são associados a cada um dos tipos de sistemas acima mencionados. Mais especificamente, quando aplicados ao caso de um sistema indirecto de *transfer*, como o protótipo desenvolvido pelo projecto GLEP, tais termos remetem para estratégias de desenvolvimento que se opõem mutuamente e que é necessário avaliar tendo em conta parâmetros como o desempenho do sistema, o tratamento do tipo de textos que se pretende traduzir, a adequação e possibilidade de implementação de cada alternativa.

As decisões que se tomam na fase de implementação de um sistema de tradução são determinadas, em parte, pela procura de um máximo de *precisão* e de *cobertura*.

Por *precisão*, entende-se o modo como as gramáticas codificam a informação linguística de forma a restringirem as frases possíveis numa determinada língua. Por *cobertura*, entende-se o número de sequências particulares que têm derivação numa gramática (cf. Pereira 1997). A primeira noção é extensionalmente equivalente ao conceito de adequação descritiva (no sentido chomskyano), enquanto a segunda é uma função do número e variedade (dentro de limites determinados) dos textos tratáveis pelo sistema.

Um sistema com grande *precisão* caracteriza-se pela qualidade dos *outputs*, sintacticamente correctos e semanticamente fiéis à língua alvo. Um sistema com larga *cobertura* caracteriza-se pela quantidade dos *outputs*, i.e., pela capacidade de dar conta de um número virtualmente infinito de sequências

Mas, se idealmente cada sistema de tradução automática procura alcançar o máximo de *precisão* e de *cobertura*, na prática, surgem algumas restrições que impedem a concretização deste objectivo. Estas restrições relacionam-se quer com a arquitectura, quer com a eficiência do sistema.

A gramática escolhida para um sistema de tradução vai determinar em parte a avaliação do sistema quanto aos requisitos de *precisão* e de *cobertura*.

Porém, antes de considerarmos mais de perto as relações existentes entre os vários níveis de eficácia e tipos de gramática, cumpre esclarecer o facto de numa discussão deste género se impor a utilização de conceitos operacionais abstractos que, em termos absolutos, não correspondem a realidades observáveis. Na verdade, é impensável conceber sistemas exclusivamente eficazes quanto à sua *precisão* ou *cobertura*. Note-se, por outro lado, que como referimos adiante, existe uma assimetria na definição dos sistemas de base lexical e de base estrutural: no primeiro caso, é possível desenhar sistemas cuja informação linguística seja codificada exclusivamente sob a forma de notação lexical, mas os sistemas assentes em gramáticas possuem necessariamente uma componente lexical associada ao sistema de regras.

Precisão e cobertura na prática: os casos dos sistemas GLEP e SYSTRAN

Como dissemos, pode considerar-se existir uma correlação, por um lado, entre *precisão* e sistemas baseados em gramáticas — que com a codificação de grande quantidade de informação sintáctico-semântica permite a geração de *outputs* mais precisos - e, por outro, entre *cobertura* e sistemas de base lexical — que com a codificação de grande quantidade de entradas lexicais permitem a geração de um maior número de sequências.

Para além disso, um sistema com uma componente lexical robusta e uma componente estrutural menos explorada terá, em princípio, resultados mais rápidos em termos de tempo de processamento uma vez que as duas componentes representam pesos diferentes para o sistema[10].

Reunir num mesmo sistema o desenvolvimento paralelo das duas componentes seria o cenário ideal. No entanto, o estado actual do conhecimento, as limitações do equipamento e a necessidade de optimizar os recursos informáticos disponíveis não permitem uma solução deste tipo. É neste sentido que se pode afirmar que *precisão* e *cobertura* estão em permanente conflito na implementação de uma gramática para um sistema de tradução (cf. Pereira 1997).

Vejamos, então, como funcionam e quais os resultados de dois sistemas de TA — o SYSTRAN e o GLEP — que adoptam respectivamente uma arquitectura directa e indirecta.

O GLEP é um projecto em curso no ILTEC que tem como objectivo global desenvolver um protótipo de um sistema de tradução automática formado por um par de gramáticas (de análise e de síntese), capazes de tratar um largo número de

estruturas linguísticas e de lidar com corpora de "texto real", e que, associadas a uma componente de *transferência*, permitem produzir traduções para o Português a partir de textos em Inglês.

As gramáticas e o módulo de *transferência* são implementados na plataforma ALEP (Advanced Language Engineering Platform), sendo a descrição linguística feita de acordo com as especificações formuladas para a gramática de análise do Português, desenvolvida no âmbito do projecto europeu LSGram (Large Scale Grammar). Trata-se de gramáticas de unificação que adoptam o formalismo das gramáticas TFS (Type Feature Structures) e que são baseadas nos princípios da teoria linguística designada HPSG (Head-driven Phrase Structure Grammar) de Pollard & Sag (1987, 1994).

De acordo com a classificação das estratégias de TA apresentada, o sistema que desenvolvemos integra-se na secção das estratégias de *transfer*, sendo constituído pelas componentes de análise, *transfer* e síntese. A componente de análise é responsável por atribuir descrições linguísticas a sequências bem formadas de caracteres na língua de origem (no caso, o Inglês). Para obter estas representações linguísticas, esta componente toma como *input* o resultado de uma série de operações que segmentam as sequências lineares e as associam a informação linguística representada sob a forma de regras. O *transfer* é a componente do sistema que permite projectar a informação linguística pertencente à língua fonte (no caso, o Inglês) nas estruturas da língua alvo (o Português) através de um sistema de regras. As regras aplicam-se sucessivamente, por encaixe, até que a estrutura esteja *saturada* e pronta para servir de *input* para a síntese. A estratégia adoptada foi a de basear o *transfer* na informação associada aos itens lexicais; assim, as entradas lexicais têm toda a informação linguística necessária para a formação das estruturas. A componente de síntese tem como objectivo produzir uma sequência linear bem-formada a partir da informação estrutural fornecida pelo módulo anterior. Num primeiro ciclo, que tem como *input* inicial o resultado do *transfer*, as regras de síntese aplicam-se sucessivamente de forma a obter uma estrutura linguística final. Em seguida, esta estrutura serve como *input* das operações que produzem sequências lineares bem formadas em Português.

Em princípio, análise e síntese são operações simétricas (e idealmente reversíveis) associadas entre si pelo módulo de *transferência*. Em análise, parte-se de sequências bem formadas de caracteres na língua de origem (Inglês) até se chegar à sua representação linguística; o *transfer* estabelece a relação entre esta estrutura da língua de origem e a estrutura correspondente da língua alvo; em síntese parte-se da estrutura obtida em *transfer* para se chegar a uma sequência bem-formada de caracteres na língua alvo (Português).

O SYSTRAN é um sistema de tradução automática comercial utilizado, entre outras empresas e organizações, pela Comunidade Europeia. Segundo os

dados publicados, este sistema traduz entre um 1 milhão e 1,5 milhão de palavras por hora e, na versão usada pela C.E., permite a manutenção e desenvolvimento de 17 pares de línguas, entre as quais o par Inglês/Português.

O SYSTRAN é um sistema de arquitectura directa constituído por três componentes: sistema básico, base de dados linguísticos e software linguístico. Fazem parte do sistema básico os programas que controlam o processo de tradução e algumas actividades periféricas. A base de dados linguísticos contém essencialmente dicionários e todos os dados que são processados pelo sistema básico e pelo software linguístico. Este último tem um tamanho de cerca de cem mil linhas de código de fonte por par de línguas e é composto por uma colecção completa de algoritmos e regras estruturadas que constituem a análise, o *transfer* e a geração.

A análise é a primeira fase do processo de tradução — cada frase é analisada da direita para a esquerda com base nas ligações entre os elementos lexicais, que são codificadas. O *transfer* é a fase bilingue do processo de tradução, dando conta dos aspectos específicos num dado par de línguas. No caso de os resultados desta fase não serem satisfatórios, são usadas regras supletivas, relativas à ordem de constituintes e relações sintácticas. A síntese ou geração é a etapa final que produz a tradução na língua alvo a partir dos resultados dos processos anteriores em associação com um dicionário.

Depois de a estrutura traduzida, o *output* final é determinado por um programa específico responsável pelo re-arranjo da sequência linear (ordem das palavras, inversão do sujeito e do predicado no caso de estruturas impessoais, inserção lexical de pronomes relativos, etc.).

Apresentamos, de seguida, alguns resultados obtidos num teste comparativo entre o GLEP e SYSTRAN, de modo a ilustrar as vantagens e inconvenientes de ambas as estratégias:

(1) Brian gives the cat to Daniel.
Brian dá o gato a Daniel./ Brian dá a gata a Daniel. (GLEP)
Brian dá o gato ao Daniel. (SYSTRAN)

O GLEP gera dois *outputs*, admitindo que *cat* possa ser traduzido por um nome no género feminino ou masculino, enquanto o SYSTRAN oferece apenas a possibilidade de uma tradução pelo nome masculino, não considerando a informação sobre a uniformidade de género de alguns substantivos do Inglês.

(2) Brian gives the Daniel the dog.
Brian dá o cão a Daniel. (GLEP)
Brian dá ao Daniel o cão. (SYSTRAN)

No caso de estruturas de *dative shift*, inexistentes em Português, ocorrem dois sintagmas nominais como complementos do verbo segundo a sequência Verbo-Tema-Alvo. Uma boa tradução para Português destas estruturas tem de assegurar não só a inserção da preposição (efectuada pelo GLEP e pelo SYSTRAN) mas também a alteração da ordem de constituintes, fazendo ocorrer o argumento Tema seguido do argumento Alvo. Em termos sintagmáticos a ordem correcta será, pois, V SN SP e não V SP SN, tal como se verifica na tradução proposta pelo SYSTRAN.

(3) Brian expects the dog to bite the cat.
Brian espera que o cão morda o/a gato/a. (GLEP)
Brian espera o cão morder o gato. (SYSTRAN)

O verbo matriz em Inglês é um verbo de elevação de objecto e em Português subcategoriza uma completiva conjuntiva. A correspondência entre as estruturas das duas línguas é tratada de forma correcta pelo GLEP. Pelo contrário, a tradução proposta pelo SYSTRAN não dá conta desta diferença, aplicando ao verbo Português a subcategorização do verbo Inglês e gerando, desta forma, uma frase agramatical.

(4) African share prices post modest gains.
Os preços de acções africanas denunciam lucros modestos. (GLEP)
Os preços de acções africanos denunciam lucros modestos. (GLEP)
Ganhos modestos do borne africano dos preços de parte. (SYSTRAN)

O GLEP gera dois *outputs* de forma a dar conta da possibilidade de o adjectivo *african* modificar os nomes *share* ou *prices*. O SYSTRAN interpreta toda a estrutura como um sintagma nominal, provavelmente por *post* estar apenas codificado como substantivo *borne* e não também como verbo *denunciar*. De acordo com o pressuposto de que em Inglês a modificação se faz à esquerda, a tradução proposta por este sistema interpreta *modest gains*, o constituinte mais à direita, como a cabeça do sintagma. Este, por sua vez, encontra-se modificado pelos restantes elementos que ocorrem, em Português, à direita. Um destes modificadores, *share*, encontra-se codificado apenas com uma entrada que lhe atribui a acepção de *parte* e não também de *acção*.

(5) Brian knows that Daniel knows the dog.
Brian sabe que Daniel conhece o cão. (GLEP)
Brian sabe que Daniel sabe o cão. (SYSTRAN)

A agramaticalidade da tradução proposta pelo SYSTRAN deve-se, provavelmente, ao facto de o verbo *knows* se encontrar codificado apenas com uma acepção. No GLEP, adoptou-se a estratégia de atribuir a tradução de acordo com os argumentos seleccionados pelo verbo em Inglês. Assim, quando o verbo selecciona um sintagma nominal a tradução é *conhecer* e quando selecciona um complemento frásico é *saber*.

(6) Brian knows that the big african dog tried to bite Daniel.
Brian sabe que o grande cão africano tentou morder Daniel. (GLEP)
Brian sabe que o cão africano grande tentou morder Daniel. (GLEP)
Brian sabe que o cão africano grande tentou morder Daniel. (SYSTRAN)

Em Inglês, os adjectivos ocorrem sempre à esquerda dos nomes que modificam. Em Português, ainda que ocorram por norma à direita do nome, há contextos específicos, como a dupla modificação, que permitem a ocorrência de um dos modificadores à esquerda do modificado[11]. O SYSTRAN não dá conta desta última possibilidade.

(7) Brian agrees that Daniel replaces the dog prices.
Brian concorda que Daniel substitua os preços dos cães. (GLEP)
Brian concorda que Daniel substitui os preços dos cães. (SYSTRAN)

Em Português, o verbo *concordar* selecciona uma completiva com verbo no modo conjuntivo. O SYSTRAN parece não codificar esta informação, propondo uma tradução agramatical com o verbo no modo indicativo.

(8) All dogs died.
Todos os cães morreram. (GLEP)
Todos os cães morreram. (SYSTRAN)
Todos os cães morridos. (SYSTRAN)

Ambos os sistemas propõem uma tradução correcta para (8). O SYSTRAN, no entanto, apresenta mais uma proposta, em que *died* é interpretado como particípio passado, gerando desta forma um sintagma nominal. Esta solução nunca poderia ser admitida pelo GLEP que identifica todas as sequências a tratar como frases, rejeitando qualquer sequência em que não ocorra um verbo.

(9) Brian persuades Daniel to buy the dog.
Brian persuade Daniel a comprar o cão. (GLEP)
Brian persuade Daniel comprar o cão. (SYSTRAN)

Em Português o verbo da oração subordinante subcategoriza dois complementos: um SN objecto e uma oração infinitiva introduzida pela preposição *a*. A tradução do SYSTRAN não dá conta desta selecção visto que não introduz a preposição, gerando assim um *output* agramatical.

O confronto dos resultados obtidos leva-nos a concluir que o SYSTRAN é um sistema de TA poderoso, munido de uma forte componente lexical que lhe permite dar resposta a grande número de sequências de texto real. No entanto, parece ter ficado claro, mesmo com a pouca representatividade das sequências testadas, que algumas das entradas lexicais não estão codificadas com toda a informação pertinente. Embora não tenhamos um conhecimento profundo do funcionamento deste sistema[12], pode afirmar-se que a sua componente estrutural é pouco desenvolvida, como ilustram os problemas encontrados na tradução de casos que requerem necessariamente o processamento de estruturas e não apenas de valores lexicais. Note-se igualmente que a componente lexical apresenta algumas deficiências, nomeadamente quanto a informação relativa a restrições de selecção de alguns predicadores lexicais.

Pelo seu lado, o GLEP, apetrechado com uma forte componente estrutural, obtém resultados de tradução mais precisos, isto é, mais correctos do ponto de vista sintáctico-semântico.

Deve, no entanto, dizer-se que a escolha dos exemplos foi determinada pela dimensão do léxico do nosso projecto na actual fase de desenvolvimento; para possibilitar uma comparação dos resultados dos dois sistemas, as sequências testadas tiveram de ser escolhidas de acordo com o léxico disponível no GLEP[13]. Assim, o grande trunfo do SYSTRAN, a sua grande *cobertura*, não pôde ser devidamente observado e a estratégia adoptada apenas permitiu a avaliação dos dois sistemas quanto à sua *precisão*, apontando para uma clara superioridade do GLEP neste campo.

Como é patente nas considerações anteriores, não entramos em linha de conta com outro dos principais parâmetros de avaliação da eficiência dos STA - o tempo de processamento, dado que, em relação a este ponto não é possível a comparação entre um produto comercial, como o SYSTRAN e um protótipo em desenvolvimento como é o caso do GLEP. Note-se, no entanto, que os sistemas directos apresentam, em princípio, uma vantagem do ponto de vista deste factor perante os sistemas indirectos, em virtude da sua arquitectura mais simples.

A arquitectura de um sistema de TA baseado numa gramática estrutural absorve grandes recursos computacionais em termos de *hardware* e será tanto mais viável quanto maiores forem esses recursos, porque o motor de tradução (a *máquina virtual*) é necessariamente implementada em linguagens de alto nível, sob uma forma altamente modular e com interacções complexas. No caso

particular das gramáticas implementadas em ALEP, isso leva a que o sistema opere segundo uma sequência que percorre toda a componente lexical para identificação do items pertencentes à expressão a tratar, para em seguida, de forma recursiva, tentar instanciar todas as regras estruturais, determinando as que se aplicam ao caso.

Teoria Linguística e Linguística Computacional

Esta comunicação defende o ponto de vista segundo o qual um sistema automático de processamento de linguagem natural, e em particular um sistema de tradução, deve incorporar informação linguística relevante, de modo a satisfazer critérios de adequação. Mais especificamente, tal informação deve ser expressa sob a forma de regras estruturais (ou seja, o sistema deve conter uma gramática) em associação com representações lexicais, único modo de garantir a capacidade de calcular as expressões linguísticas complexas; a esta capacidade nos referimos aqui sob a designação de *precisão*. Evidentemente, isto significa que a investigação nesta área deve acompanhar a investigação linguística.

A relação entre as áreas da Linguística Computacional e da Linguística Teórica é complexa e nem sempre são óbvias as conexões entre os avanços e aquisições em cada uma delas. Isto é particularmente evidente no caso particular da TA, dado que os processos de tradução em geral nunca foram objecto de estudo nas áreas linguísticas tradicionais.

Umas das razões que conduz a este estado de coisas é a não coincidência de objectivos e propósitos inerente à distinção entre uma abordagem de tipo teórico e outra necessariamente restringida pelas condicionantes ligadas à construção de um objecto empírico, mencionadas na Introdução.

Contudo, a observação mostra que, por um lado, os STA que assentam no Processamento do Conhecimento ou noutra abordagem — não-linguística — do problema defendida pela AI não são (ainda?) funcionais, enquanto, que por outro, os sistemas desenhados exclusivamente para satisfazerem critérios de eficiência, que assentam na capacidade de processar rapidamente enormes volumes de dados, frequentemente contêm representações linguísticas *ad hoc*, pelo que necessariamente falham ao processar estruturas não triviais (pense-se, por exemplo, na complexidade da computação das construções em que ocorrem "anáforas ligadas a longa distância" ou das chamadas *garden path sentences*). Estes factos sugerem a necessidade de incorporar o conhecimento linguístico sob uma forma (relativamente) próxima dos enunciados das teorias linguísticas.

Para além disso, há pelo menos uma relação histórica entre a Linguística Teórica e as variantes da Linguística Computacional que assentam no processamento da linguagem, dado que a possibilidade de formalizar o conhecimento linguístico foi originalmente demonstrada por linguistas como Chomsky, Harris e Bar-Hillel, entre outros.

No entanto, em determinados casos, a implementação de uma Teoria Linguística — que nunca é tarefa trivial — nem sempre é possível dado que, muito frequentemente, tais teorias contêm enunciados que não são totalmente explícitos ou utilizam representações que não são expressáveis em formalizações computáveis.

Neste contexto, tem um lugar particular a teoria de Pollard e Sag (1987, 1994), cuja formulação obedece a critérios de computabilidade (grau de explicitação dos princípios, formalização dos teoremas e adopção de um formalismo expressivo directamente computável - as gramáticas de unificação). Estas características levaram a que este modelo tivesse sido adoptado em vários projectos de Linguística Computacional baseados em gramáticas; o projecto GLEP pretende alargar o domínio da sua aplicação ao caso dos STA.

A ideia de basear um STA num modelo teórico abstracto confronta-se, no entanto, com o facto de a natureza explicativa e universal de um modelo teórico ser naturalmente diferente do carácter funcional e particular de um STA, o que leva a que aquele tenha de sofrer adaptações que o tornem implementável num contexto específico (neste caso, motivadas pelas características e restrições da plataforma de implementação ALEP).

Notas

1 A discussão que se segue será restringida aos aspectos dos sistemas da TA directamente relacionados com a representação do conhecimento linguístico. Outros aspectos, como, por exemplo, os algoritmos de base probabilística que alguns sistemas incorporam, não são aqui tidos em conta.

2 O projecto GLEP (Gramática de Larga Escala do Português) é financiado pelo programa PRAXIS XXI (PCSH/P/CLC/123/96.).

3 Cf. Slocum (1988). Podemos considerar que esta distinção é aplicável a qualquer forma de tradução, humana ou automática; contudo, recorde-se que embora a tradução literária possa ser considerada como um caso de disseminação de informação, os sistemas de TA aplicam-se exclusivamente a textos técnicos.

4 É o caso de uma versão do sistema Systran, acessível em conjunto com os resultados da pesquisa feita utilizando o AltaVista.

5 Esta caracterização diz respeito ao processo de tradução propriamente dito, não excluindo a intervenção humana na fase de pré-processamento (desde que não diga respeito à manipulação de informação linguística, como, e.g., a delimitação manual de constituintes) ou na fase de pós-edição.

6 Sobre a avaliação dos STA, ver os documentos produzidos pelo EAGLES Working Group on Evaluation.

7 Estamos a considerar exclusivamente o caso da tradução de frases no domínio da tradução técnica, deixando de lado certos aspectos estilísticos

8 Nestes termos, poder-se-ia ser tentado a considerar os sistemas indirectos como mais inclusivos que os sistemas directos, já que para além dos dicionários (que estes também têm), possuem uma gramática; no entanto, esta visão não seria adequada, dado que aquilo que os distingue é a diferente concepção da operação de tradução, que se reflecte na estrutura do motor de tradução (ou máquina virtual) e na forma de implementação (arquitectura da plataforma, linguagem de programação).

9 O tratamento dado a este caso segue a estratégia decidida no âmbito do projecto GSP (Gramática de Síntese do Português), desenvolvido no ILTEC em 1996-1997.

10 Isto torna-se evidente se, por exemplo, considerarmos como uma medida da eficiência de cada tipo de sistema o tempo de resposta - o intervalo entre o momento em que o *input* do utilizador é fornecido e o momento em que o sistema fornece um *output* (note-se que o tempo deve ser medido em termos de ocupação de CPU e não em tempo de relógio). A informação representada sob a forma de entrada em dicionários é mais rapidamente acedida que a informação introduzida sob a forma de regras, uma vez que requerem diferentes formas de armazenamento, codificação e processamento.

11 Para o tratamento das sequências de adjectivos neste contexto, ver Eliseu *et al.* (1997).

12 Dada a natureza comercial, do SYSTRAN, a informação pública sobre as suas características técnicas é reduzida.

13 Na actual fase, o objectivo prioritário da equipa é o desenvolvimento da gramática de síntese e da componente de *transfer*, bem como da incorporação dos *taggers* destinados a lidar com sequências não tratáveis pelas regras linguísticas, como datas, títulos, legendas, nomes próprios, etc. (cf. Genclioux & Murujo 1998, este volume). Por este motivo, e igualmente por razões práticas, tem sido utilizado um léxico de pequena dimensão; contudo, dada a estratégia utilizada para codificar a informação lexical, os dicionários poderão ser facilmente expandidos.

Bibliografia

Arnold, Doug, Lorna Balkan, R. Lee Humphreys, Siety Meijer and Louisa Sadler (1995) *Machine Translation: An Introductory Guide* [Online], disponível: http://clwww.essex.ac.uk /~doug/book/book.html.

Blache, Philippe and Jean-Yves Morin (1990), Bottom-Up Filtering: a Parsing Strategy for GPSG, *Proceedings of the 12th International Conference on Computational Linguistics*, vol. II, COLING, Budapeste.

Black, Ezra (1997) Evaluation of Broad-Coverage Natural-Language Parsers, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology*.

Boitet, Christian (1997) (Human-Aided) Machine Translation: A Better Future?, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology*.

Boitet, Christian (1997) Machine-Aided Human Translation, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

Briscoe, Ted (1997) Robust Parsing, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

Carbonell, Jaime G. and Masaru Tomita (1987) Knowledge-based machine translation, the CMU approach, in Nirenburg org. (1987).

Cullingford, Richard E. and Boyan A. Onyshkevych (1987) An experiment in lexicon-driven machine translation, in Nirenburg org. (1987).

Eliseu, André, Ana Lúcia Santos e Ermelinda Gonçalinho (1997) O problema da ordem dos modificadores adjectivais no contexto de um sistema de Processamento de Linguagem Natural Bilingue, *Actas do XIII Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa.

Eliseu, A. org. (1998) *Projecto (GLEP) Relatório de progresso. Implementation Report*, ILTEC, Lisboa.

Genelioux, Susana e Vitória Murujo (1998) Taggers em Gramáticas de Análise e Síntese, *Actas do XIV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa.

Hutchins, John (1997) Evaluation of Machine Translation and Translation Tools, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

Jensen, Karen (1988) Why Computational Grammarians Can Be Skeptical About Existing Linguistic Theories, in *Proceedings of COLING*, Budadpeste, vol. II.

Jordan, Pamela, Bonnie Dorr, John Benoit (s. d.) *A First-Pass Approach for Evaluating Machine Translation Systems*, ms., University of Maryland, UMIACS.

Joshi, Aravind (1997) Parsing Techniques, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

Kay, Martin (1997) Machine Translation: The Disappointing Past and Present, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

King, Margaret (1997) Human Factors and User Acceptability, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

King, Margaret & K. Falkedal (1990) Using test suites in evaluation of machine translation systems, *Proceedings of COLING.*

Muthusamy, Yeshwant K. and A. Lawrence Spitz (1997) Automatic Language Identification, *Linguistica Computazionale,* vol. XII-XIII - *Survey of the state of the art in human language technology.*

Nagao, Makoto (1987) Role of structural transformation in a machine translation system, in Nirenburg org. (1987).

Nirenburg, Sergei org. (1987) *Machine Translation. Theoretical and Methodological Issues*, Cambridge University Press, Cambridge.

Panevová, Jarmila and Sgall, Petr (1987) Machine Translation, Linguistics, and Interlingua, *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, Copenhagen.

Pereira, Fernando (1997) Sentence Modeling and Parsing, *Linguistica Computazionale*, vol. XII-XIII - *Survey of the state of the art in human language technology*.

Pollard, Carl e Ivan A. Sag (1994) *Head-Driven Phrase Structure Grammar*, University of Chicago Press, Chicago.

Roesner, Dietmar (1988) Why Implementors of Practical NLP Systems Can not Wait for Linguistic Theories, *Proceedings of the 12th International Conference on Computational Linguistics*, vol. II, COLING, Budapeste.

Raskin, Victor (1987) Linguistics and natural language processing, in Nirenburg org. (1987).

Sanfilippo, Antonio (1997), Lexicons for Constrains-Based Grammars, *Linguistica Computazionale*, vol. XII-XIII - *Survey of the state of the art in human language technology*.

Santos, Diana (1992) Broad-coverage machine translation, *The INESC Journal of Research & Development*, 3-1, Lisbon.

Santos, Pedro (1995) Tradução Automática, in Mateus, Maria Helena & António Branco orgs. (1995) *Engenharia da Linguagem*, Edicões Colibri, Lisboa.

Slocum, Jonathan (1988) A survey of Machine Translation: its History, Current Status, and Future Prospects, in Slocum org. (1988).

Slocum, Jonathan org. (1988) *Machine Translation Systems*, Cambridge University Press, Cambridge.

Sproat, Richard (1997) Text Interpretation for TtS Synthesis, *Linguistica Computazionale*, vol. XII-XIII - *Survey of the state of the art in human language technology*.

Tomita, Masaru (1988), "Linguistic" Sentences and "Real" Sentences, *Proceedings of the 12th International Conference on Computational Linguistics*, vol. II, COLING, Budapeste.

Tsujii, Jun-ichi, (1988) Reasons why I do not care grammar formalism, *Proceedings of the 12th International Conference on Computational Linguistics*, vol. II, COLING, Budapeste.

Uszkoreit, Hans and Annie Zaenen (1997) Grammar Formalisms, *Linguistica Computazionale*, vol. XII-XIII - *Survey of the state of the art in human language technology*.