LINGUÍSTICA E INFORMÁTICA:

PERSPECTIVAS RECENTES DO USO DO COMPUTADOR EM LINGUÍSTICA APLICADA E DESCRITIVA [1]

1. Introdução

A «interface» entre linguística e informática pode ser analisada sob prismas diversos. Uma área das ciências da computação que tem dedicado uma atenção particular ao estudo das línguas naturais é, como se sabe, a inteligência Artificial, cujas relações com a Linguística e a Psicologia Cognitiva são de tal forma estreitas que se pode hoje falar de um metadomínio científico em que elas se aglutinam, a Ciência Cognitiva.

Não é esse, contudo, o aspecto que desejamos abordar neste modesto trabalho. É apenas nosso propósito considerar algumas aplicações da informática, na sua maioria muito recentes, em diversos domínios da linguística, caso da lexicografia, tradução, reconhecimento/sintetização de voz e aprendizagem de línguas assistida por computador (secção 2). A própria metodologia da linguística tem sido afectada pelo crescente uso do computador, pelo que, numa segunda parte (secção 3), queremos avaliar alguns aspectos dessa transformação. Reflectiremos em particular sobre os contornos do que parece ser uma (re)emergência do paradigma descritivista na linguística actual, e apontaremos aspectos da sua interacção com a

---

[1] Devo à minha participação na «Third European Science Foundation summer school» (Pisa, Julho/Agosto de 1988), sobre «Computational Linguistics and Lexicography», muito do estímulo que motivou a realização deste trabalho. Ao Instituto Nacional de Investigação Científica e à Fundação Calouste Gulbenkian expresso a minha gratidão por terem tornado possível essa participação.

linguística computacional e a linguística aplicada. Subjacente a esta análise está a convicção de que muitas das recentes e futuras alterações da linguística se encontram determinadas pela sua necessária integração num novo quadro disciplinar articulado pelo *paradigma computacional*, o paradigma metodológico e epistemológico que está a caracterizar genericamente a ciência moderna.

2. Recentes desenvolvimentos de aplicações linguístico-computacionais

A evolução rápida da informática e da tecnologia de modo geral conduziram, como seria de esperar, ao desenvolvimento de aplicações linguístico-computacionais de tipos muito variados, desde os utilitários simples que equipam hoje os processadores de texto até aos produtos sofisticados produzidos em laboratórios de investigação universitária ou industrial. Daremos a seguir a indicação de algumas dessas aplicações, designadamente as que têm neste momento um especial interesse para a linguística em geral e para certos domínios da linguística aplicada em particular [2].

2.1. Utilitários linguísticos

Existe neste momento, em alguns casos já há alguns anos, um conjunto de utilitários linguísticos que se encontram na sua maioria associados ao processamento de texto. Referimo-nos, por exemplo, aos **verificadores** («checkers»), como os verificadores ortográficos («spellcheckers») de uso cada vez mais popular. De facto, muitos programas de processamento de texto incluem já verificadores deste tipo que, mediante comparação do texto introduzido com uma lista de palavras armazenadas em memória, assinalam erros e apresentam propostas de correcção.

De um ponto de vista linguístico, a verificação ortográfica é bastante rudimentar. O mesmo não se poderá dizer de dois outros tipos de verificadores que têm vindo a ser lançados mais recentemente: os verificadores gramaticais e estilísticos. Os primeiros efectuam um

---

[2] Grande parte das observações que se seguem procedem de informações contidas em revistas da especialidade, como *Language Monthly*, *Language International* e *Language Technology*.

controlo sobre falhas habituais, sobretudo ao nível da concordância e da pontuação, mas envolvem também processos mecânicos de análise morfológica e sintáctica já de alguma complexidade [3]. Neste momento, alguns processadores de texto já incluem verificadores gramaticais [4] e mesmo verificadores de estilo. Embora os verificadores gramaticais sejam considerados ainda muito incipientes, aguarda-se com alguma expectativa o lançamento do **Critique Grammar** da IBM, que está já há algum tempo a ser usado e desenvolvido por investigadores do «Natural Language Group» desta empresa americana. Ao grupo dos verificadores de estilo pertence, por exemplo, o programa **Readability,** que compara o texto introduzido no computador com nove estilos pré-programados, desde o tipo artigo jornalístico até à linguagem burocrática e publicitária, passando pela linguagem literária (!). Convém notar que este programa foi desenvolvido por um matemático sueco com base no trabalho de um linguista, Edmond H. Weiss, estando neste momento em preparação, para além da versão inglesa original, versões para o alemão, francês e espanhol.

Podemos incluir neste grupo de utilitários programas que elaboram automaticamente concordâncias e listas de palavras por ordem de frequência. Um dos mais conhecidos é o **Oxford Concordance Program,** de que existe já uma versão para microcomputadores DOS [5]. Um utilitário particularmente interessante é o desenvolvido por P. Legrand na Universidade de Nantes: baseando-se na «lexigramática» de M. Gross, este programa permite analisar as proprie-

---

[3] Programas como Grammatik, da casa americana «Reference Software», para a família IBM; Sensible Grammar, que possui uma vasta lista de construções indesejáveis de palavras e de frases e inclui sugestões a nível da pontuação (para o Mac); ou o potente DEC'S VAX Grammar Checker, que analisa frases detectando erros comuns ao nível da sintaxe, pontuação, capitalização, hifenização e ortografia, e apresenta sugestões de correcção.

[4] Caso da versão 4 do «Volkswriter», com a aplicação CorrecText, que corrige documentos analisando a estrutura das frases e identificando relações gramaticais das palavras na frase. Usa uma base de dados de 135 000 palavras, em que cada entrada é codificada com informação morfológica, sintáctica, ortográfica, e ainda com informações sobre hifenização, flexão e outras características linguísticas.

[5] Um programa semelhante, para computadores Mac, é o Stablex, desenvolvido pelo prof. A. Camlong em Toulouse e recentemente apresentado na Universidade do Porto, que transporta texto para uma folha de cálculo e contém facilidades adicionais no campo do tratamento estatístico do léxico.

dades formais contextuais de palavras, assinalando com «+» ou «—» a realização dessas propriedades; um sistema de busca permite definir subclasses gramaticais, como, por exemplo, as que pertencem ao grupo N-V-prep-V-Nhum.

2.2. Lexicologia e lexicografia assistida por computador

No domínio da lexicografia, a utilização dos recursos da informática tem sido notória e, em alguns casos, geradora de mudanças assinaláveis a nível da metodologia tradicional.

É o caso do dicionário lançado em 1987 pela editora britânica Collins, o **Collins COBUILD English language dictionary,** que resultou da cooperação da Collins com a universidade de Birmingham (mais especificamente, a «Birmingham University International Language Database»). Este dicionário trata os 2000 vocábulos de maior frequência do inglês e baseia a descrição linguística nos dados fornecidos por um 'corpus' de 20 milhões de palavras, cujas origens são muito diversificadas — desde romances e novelas, textos de jornais e revistas, documentos burocráticos de empresas, a textos orais transcritos, obtidos na rádio e televisão. Os seus destinatários são, primariamente, professores e aprendentes de inglês.

Embora possam ser-lhe tecidas críticas a diversos níveis [6], a verdade é que um dicionário deste tipo possui características inovadoras que o tornaram já um marco de referência obrigatória. Como observa um dos seus responsáveis [7]: *«What is new about the project, apart from technology, is the ability to get for the first time a view of a language which is both broad and comprehensive [...] based on a thorough study of the way words are used.»*

De facto, este dicionário assenta numa **evidência textual** que lhe confere uma autoridade muito particular. Para cada entrada é feita uma análise semântica (que inclui alguma informação de ordem pragmática), gramatical e fonética. No que respeita ao tratamento

---

[6] Cf. os comentários críticos na recensão feita em *Language Monthly,* 47, Agosto de 1987, p. 9, sobretudo no que respeita à apresentação da informação sintáctica. V. também a entrevista feita a Pierre Cousin, um dos principais responsáveis do projecto, em «Language Technology», 12, Março/Abril de 1988, p. 14.

[7] SINCLAIR, J. M. (ed.) — *Looking Up, an account of the COBUILD Project in lexical computing,* Collins, 1987, p. vii.

do significado, há um dado que torna a descrição semântica diferente da generalidade dos dicionários: o recurso sistemático ao contexto, propiciado pela base de dados linguísticos. Este facto é importante na resolução de diversos problemas clássicos da lexicografia, como, por exemplo, o da ambiguidade. Como observa um dos editores: *[...] context desambiguates. In continuous discourse, whether written or spoken, true ambiguity occurs rarely, except where a writer or speaker deliberately wants to be ambiguous. [...] A particular word is unlikely to be ambiguous at the moment of utterance, irrespective of how many different senses for it are recorded in a dictionary* [8].

O dicionário é apenas o primeiro dos projectos elaborados a partir da base COBUILD. Outras realizações estão previstas, em variadas áreas de investigação, como, por exemplo, no campo do ensino do inglês, o «Collins Cobuild English Course».

Vários projectos de estudo computacional do léxico estão a ser realizados para outras línguas, caso do «Trésor de la Langue Française» e o «Lessico della Lingua Italiana». Para o português, para além das realizações do «Português Fundamental», existem projectos sobretudo em torno de terminologias, e outros de alcance mais amplo mas ainda embrionários. Há também notícia de que está neste momento em desenvolvimento, em Araraquara, Brasil, um plano de criação de uma base de dados lexical de 5 milhões de palavras portuguesas, distribuídas da seguinte forma: 3 milhões de palavras extraídas a partir de textos brasileiros, 1 milhão a partir de textos portugueses e 1 milhão a partir de textos da África de língua portuguesa [9]. O objectivo deste projecto é a elaboração de um dicionário de frequências de palavras e uma lista de concordâncias.

Uma versão de dicionário que está a ganhar cada vez mais interesse é o chamado dicionário electrónico, isto é, um dicionário que a máquina pode ler e que, uma vez em memória, pode ser relacionado com outras aplicações (como, por exemplo, 'software' de tradução automatizada). Existem diversas realizações neste campo, desde o projecto de M. Gross na Universidade de Paris VII, em que se trabalha já há quase 20 anos, em associação com a construção de uma base de dados linguística, até aos mais recentes dicionários

---

[8] *Ibidem*, p. 87.

[9] Cf. «Language International», vol. 1, n.º 1, 1989, p. 33.

de tecnologia CD-ROM, como o **Oxford English Dictionary.** Este último propicia, num único disco compacto, o acesso electrónico ao conteúdo dos 12 volumes originais do OED, com possibilidades adicionais de pesquisa de palavras segundo critérios pré-definidos. Também o Dicionário **Webster** se encontra numa versão para computadores DOS, contendo 80 000 entradas e 470 000 sinónimos para 40 000 entradas do «thesaurus». Além destes, estão em produção dicionários bilingues da editora Collins, que têm a vantagem de poderem ser carregados na memória activa do computador, permanecendo «on-line» à disposição do utilizador.

Outros lançamentos recentes de dicionários CD-ROM, além do já referido OED, são: o **Harrap's Multilingual Dictionary,** com acesso a 5 milhões de palavras em oito línguas (inglês, alemão, espanhol, francês, holandês, italiano, japonês e chinês), estando previsto o acréscimo do sueco, dinamarquês, norueguês e finlandês; o **Termdok,** de origem sueca, que fornece traduções de 75 000 palavras do sueco, norueguês e finlandês para inglês, alemão, espanhol e russo. Um projecto britânico ainda em preparação, que deverá ser lançado em 1990, **Facts on File Visual Dictionary,** elaborado a pensar primariamente no ensino da língua, combina texto, som e imagem, permitindo a pesquisa rápida de palavras, acompanhadas da respectiva pronúncia e imagem. As versões iniciais são para inglês-francês, inglês-alemão e inglês-espanhol.

No domínio das terminologias, assiste-se a um rápido crescimento, quer de bases de dados, que de 'software' apropriado. É conhecida a base **Eurodicautom,** da ECHO («European Commission Host Organization»), que contém cerca de meio milhão de termos científicos e técnicos, frases contextuais e 90 000 abreviaturas, em todas as línguas da Comunidade Europeia. Em França, a base **Normaterm,** acessível pelo popular serviço francês de videotexto, o Minitel, possui 90 000 termos de domínios técnicos, como a agricultura e os têxteis, em inglês e francês. Outros programas constituem valiosos auxiliares no domínio da tradução técnica, como o **Superlex,** uma espécie de glossário residente na memória que substitui as velhas fichas pessoais elaboradas penosamente pelos tradutores.

A leitura óptica, com digitalização de imagem, feita por 'scanners' com uma capacidade de resolução cada vez maior, é um recurso tecnológico que está já e vai sem dúvida, nos anos que se seguem, facilitar a compulsão de vastos 'corpora' linguísticos. Embora

ainda incipientes, alguns programas de reconhecimento óptico de caracteres (OCR) apresentam já capacidades notáveis de reconhecimento de um largo espectro de tipos de letra. Este facto, para dar apenas um exemplo, está a dar novas perspectivas aos estudos de textos em línguas clássicas, dadas as facilidades de edição que propicia.

2.3. Tradução automática

A tradução automática é uma velha aspiração em que hoje se concentram grandes meios científicos e tecnológicos. Convém distinguir a simples tradução assistida por computador, em que se integram os conhecidos como sistemas CAT («computer-aided translation»), da tradução automatizada propriamente dita (MT, «machine translation»).

No primeiro grupo podemos incluir um vasto conjunto de 'software' que tem vindo a ser desenvolvido como apoio aos diversos tipos de tradução. Alguns desses programas assistem a tradução no campo da correspondência comercial geral [10], da correspondência bancária [11], dos manuais técnicos e científicos [12], ou possibilitam a gerência de grandes bases de dados de terminologias multilingues [13]. Na maioria dos casos, trata-se de uma tradução semi-automática, em que é exigida a intervenção do tradutor ao nível da sintaxe e da correcção da própria tradução sugerida pelo computador. Seja como for, temos a redução de pelo menos parte do trabalho do tradutor: este terá sempre a seu cargo a tarefa da pós-edição.

A par destes podemos colocar potentes sistemas de tradução automática, como o **SYSTRAN,** desenvolvido em França, largamente usado pelos serviços de tradução da Comunidade, mas cujos serviços são também acessíveis ao público (por Minitel). Este sistema de MT inclui dois tipos de dicionários, gerais (o maior dos quais é o Russo-Inglês, com 500 000 entradas) e específicos, cobrindo estes últimos cerca de 30 diferentes domínios, desde a electrónica à medicina.

---

[10] Caso do Lingua Write, que traduz para cinco línguas 2000 blocos de texto comercial. (Cf. «Language Technology», 11, 1988, p. 10).

[11] Opção contida pelo conhecido PC-Translator, cf. «Language Technology», 13, 1989, p. 34.

[12] Como o MicroCat, que, entre muitas outras, contém uma opção de tradução de inglês para português.

[13] Caso do Interdoc, abreviatura de «Integrated Terminology and Document Control System» (cf. «Language Technology», 13, 1989, p. 3).

Este programa possibilita a tradução, sobretudo a partir do inglês, para todas as línguas da Europa Ocidental e o árabe; do alemão para inglês, francês, italiano e espanhol; do francês para inglês e alemão; e do russo, espanhol, italiano e português para inglês. A sua 'performance' é considerada razoável, com percentagens de correcção entre os 70 % e os 95 % e uma velocidade de cerca de 8000 palavras/hora.

Outros sistemas comerciais de CAT e MT estão em lançamento, embora o interesse actual de empresas privadas pelos sistemas de tradução automático seja ainda relativamente reduzido [14].

No campo da tradução automática, um dos projectos de maior alcance existentes neste momento a nível mundial é o EUROTRA, um projecto da CEE que se desenvolve já há sete anos e a que Portugal se encontra associado. Embora não exista muita informação sobre os progressos do trabalho em curso, o que alguns interpretam como falta de abertura e transparência por parte dos responsáveis [15], são conhecidas as dificuldades que este projecto tem enfrentado, prevendo mesmo alguns o reconhecimento público da sua inviabilidade (como já aconteceu anteriormente com outros projectos, como o «Ikaros», «Palabre» ou «Spin»), o que, até agora, não se verificou. O EUROTRA pretende obter um protótipo de uma máquina avançada de tradução automática para as línguas da Comunidade, alvo que, todavia, parece situar-se ainda num horizonte longínquo [16]. Contudo,

---

[14] Um estudo da viabilidade comercial de projectos de tradução assistida por computador conclui que «*no CAT system is cost effective yet, and no CAT company ever made a real profit*» (in «Language Technology», 13, 1989, p. 67).

[15] Crítica, por exemplo, de um conhecido traductologista, o Prof. Knowles, que, na 10.ª conferência anual sobre «Tradução e o computador» (10 e 11 de Novembro de 1988), observou o seguinte: «[...] developing a multilingual, multifunctional system such as EUROTRA is really akin to NASA trying to launch a space-shot to Mars. The true nature of the complexity is not known beforehand and those funding it having to accept that quite openly at the beginning». (In «Language International», vol. 1, n.º 1, pp. 23-28).

[16] Uma das deficiências estruturais do projecto, segundo PIGGOT, Ian — (*EC Policies for Coping with Multilingualism*, in «Language Technology, 13, 1989, p. 27), seria a seguinte: «[...] one of the biggest mistakes the Eurotra people made is that they had the national universities research the specific aspects of their languages, instead of searching for what these have in common. The numerous exceptions are impossible for the software to handle».

dado o envolvimento político e financeiro que rodeia este projecto, mesmo que se revele irrealizável a curto prazo, o mais provável é que continue enquanto projecto, de inegável interesse, de investigação europeia na área da linguística computacional.

O EUROTRA é bem representativo das dificuldades que os actuais sistemas de MT enfrentam. Uma dessas dificuldades prende-se com a necessidade de ultrapassar as barreiras da frase. As gramáticas desenvolvidas para computador são gramáticas de frase, o que faz com que a tradução trate as frases como partes distintas e não relacionáveis entre si. Torna-se muito complexa a tarefa de levar em conta, na tradução automática, fenómenos que dizem respeito à coerência e coesão textuais, em que se destacam a referência pronominal, relações temáticas, determinação do foco, análise dos conectores lógicos e a anáfora em geral. Um dos grandes desafios que se colocam à tradução automática é, reconhecidamente, a passagem de uma gramática de frase à análise do discurso.

2.4. Reconhecimento e sintetização da fala

Um domínio em que se colocam hoje também grandes esforços é o do reconhecimento e sintetização da fala por computador. Esta área tem tido grandes avanços nos últimos anos, existindo neste momento uma gama de sistemas que estão inclusive a ser comercializados, desde sistemas independentes do falante, isto é, que aceitam uma grande diversidade de realizações linguísticas, dentro de certos limites dialectais e etários, aos sistemas adaptáveis ao falante que necessitam de ser «treinados» para reconhecerem a voz de determinado sujeito; ou desde sistemas de palavras isoladas, isto é, sistemas que apenas reconhecem palavras ou frases curtas articuladas separadamente, a sistemas que aceitam discurso contínuo, a fala em sequência normal.

As fases típicas do reconhecimento automático são as seguintes [17]: processamento do sinal, em que o som captado por um microfone é interpretado de acordo com um (elevado) número de parâmetros; segmentação e classificação dos sons, isto é, elaboração de hipóteses

---

[17] Cf. «Language Technology», 10, 1988, p. 25.

sobre a natureza dos fonemas articulados; hipóteses de identificação das unidades lexicais a que podem corresponder os fonemas interpretados; e eliminação de ambiguidades, geralmente recorrendo a critérios de ordem sintáctica.

O mercado para os sistemas deste tipo é muito vasto, pelo que praticamente todas as grandes empresas de informática a nível mundial têm projectos de investigação em curso nesta área. Alguns sistemas, os chamados «voicewriters», máquinas que reconhecem alguns milhares de palavras pertencentes a domínios especializados do léxico (caso da linguagem comercial), estão já a ser usados e comercializados, embora exijam na sua vasta maioria a articulação de palavras isoladas.

2.5. Ensino/aprendizagem de línguas assistida por computador

Os sistemas «CALL» («Computer Aided Language Learning») estão em franco desenvolvimento, sobretudo graças às novas e apaixonantes perspectivas que a evolução dos «hipermedia» [18] proporciona neste momento. Vai já longe o tempo em que o computador era apenas uma espécie de extensão do laboratório de línguas. Diversos tipos de aplicação estão a ser usados, dentre os quais destacamos [19]: exercícios estruturais de tipo clássico, com manipulação mecânica de palavras e frases, espaços para preencher, etc.; programas tutoriais, em que o computador explica uma certa quantidade de material, de forma interactiva; simulações, em que o aprendente é mergulhado numa situação (criada apenas por texto ou também com o auxílio de som e imagem) que solicita o exercício linguístico; jogos, geralmente relativos à resolução de problemas de tipo variado; e programas utilitários diversos que assistem o estudante no aperfeiçoamento dos seus desempenhos a nível da ortografia, pontuação e escrita de modo geral.

---

[18] Os «hipermedia» possibilitam o acesso relacional e integrado a diferentes «media», desde o áudio e vídeo a gráficos e animação. Seleccionando, por exemplo, uma palavra em computador, é possível obter vários tipos de respostas: um excerto de texto, a pronúncia da palavra, uma imagem gráfica ou uma passagem de vídeo.

[19] Cf. UNDERWOOD, John — *CALL. Computer aided language learning*, in «Language Technology», 9, 1988, pp. 29-32.

Há neste domínio muitas hesitações e problemas a resolver [20]. Um deles diz respeito a certas limitações do computador enquanto tutor interactivo, como, por exemplo, a impossibilidade de aceitar como válidas respostas correctas que sejam diferentes das que tem armazenadas em memória. Torna-se assim necessário desenvolver programas cada vez mais inteligentes, que, por exemplo, sejam capazes de interpretar a origem de determinados erros, — mas isso já excede a própria linguística computacional, entrando no campo dos sistemas periciais baseados na estrutura do conhecimento.

A aprendizagem de línguas assistida por computador está destinada a grandes desenvolvimentos no futuro, pois trata-se de uma área crucial que pode ajudar a resolver muitos problemas presentes, mas que está dependente dos progressos que se venham a verificar noutras áreas. De qualquer forma, as expectativas são imensas neste domínio. Como observa um investigador [21], *«we are beginning to create language-learning tools like no others have had: tools that will ultimately change the way we think and feel about learning a language.»*

3. Linguísticas descritiva, computacional e aplicada: um quadro de interacção

As aplicações a que nos referimos na secção anterior, bem como muitas outras que não mencionámos, continuarão a desenvolver-se no futuro próximo, atraindo a uma cooperação especial três disciplinas linguísticas a que temos feito particular referência — a linguística descritiva (LD), a linguística computacional (LC) e a linguística aplicada (LA). Queremos a seguir considerar algumas perspectivas de interacção destas disciplinas linguísticas, considerando um quadro disciplinar comum em que os seus contributos específicos se podem integrar.

3.1. Linguística descritiva

O descritivismo caracterizou boa parte da actividade linguística deste século, embora de certa forma secundarizado pela emergência,

---

[20] Cf. HAYET, Marie C. — *Introducing CALL*, in «Language International», vol. 1, 2, 1989, p. 22.

[21] UNDERWOOD, J. — In *op. cit., ibidem.*

sobretudo a partir do início dos anos 60, de uma linguística de cariz fundamentalmente teórico e especulativo. Hoje, porém, sem embargo dos investimentos que se verificam no domínio da linguística teórica, torna-se óbvia uma forte reafirmação do paradigma descritivista, cremos que em grande parte motivada pela premente necessidade de informações fidedignas a respeito da estrutura das línguas naturais manifestada, em primeiro plano, pelas linguísticas aplicada e computacional.

De facto, no que respeita à LC, é óbvio que esta necessita, como seu «input», de todo um conjunto de descrições prévias das línguas naturais, talvez em medida igual ou suerior à sua necessidade de algoritmos mais potentes ou de novos recursos tecnológicos. Os grandes problemas que a LC enfrenta não são apenas os computacionais, mas também os linguísticos. Poderíamos mesmo fazer a suposição de um momento em que a técnica e algoritmia avançassem a níveis considerados satisfatórios, mas em que a evolução da LC seria detida por ausência de descrições relevantes das línguas naturais. O mesmo pode ser afirmado, *mutatis mutandis,* da LA. Veja-se o caso do ensino de línguas: é fundamental que esse ensino se baseie na língua **em uso,** sobretudo no que respeita à gramática e léxico.

Compete à LD dar resposta a estas solicitações, sendo certo que ela mesma não pode passar indemne à revolução informática. Hoje estão ao seu dispor instrumentos técnicos e conceptuais que podem redimensionar a sua actividade. Especificamente, interessa à LD (por todas as razões, incluindo provavelmente a da sua própria sobrevivência como ciência) articular-se no que hoje se designa por *paradigma computacional* da ciência. Tal perspectiva não significa o abandono do seu objecto específico de estudo — as línguas naturais, enquanto sistemas simbólicos de ordem física, fisiológica, mental e social —, antes um reforço e, talvez em alguns casos, um regresso a esse estudo. Procuramos a seguir definar os contornos dessa perspectiva.

Em primeiro lugar, a LD necessita de utilizar em seu favor os actuais recursos propiciados pela informática, o que implica que a sua actividade seja *assistida por computador.* Isso envolve, por exemplo, um reconhecimento das actuais capacidades do computador a nível do armazenamento de grandes volumes de dados e de pesquisa económica de informação. Essa capacidades devem ser exploradas fundamentalmente em favor da construção de grandes *bases de dados*

linguísticas (BDL), onde se organizem 'corpora' não apenas de linguagem escrita, mas também conversacional, bem como de terminologias técnicas e científicas, em constante mutação. As BDL devem incluir facilidades de utilização, como acesso rápido aos contextos, análise de frequências ou elaboração de concordâncias. Em segundo lugar, cabe à LD a exploração linguística dos 'corpora' contidos de forma organizada nas BDL, propondo descrições dos níveis fonético e fonológico, morfológico, sintáctico, semântico, pragmático, retórico e textual. Estas descrições devem ser *formalizáveis*, embora, a nosso ver, não necessariamente formalizadas [22].

Como vemos, a reafirmação nestes moldes da tradição descritivista conduz necessariamente ao retomar da análise do 'corpus' como elemento central do estudo da linguagem. Recordemos que, sobretudo a partir da distinção competência/«performance», os dados provenientes do 'corpus' foram considerados reflexo da «performance», e, por esse motivo, relegados para um plano secundário. O acesso à competência seria feito através da introspecção, devendo as intuições do sujeito falante substituir os dados empíricos, dados considerados essenciais por modelos descritivistas (como, por exemplo, o de Z. Harris, paradigmático em muitos aspectos [23]). A rejeição do 'corpus' foi feita em nome de alguns princípios, como sendo a falta de valor generalizável das descrições baseadas em 'corpus', a suposta exclusão do recurso à intuição que o uso de 'corpus' implicaria e, sobretudo, a fácil degeneração dos dados empíricos obtidos por processos mecânicos.

---

[22] «Formalização» não implica necessariamente «lógica». É sabido que uma boa área da inteligência Artificial que se especializou no processamento das línguas naturais preferiu o modelo, algo intuitivo, dos casos semânticos de Fillmore aos modelos da gramática generativa ou da semântica lógica formal de Mantague (cf. a este respeito, por ex. WINOGRAD, T. — *On some contested suppositions of generative linguistics about the scientific study of language*, in «Cognition», Maio de 1977, e SCHANK, R., WILENSKY — *Response to Dresher and Hornstein, ibidem*). Por outro lado, a tarefa essencial do linguista reside na área da pré-formalização, isto é, da descrição propriamente dita. Poderíamos pois concluir que a linguística que pode funcionar como «input» da linguística computacional não é necessariamente uma linguística formalizada, mas, de qualquer modo, um conjunto de informações formalizáveis.

[23] HARRIS, Z. — *Methods in Structural Linguistics*, Chicago, University of Chicago Press, 1951.

Hoje, graças ao desenvolvimento da informática, muitos destes argumentos perdem a sua validade. Como observámos na secção anterior, é hoje possível registar, armazenar, ordenar e pesquisar facilmente um imenso volume de informação linguística **real**. Pode-se contestar a representatividade de um 'corpus' de alguns milhares de palavras, mas não os actuais 'corpora' de vários milhões. De facto, dificilmente se poderá justificar, com a existência dos recursos actuais, que a teorização linguística se baseie em exemplos casuísticos construídos em muitos casos pelo próprio linguista (as famosas frases-de--linguista), de aceitabilidade duvidosa, sem reconhecimento da intervenção de factores que determinam a produção linguística, a variedade sócio-linguística, o próprio «medium» utilizado para veicular a mensagem e o contexto de um modo geral [24].

Por outro lado, a nova linguística descritiva não pode regressar ao mecanicismo dos modelos americanos dos anos 50. Um grupo de linguistas holandeses, envolvidos em linguística descritiva computacional, em moldes que nos parecem bem representativas do que poderíamos chamar hoje linguística neo-descritiva, aponta algumas diferenças básicas a esse nível [25]. Em primeiro lugar, diferente do descritivismo dos anos 50 é a não restrição aos dados do 'corpus', admitindo como válido o recurso à intuição (que, afinal, nunca foi totalmente suprimida dos modelos empiristas) e atribuindo-lhe mesmo um papel fundamental. Em segundo lugar, a análise do 'corpus' não pode ser feita mediante procedimentos mecânicos do tipo, por exemplo, dos propostos por Harris, mas assume a forma de **hipóteses** formuladas a respeito da estrutura do 'corpus'. Finalmente, o 'corpus' não é considerado como um bloco autónomo (fechado) de dados, conducente à sua própria descrição, mas como uma forma de testar as hipóteses.

---

[24] É sabido que, em muitos casos, um estudo dos dados fornecidos pelo 'corpus' desmente muitas conclusões estabelecidas. Um dos responsáveis do COBUILD, por exemplo, assinala o facto de «know» ocorrer mais frequentemente em expressões tipo bordão, como «you know», do que como verbo lexical pleno.

[25] «The Nijmegen Research Group for Corpus Linguistics», especializado na construção de bases de dados sintácticas. Cf. AARTS, J.; MEIJS, W. (eds.) — *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research*, Amsterdam, Rodopi, 1984.

3.2. **Linguística computacional**

Muito do que por vezes se chama linguística computacional é, de facto, *linguística assistida por computador,* caracterizada pelo uso ingénuo e meramente instrumental da máquina, enquanto a LC propriamente dita implica a construção de algoritmos em linguagem de programação e a sua implementação em computador. A diferença entre uma e outra pode de algum modo ser colocada em paralelo com a que se estabelece em informática entre *utilização* e *programação.*

No interior da LC podemos observar duas perspectivas distintas quanto à forma de abordar a língua. Uma que aborda a linguagem numa perspectiva mais próxima da engenharia da linguagem, em que se encara o processamento da língua natural essencialmente como meio de realização de projectos computacionais específicos (na área, por exemplo, do desenvolvimento de sistemas periciais). Outra que, embora apontando para objectivos semelhantes, mantém uma orientação que diríamos ser a da linguística clássica, em que o desenvolvimento de algoritmos que possam ser implementados em computador para fins de tratamento automatizado de domínios restritos da linguagem é articulado com a consideração do **sistema** da língua, na sua complexa diversidade e, sobretudo, na sua **irredutibilidade.**

Para além desta diferença básica de perspectiva, podemos observar outra de carácter mais restrito mas não menos importante: em alguma linguística computacional a língua é analisada directamente através de programas de computador criados geralmente «ad hoc», não sendo muitas vezes clara a teoria linguística que lhes subjaz; outra faz preceder a automatização da elaboração de gramáticas formais, que, em si mesmas, são independentes do computador. É evidente que é esta última que está mais de acordo com a sensibilidade do linguista no que respeita ao processamento automático da língua e a que nos referimos neste trabalho. Se a primeira tende em muitos casos a simplificar aspectos da língua, sacrificando o rigor descritivo em favor da capacidade de execução, a segunda recusa por princípio essa atitude reducionista e sublinha o que, na língua, é diversidade e complexidade, e o muito que é ainda objecto de interrogação e dificilmente computável (como é o caso de quase tudo o que se prende, por exemplo, com as dimensões discursiva e pragmática da linguagem). Em muitas aplicações linguístico-computacionais verifica-ce facilmente a existência destas duas tendências.

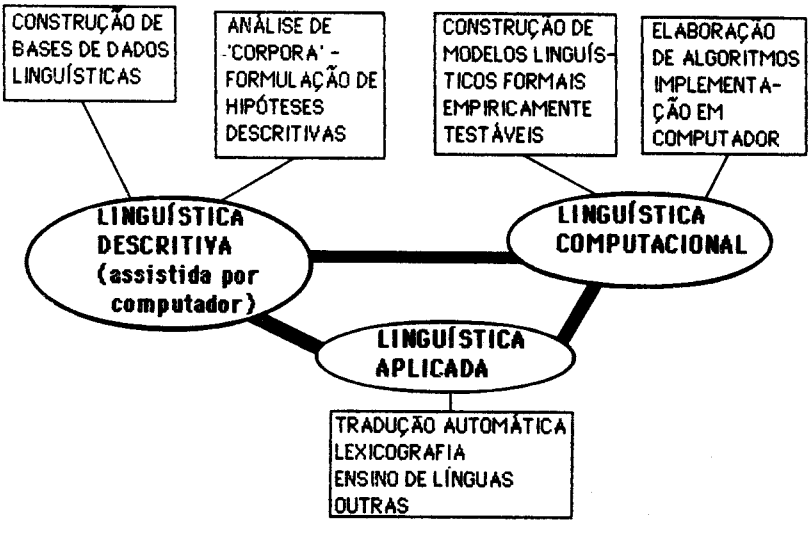
Contudo, dado o crescente envolvimento de linguistas «puros» nestes domínios, cremos que esta segunda tendência irá progressivamente adquirir mais peso e equilibrar a LC em muitos aspectos.

É a LC considerada nesta segunda perspectiva que necessita de dados descritivos da **língua real,** dados que a LD deverá tornar disponíveis. A partir daí, é necessário elaborar gramáticas formais, independentes da programação em si mesma. Finalmente, os algoritmos construídos necessitam de ser empiricamente testados, havendo para tal a possibilidade de recorrer às BDL.

3.3. Linguística aplicada

Neste momento, os maiores estímulos ao desenvolvimento da LC provêm de áreas da linguística aplicada. Vejam-se os actuais projectos de tradução automática, de elaboração de dicionários electrónicos, de 'parsers' que incluam informação semântica, de ensino da língua, de estudos contrastivos, de análise automática de erros e de muitos outros. É certamente em função destas e outras áreas da LA que hoje se trabalha em LC e mesmo em LD.

Podemos, pois, articular estas três disciplinas num contexto de relações e interdependências que, esquematicamente, teria uma configuração do tipo que sugerimos no quadro seguinte:



4. Conclusão

Procurámos neste trabalho observar a recente evolução de aplicações informáticas na área do processamento automático das línguas naturais com interesse especial para a linguística, em domínios como a lexicografia, a tradução automática e o ensino de línguas. Verificámos que os desenvolvimentos que se processam nestas áreas solicitam uma articulação em novos moldes das linguísticas descritiva, computacional e aplicada. No que respeita à primeira, observámos os novos contornos de que se reveste o paradigma descritivista na actualidade.

Cremos que este quadro disciplinar que tentámos esboçar cobre áreas de reconhecida carência no que respeita à língua portuguesa. Supomos que, futuramente, à semelhança do que ocorre internacionalmente, constituirão domínios privilegiados da investigação linguística no nosso país.

Porto, Maio de 1989

Sérgio Matos