Ambientes de processamento de corpora em português: Comparação entre dois sistemas

Diana Santos[1] and Elisabete Ranchhod[2]

[1] SINTEF Telecom and Informatics,
Box 124 Blindern, N-0314 Oslo, Noruega,
Diana.Santos@informatics.sintef.no, http://www.portugues.mct.pt
[2] Faculdade de Letras da Universidade de Lisboa e Centro de Automática da UTL
Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal,
elisabet@label.ist.utl.pt, http://label2.ist.utl.pt/LabEL/LabEL.htm

Abstract. Comparamos dois ambientes de processamento de corpora: o INTEX e o IMS Corpus Workbench. Depois de apresentadas as especificidades de cada um deles, discutimos as suas diferentes capacidades e a forma como se adequam ao processamento do português.

1 Introdução

O INTEX e o IMS-CWB são ambientes informáticos especialmente desenhados para facilitar o trabalho linguístico baseado na observação da forma como a língua é empregue – ou seja, recorrendo a grande quantidade de textos. Concebidos de forma totalmente independente e respondendo a desejos de comunidades distintas, apresentam-se actualmente como alternativas no mercado mais genérico das ferramentas de processamento de corpora ("corpus processing tools").

Visto que as autoras têm, cada uma, uma longa experiência com um destes ambientes, propusemo-nos descrevê-los e compará-los, por forma a ilustrar as possibilidades de cada sistema e clarificar, por um lado, as funcionalidades em que concorrem e, por outro, aquelas em que são complementares. É preciso, contudo, esclarecer que, enquanto que os recursos linguísticos do português do INTEX foram elaborados pela equipa da segunda autora (o LabEL), os corpora portugueses codificados com o IMS-CWB não foram compilados pela primeira autora.

Não pretendemos formular qualquer veredicto final sobre os sistemas, mas apenas identificar vários critérios e características que possam iluminar uma escolha consciente por parte de um público interessado nos problemas que discutimos aqui. O artigo está estruturado do seguinte modo: nos pontos 2 e 3, apresentamos uma pequena introdução ao INTEX e ao IMS-CWB, respectivamente. Em 4. discutimos várias diferenças e semelhanças a nível técnico, linguístico e prático, ilustrando, com exemplos simples do português, o modo de funcionamento e de utilização de cada um dos sistemas.

2 O INTEX

O INTEX é um sistema modular desenvolvido inicialmente para NextStep [32] e, mais recentemente, para Windows NT - 95/98, que pode ser usado para analisar corpora de muitos milhões de palavras [34]. Utiliza léxicos e gramáticas de grandes dimensões, representados por autómatos de estados finitos. Os dicionários e gramáticas são aplicados, em combinação, ao processamento de corpora a fim de, entre outras coisas: (i) reconhecer unidades lexicais, simples e compostas; (ii) identificar estruturas sintácticas ou léxico-sintácticas; (iii) resolver ambiguidades; (iv) etiquetar palavras ou expressões. Permite elaborar concordâncias lematizadas de estruturas linguísticas variadas, bem como indexar textos de modos diversos. Inclui igualmente ferramentas que auxiliam a manutenção e criação de dicionários e gramáticas e ferramentas que permitem obter dados estatísticos sobre os textos.

Estão actualmente implementadas neste sistema descrições de mais de uma dezena de línguas, entre as quais: alemão, espanhol, francês, inglês, italiano e português. Os recursos linguísticos do português são constituídos por [25, 23, 24] dicionários de palavras simples (110.000 lemas) e compostas (30.000 lemas) e por uma biblioteca de gramáticas, que inclui fundamentalmente gramáticas de análise e reconhecimento de estruturas léxico-sintácticas, gramáticas de resolução de ambiguidades e gramáticas de normalização de texto (divisão do texto em frases, separação de formas contraídas, etc.).

O modo como o sistema trata um dado texto é o seguinte: primeiro identifica todos os "tokens" (palavras, sinais de pontuação, marcadores de frase e algarismos), depois aplica os dicionários de palavras simples e compostas do português, indexando todas as unidades lexicais, ou seja, associando a cada uma delas a informação constante dos dicionários que aplicou.

O sistema associa todas as formas flexionadas ao seu lema (mais do que um, no caso das ambiguidades resultantes de homografia, por exemplo: entre,entrar.V:S1s:S4s:S3s:Y4s; entre,entre.PREP) e especifica os seus atributos linguísticos. Por exemplo: V:S1s:S4s:S3s:Y4s da forma verbal *entre* significa que esta forma corresponde à primeira, segunda (tratamento "você") e terceira pessoas do singular do presente do conjuntivo e à segunda pessoa do singular (tratamento "você") do imperativo. A ambiguidade nos compostos é muito menor. Em todo o caso, os compostos ambíguos estão incluídos num dicionário, e os não ambíguos num outro dicionário. Por exemplo, o ADV *em combinação* é ambíguo (devido à polissemia do nome *combinação* e à possibilidade de poder ser precedido pela preposição *em*, quer quando é um nome predicativo, relacionado com *combinar*, quer quando é um nome concreto).

Indexado o texto, várias outras operações simples podem ser efectuadas, como a lematização do texto e a etiquetagem das palavras não ambíguas. O sistema lematiza as palavras simples que têm o mesmo lema, independentemente de esse lema poder ser ambíguo, como seria o caso por exemplo da forma *grandes*, cujo lema, quer seja adjectivo ou nome, é GRANDE.

Da mesma forma, a etiquetagem das palavras é realizada pelo sistema quando estas só pertencem a uma categoria. Assim, nem a palavra *grandes* seria etique-

tada (por poder ser um adjectivo ou um nome) nem o advérbio *em combinação* seria etiquetado como composto (embora o fossem os seus constituintes).

Para tratar adequadamente estas situações, é preciso, por um lado, refinar as informações dos dicionários (de modo a distinguir com traços sintáctico semânticos os dois nomes homógrafos *combinação*) e, por outro, aplicar ao texto gramáticas de resolução de ambiguidades (para, por exemplo, distinguir o valor adjectival de *grande* do valor nominal que também pode ter). Na Fig. 1 apresentamos os autómatos (Finite State Transducers, FST) correspondentes ao texto *utiliza léxicos e gramáticas de grandes dimensões* antes e depois da aplicação de uma destas gramáticas [4], que elimina a leitura errada.

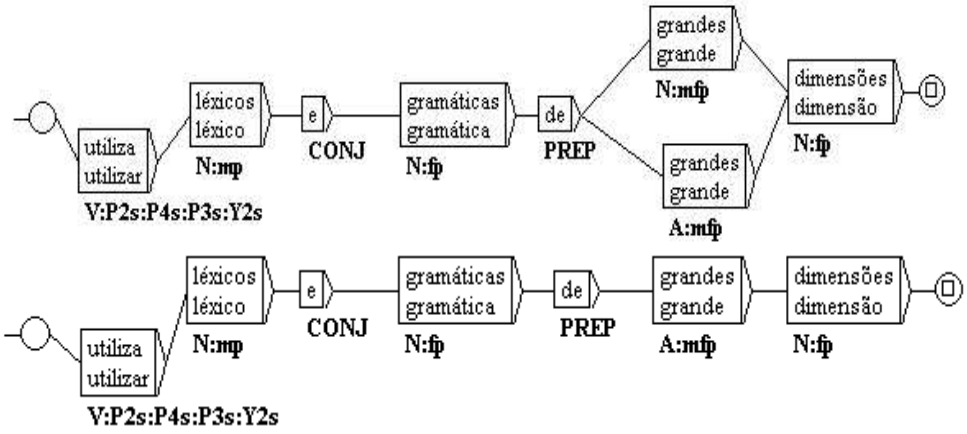

Fig. 1. Antes e depois da aplicação das gramáticas de resolução de ambiguidades

3 O IMS-CWB

O sistema de processamento de corpora do IMS[1] (por vezes referido como "corpus workbench" (CWB) ou como "corpus query system") é um conjunto de ferramentas computacionais pensadas para implementar eficientemente o tratamento de corpora de grandes dimensões com níveis de anotação variados. A arquitectura e a filosofia que presidiu ao desenho do sistema encontram-se descritas em [5, 29].

O módulo mais importante é o Corpus Query Processor (CQP), apresentado em [8] como um motor de procura especializado para a investigação linguística. Outros módulos que fazem parte integrante do sistem são Xkwic (uma interface gráfica para X [6]) e o Macro Processor (uma interface que estende as capacidades do CQP através da reutilização de padrões complexos de procura [30]).

[1] Institut für Maschinelle Sprachverarbeitung da Universidade de Stuttgart

Apesar de ter sido originalmente aplicado no processamento do alemão e do inglês, este ambiente é hoje em dia usado para línguas tão diferentes como o shona, o sueco, o bósnio, o norueguês e o checo, para não falar do português. Para a nossa língua, o IMS-CWB foi usado na disponibilização de corpora pela Internet (http://cgi.portugues.mct.pt/acesso.html). Os vários corpora codificados neste ambiente [28], todos eles acessíveis publicamente[2], são os seguintes:

NATURA Corpus jornalístico Natura-PUBLICO [19], 6 milhões de palavras

ENPCport Parte portuguesa do English-Norwegian Parallel Corpus (ENPC), [21], 200 mil palavras

MINHO Corpus jornalístico Natura-Diário do Minho [19], 1 milhão de palavras

ECI-EBR e ECI-EE Respectivamente a parte do corpus Borba-Ramsey e a da apresentação do programa Esprit do European Corpus Initiative, the Multilingual Corpus 1 (ECI/MCI) [35], 750 mil palavras

MLCC-DEB MLCC - Multilingual Corpora for co-operation, Sub-corpus: Official Journal of the European Commission, Annex: Debates of the European Parliament 1992-1994 [2], 9 milhões de palavras

SAOCARLOS Corpus NILC/São Carlos (parte corrigida) [20], 31 milhões de palavras

ANOTINESC Corpus INESC anotado [27], 17 mil palavras

4 Comparação entre os dois ambientes

Uma comparação entre sistemas tão complexos como os dois mencionados pode ser feita a vários níveis. Veja-se, por exemplo, a comparação de "corpus query tools" [31] que analisa mais de vinte sistemas diferentes[3] ou as recomendações do ponto de vista do utilizador de [3][4]. Não existe contudo, que nós saibamos, uma comparação entre o IMS-CWB e o INTEX (incluindo ou não outros sistemas também), que se torna relevante por serem aqueles em que mais trabalho existe para o português. Tentaremos estabelecê-la aqui de várias maneiras distintas.

4.1 Questões fundamentais

Para além do processamento de textos, o INTEX foi pensado para permitir o desenvolvimento da investigação em linguística. Um dos objectivos dos investigadores que com ele trabalham é, pois, a criação e/ou melhoramento de recursos linguísticos (dicionários e gramáticas de ampla cobertura lexical) formalizados em autómatos de estados finitos. Apresenta assim uma série de ferramentas dedicadas, para esse efeito.

[2] Nem todos possuem ainda autorização para ser disponibilizados através da Web.

[3] Os vinte e dois sistemas estão listados em [31]. O IMS-CWB faz parte de lista, mas não o INTEX. Veja-se também [12, 13].

[4] Neste artigo as autoras apresentam critérios que consideram importantes do ponto de vista do utilizador recorrendo a três sistemas diferentes como ilustração, sendo Xkwic, do IMS-CWB, um deles.

Por outro lado, o IMS CWB foi desenhado para ser o mais possível independente de teorias – quer sobre a língua quer sobre o seu processamento – não favorecendo, pois, nenhuma escola particular. Por exemplo, qualquer desambiguação ou refinamento da anotação de um dado corpus não faz pois estritamente parte das potencialidades do IMS-CWB. Um dos seus objectivos – ou aplicações preferenciais – é também, no entanto, a extracção de informação relevante para o trabalho lexicográfico, seja ele computacional [11, 15] ou tradicional [9]. Outra das aplicações é a extracção de terminologia [16].

Em termos de concepção, pode dizer-se que o INTEX vê o corpus como um objecto autónomo ao qual se aplicam vários recursos linguísticos (dicionários, gramáticas de análise, de resolução de ambiguidades, de normalização de texto, etc.) que o preparam para ser explorado de diversos modos.[5] O corpus é pois uma entidade que pode ser trabalhada de acordo com as necessidades e objectivos do utilizador: obtenção de resultados para utilização imediata (por exemplo: pesquisa de todas as unidades lexicais complexas, integradas ou não no contexto), ou de resultados necessários para poder prosseguir a investigação, usando instrumentos linguísticos cada vez mais sofisticados.[6]

Pelo contrário, o IMS-CWB vê o corpus como uma entidade com uma integridade própria, sobre o qual pode interrogar mas nunca alterar. Embora permita várias visões diferentes sobre o mesmo corpus (usando vários níveis de anotação diferentes, e mesmo entrando em conflito, para o mesmo corpus), todas essas visões foram criadas antes da fase de interrogação.

4.2 Questões técnicas

Uma diferança óbvia entre os dois sistemas, e que poderia levar, se não fosse a ubiquidade da WWW, à perfeita ignorância mútua, é o facto de os sistemas operativos em que os dois sistemas foram concebidos, e existem na prática, serem completamente diferentes. O INTEX foi desenvolvido para NextStep, contando também actualmente com uma versão para Windows 95/98 e NT, enquanto que o IMS-CWB foi desenvolvido para Sun OS 4.1.3, Solaris2 e Linux. Contudo, além das várias interfaces para WWW existentes para o CQP, existe agora uma versão em Java que possibilita o acesso remoto através da Web ao IMS-CWB[7], enquanto que o LADL, por seu turno, está a desenvolver um "Web-frontend" para o INTEX[10].

Não admira, pois, que as filosofias de desenho dos dois sistemas sejam tão diferentes, visto que concordam e conformam com as perspectivas dos sistemas operativos em que estão integradas. Temos, assim, para o IMS-CWB, a filosofia

[5] Os resultados dos vários tratamentos são escritos em ficheiros auxiliares, permanecendo o texto inicial sem qualquer alteração.

[6] Paskaleva [22] usa o INTEX para facilitar a criação de corpora anotados de búlgaro (desambiguados, portanto) enquanto que Abeillé et al. [1] usam o INTEX apenas como fornecedor de informação lexical – a validar por outros meios.

[7] Veja-se http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/CQPDemos.html.

de caixa de ferramentas e acesso remoto, típica do Unix, e para o INTEX, na interface Windows, a filosofia de ambiente para o qual se importam e dentro do qual se visualizam todos as funcionalidades, típica deste ambiente.

4.3 Questões linguísticas e de utilização

Independentemente do apanhado acima, que apresenta os dois sistemas como bons representantes, cada um, de um modo de estar na sociedade de informação, algumas questões há que transcendem a mera análise sociológica e que correspondem a capacidades linguísticas apresentadas pelos dois sistemas, que convém referir. Porque o utilizador destes dois sistemas é primordialmente o linguista, não separamos aqui as questões linguísticas (correspondentes aos objectivos teóricos do utilizador) e as questões práticas (correspondendo à sua forma de interagir com o sistema).

Em ambos os sistemas, é possível fazer pesquisas lexicais, morfológicas e/ou sintácticas, desde que o sistema tenha essa informação (no INTEX, associada ao léxico ou como resultado da aplicação de gramáticas, no IMS-CWB, na forma de anotação do corpus). Usando o INTEX sobre um corpus de 4 milhões de palavras, constituído por textos jornalísticos, recolhidos na Web, ilustramos alguns tipos de pesquisa. As mais simples são de ordem lexical: pesquisa de palavras já flexionadas ou de lemas, escritos entre angulares. Pesquisas mais complexas fazem uso de outros campos: Assim, a expressão regular: `(<ser> + <estar>) <A>` retira do texto todas as ocorrências dos auxiliares SER e ESTAR seguidos de adjectivo (apresentam-se as primeiras quatro ocorrências):

o o seu pessoal de Maliana por não estarem garantidas as condições de segurança e as
rrinho foi acordado de emergência. Estava iminente um ataque das milícias, e todos os
resultados". Para a mesma fonte "é difícil avançar qualquer cenário", mas se "os au
de ir. Não voltam para casa porque é inútil. A milícia garantiu-lhes que, se reinicia

Fig. 2. Pesquisa de adjectivos auxiliados por SER e ESTAR (INTEX)

Se o utilizador precisar de resultados que contemplem a possibilidade de haver um advérbio entre o auxiliar e o adjectivo (situação frequente), deverá escrever uma expressão regular como: `(<ser> + <estar>) (<E>+<ADV>)<A>`. Se quiser restingir esta pesquisa apenas aos casos em que os verbos estão no imperfeito do indicativo, especificará: `(<ser:I> + <estar:I>) (<E>+<ADV>)<A>`. As quatro primeiras ocorrências são apresentadas por ordem alfabética na Fig. 3.

Assim como é possível pedir ao INTEX uma procura especificando vários níveis de análise diferente (lexical, morfológica, sintáctica e léxico-sintáctica, etc.), a mesma potencialidade actualiza-se no IMS-CWB ao permitir que um corpus tenha várias anotações (níveis de anotação) e aceitando procuras multi-nível. Por exemplo, a instrução seguinte procuraria um adjectivo seguido de um nome (nível gramatical, marcado pelo atributo POS no exemplo) em que o

eniente. Ontem à noite, não era ainda possível perspectivar o sucesso da
a edição a situação em Timor era bastante instável. A Unamet evacuou todo
om os timorenses, a situação era muito diferente: a milícia estava na rua
foi acordado de emergência. Estava iminente um ataque das milícias, e tod

Fig. 3. Pesquisa de adjectivos auxiliados por SER e ESTAR no imperfeito com um advérbio opcional entre o auxiliar e o adjectivo (INTEX)

adjectivo fosse conotativo (nível semântico, especificado como valor do atributo SEM): `SINTSEM> [pos="ADJ.*" & sem="CON.*"] [pos="N.*];.`

Quanto à apresentação dos resultados, o INTEX fornece quatro alternativas: texto completo com os resultados das pesquisas sublinhados, concordância (especificada por número de caracteres parametrizável), frases (definidas por uma gramática especificamente construída) e parágrafos. O CQP não apresenta a primeira opção, permitindo por outro lado a especificação do contexto da concordância em termos de caracteres, "tokens" ou quaisquer atributos estruturais definidos no corpus. (É portanto comparável ao INTEX para um corpus apenas dividido em frases e parágrafos, mas para corpora com outras divisões tem maior flexibilidade: pode-se pedir para apresentar como contexto à esquerda um verso, contexto à direita 2 versos, etc..) A seguinte instrução procura todas as frases que têm simultaneamente o lema HOMEM e a forma *casa* e apresenta o contexto do resultado em duas frases (`set c 2 s`) (mostramos os dois primeiros casos):

```
ECI-EBR> set c 2 s; "home[mn]s*" []* "casa" within s;
2204: Soube mesmo pela empregada , que Dona Leonor não quis se levantar da cama
e pediu , apenas , um chá com torradas . Quanto aos <homens de casa> , podem
perfeitamente ter saído e comido fora . ( Não perguntei por eles , é claro .
73334: Foi na última chuvarada do ano , e a noite era negra . O <homem só estava
em casa> ; chegara tarde , exausto e molhado , depois de uma viagem de ônibus
mortificante , e comera , sem prazer , uma comida fria . Vestiu o pijama e ligou o rádio
, mas o rádio estava ruim , roncando e estalando .
```

Fig. 4. Pesquisa de frases contendo HOMEM e *casa* (CQP)

Além da apresentação dos resultados, a questão dos marcadores estruturais permite no IMS-CWB fazer perguntas restritas a partes diferentes de um corpus, desde que o corpus tenha sido anotado como tal. (Por exemplo, é possível procurar só em títulos, ou só em notas de rodapé, ou só em literatura do tipo oral, etc. Tal pode também pode ser feito no INTEX desde que tenham sido previamente introduzidos marcadores adequados). Veja-se as primeiras linhas do resultado da procura da preposição *em* a uma ou duas palavras do fim de um título, na figura seguinte.

Podemos também procurar todos os títulos em que aparece a palavra *em*, e apresentá-los sem mais contexto: `"em" %c within titulo expand titulo;.`

```
MINHO> "em" []1,2 </titulo>;
3184: il . Estímulos primários <em ambiente atractivo> A Associação possui uma
3760: nsformação do Sp . Braga <em SAD> Evidentemente que o fute
8444: LIA Angola : ano termina <em nova guerra> Angola chega hoje ao fin
9564: o de Acolhimento arranca <em 1999> Partindo das necessidade
24205: arkka finlandesa . Vamos <em frente> Salve 1-1-1999 P.e José
25141: talidade infantil baixou <em Cuba> Cuba atingiu no ano de 1
27318: Portuguesa redenominada <em euros> A redenominação da dívid
```

Fig. 5. Pesquisa da preposição *em* no fim de um título (CQP)

Poder parametrizar o tamanho da concordância é uma capacidade considerada importante, do ponto de vista do utilizador, por [3], que também refere a facilidade de manipulação das próprias concordâncias (possibilidade de ordenação, de extracção de um subconjunto aleatório – ambas neste caso satisfeitas pelos dois sistemas).

Os dois ambientes equivalem-se, também, no que se refere ao manuseamento da pontuação, ao tratarem os sinais de pontuação como "tokens" normais. Para, no INTEX, extrair os casos de aposição e coordenação de nomes, eventualmente seguidos de adjectivo, construiu-se o grafo da Fig. 6.

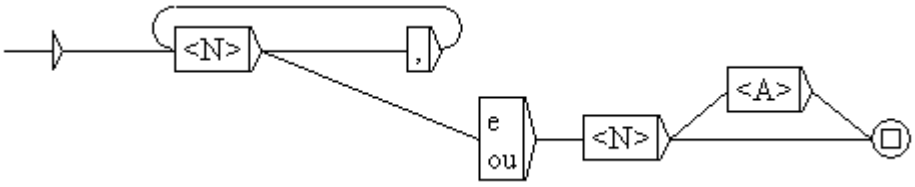

Fig. 6. Autómato simples para extracção de nomes em aposição e coordenação

Os primeiros resultados dessa pesquisa são apresentados na Fig. 7.

unto à qual se amontoam restos de alguidares e utensílios de cozinha, em ferro, com aos gritos. Um homem dos seus 30 anos, barba e cabelo hirsuto, saiu violentamente, e, os jovens, quatro ou cinco, com catanas ou facões na cintura. Depois os velhos, fi , uma cidade a 140 quilómetros de Díli e capital da zona de maior influência das mil ícias controlam todos os acessos a Ermera, Maliana e Lospalos. Desde segunda-feira re

Fig. 7. Resultados da pesquisa descrita no autómato anterior

Parece-nos que uma das maiores potencialidades do INTEX é o facto de permitir a criação de gramáticas "on the fly" e correspondente anotação, ou não, dos corpora. No CQP, é possível criar expressões regulares (equivalentes formalmente às gramáticas de estados finitos), mas não criar a anotação correspondente

no corpus, e extrair apenas os casos que obedecem a essa gramática. Ou seja, não é possível identificar certo tipo de características, e considerá-las como nova anotação, de uma forma tão fácil como o é no INTEX.

É, contudo, possível criar corpora locais que obedeçam a um certo critério, e refinar a procura apenas sobre esses resultados: Por exemplo, se estivéssemos interessados em extrair orações integrantes de um corpus (não anotado sintacticamente), uma forma de agir seria procurar primeiro as frases contendo um verbo declarativo (procurando os lemas DIZER, DECLARAR, AFIRMAR, ...), criando um subcorpus (um corpus local), sobre o qual a procura de uma oração iniciada por *que* (um item muito mais frequente) seria então executada.

Outras características que apenas podemos aflorar:

O IMS-CWB permite a invocação de programas externos, permitindo a consulta a informação externa ao corpus e produzindo o que se pode chamar anotações virtuais [7]. Também permite a procura em corpora alinhados – [17] descreve um protótipo de extracção de terminologia bilingue baseado no IMS-CWB. Estas duas características não se encontram na versão actual do INTEX.

A forma gráfica – muito amigável do ponto de vista do utilizador; cf. Fig. 6 – com que no INTEX é possível especificar gramáticas complexas [14] não tem paralelo no IMS-CWB, embora o mesmo poder expressivo esteja disponível através de etiquetas associadas a expressões regulares e recorrendo à definição de macros [30] com capacidades recursivas.

Finalmente, com o seu dicionário de palavras compostas e a possibilidade de análises alternativas de compostos ambíguos, o INTEX oferece uma modelação de compostos impossível no CQP, visto que este pressupõe a definição de um constituinte mínimo ("token"), exigindo, no processo de codificação de um corpus, que cada "token" (eventualmente constituído por um número arbitrário de palavras gráficas) esteja numa linha separada.

5 Conclusões

Neste artigo foram apresentadas, de forma sucinta, as características fundamentais de dois sistemas de processamento de corpora, o INTEX e o IMS-CWB. O seu modo de funcionamento e de utilização foi ilustrado com exemplos do português, tão elucidativos e completos quanto o permitiam o espaço de que dispúnhamos.

Apesar de terem ficado de fora muitos aspectos interessantes (limitámo-nos praticamente à apresentação de resultados em forma de concordância), terão ficado claras as potencialidades de cada um dos sistemas e a possibilidade de tirar partido de ambos para fazer progredir os estudos do português baseados em corpora.

O paralelismo estabelecido mostra que os dois sistemas obedecem a concepções diferentes e, por isso, põem a ênfase em tarefas e problemas distintos. O IMS-CWB está concebido para interrogar o corpus e obter informação que possa servir para estudos sintácticos, lexicográficos, estilísticos ou outros, cujo processamento e manipulação têm, contudo, sempre de passar pelo crivo do investi-

gador. O INTEX, embora permitindo igualmente aos linguistas, não especialistas em linguística computacional, utilizar o trabalho de investigação realizado para apoiar os seus estudos lexicográficos, morfo-sintácticos, léxico-sintácticos, etc., está primordialmente vocacionado para a elaboração de recursos linguísticos que possam ser usados por sistemas computacionais de forma automática ou semi-automática.

Seria interessante delimitar bem o estudo de um determinado fenómeno linguístico e, utilizando um único corpus, avaliar comparativamente os resultados obtidos por cada um dos sistemas. Esse trabalho ficará para uma próxima oportunidade.

Agradecimentos: Agradecemos à Cristina Mota os comentários e sugestões, que ajudaram a tornar mais claros e precisos vários pontos do texto. A investigação realizada pela segunda autora foi parcialmente financiada pelo projecto PRAXIS XXI 2/2.1/CSH/775/95.

References

1. Abeillé, A., Clément, L., Reyès, R.: TALANA Annotated Corpus : first results. In: Rubio et al. (eds.): Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 2, 993–7
2. Armstrong, S., Kempen, M., McKelvie, D., Petitpierre, D., Rapp, R., Thompson, H. S.: Multilingual Corpora for Cooperation. In: Rubio et al. (eds.): Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 2, 975–80
3. Brines-Moya, N., Hartill, J.: Criteria for user-oriented Evaluation of Monolingual Text Corpora. In: Rubio et al. (eds.): Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 2, 893–98
4. Carvalho, Paula: Elaboração de gramáticas para a resolução de ambiguidades entre Nomes e Adjectivos. Tese de Mestrado, Universidade de Lisboa (em preparação)
5. Christ, Oliver: A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research (Budapest, July 7-10, 1994), 23–32
6. Christ, Oliver: The Xkwic User Manual. Universität Stuttgart, IMS, August 1995
7. Christ, Oliver: Linking WordNet to a Corpus Query System. In: Nerbonne, John (ed.): Linguistic Databases. CSLI Publications, CSLI Stanford (1998) 189–202
8. Christ, O., Schulze, B., Hofmann, A., Koenig, E.: The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. IMS, University of Stuttgart, March 8, 1999 (CQP V2.2)
9. Docherty, V. J., Heid, U.: Computational Metalexicography in Practice - Corpus-based support for the revision of a commercial dictionary. In: Proceedings of the 1998 Euralex International Congress (Liège, August 4-8, 1998)
10. Fairon, Cédrick: Extraction automatique d'information sur le WEB avec INTEX. In: Proceedings of the Second Workshop of INTEX users (Paris, June 7-8, 1999) (no prelo)

11. Fillmore, C. J., Atkins, B.T.S.: FrameNet and Lexicographic Relevance. In: Rubio et al. (eds.): Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 1, 417–23
12. Grefenstette, G., Teufel, S., Gaschler, J. Schulze, M.B.: Deliverable D-2b of DECIDE (MLAP 93/19): Specifications for collocation extraction tools. Nov. 1994.
13. Grefenstette, G., Schulze, M. B.: Deliverable D-3a of DECIDE (MLAP 93/19): Prototype tools for extracting collocations from corpora. March 1995
14. Gross, Maurice: The Construction of Local Grammars. In: Roche, E., Schabes, Y. (eds.): Finite-State Language Processing, Cambridge, Mass./London: MIT Press (1997) 329–54
15. Heid, Ulrich: Building a Dictionary of German Support Verb Constructions from Text Corpora. In: Rubio et al. (eds.): Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 1, 69–73
16. Heid, U., Jauss, S., Krüger, K., Hohmann, A.: Term extraction with standard tools for corpus exploration: Experience from German. In: Proceedings of the 4th International Congress on Terminology and Knowledge Engineering - TKE'96 (Wien, 26-30/8/96)
17. Krüger, Katja: Mehrsprachige computergestütze Texterschliessung für Übersetzer und Terminologen. In: Proceedings of GLDV (Leipzig, 1997)
18. Mota, Cristina: Enhancing the INTEX morphological parser with lexical constraints. In: Proceedings of the Second Workshop of INTEX users (Paris, June 7-8, 1999) (no prelo)
19. Projecto Natura: http://www.di.uminho.pt/$\tilde{\text{j}}$j/pln/pln.html
20. Nunes, M.G.V., Vieira, F.M.C., Zavaglia, C., Sossolote, C.R.C., Hernandez, J.: A construção de um léxico para o português do Brasil: lições aprendidas e perspectivas. In: Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese (Curitiba, 23 a 25/10/96) 61–70
21. Oksefjell, Signe: A description of the English-Norwegian Parallel Corpus: Compilation and further developments. International Journal of Corpus Linguistics (no prelo)
22. Paskaleva, Elena: The lexical resources of highly inflected Slavonic languages in European standards and implementation formats. In Proceedings of The First International Conference on Language Resources and Evaluation (Granada, 28-30 May 1998), Vol. 2, 815–819
23. Ranchhod, Elisabete M.: Dicionários Electrónicos e Análise lexical Automática. In: Marrafa, P., Mota, M.A. (eds.): Linguística Computacional: Investigação Fundamental e Aplicações. Lisboa: Colibri (1999) 207–220
24. Ranchhod, E. M., Mota, C., Baptista, J.: A Computational Lexicon of Portuguese for Automatic Text Parsing. In: SIGLEX99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL (College Park, Maryland, USA, June 20-26, 1999), 74–80
25. Ranchhod, E. M., Mota, C.: Elaboração de dicionários terminológicos. Seguros.In: Marrafa, P., Mota, M.A. (eds.): Linguística Computacional: Investigação Fundamental e Aplicações. Lisboa: Colibri (1999) 221–23
26. Ranchhod, Elisabete M. Ressources linguistiques du portugais implémentées sous INTEX. In: Proceedings of the Second Workshop of INTEX users (Paris, June 7-8, 1999) (no prelo)
27. Santos, Diana (ed.): Processamento de corpora de texto no INESC. Relatório INESC no. RT/65-92, Dezembro 1992

28. Santos, Diana: Comparação de corpora em português: algumas experiências. In: Berber Sardinha, T. (ed.): A língua portuguesa no computador. São Paulo (no prelo)
29. Schulze, Bruno Maximilian: Entwurf und Implementierung eines Anfragesystems für Textcorpora. Diplomarbeit Nr. 1059, Universität Stuttgart (1994)
30. Schulze, Bruno Maximilian: MP User's Manual. IMS, Universität Stuttgart, 16 April 1996.
31. Schulze, M., Heid, U., Schmid, H., Schiller, A., Rooth, M., Grefenstette, G., Gaschler, J., Zaenen, A., Teufel, S.: Deliverable D-1b of DECIDE (MLAP 93/19): Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-oriented Tools. November 1994.
32. Silberztein, Max: Dictionnaires électroniques et analyse automatique de textes : le système INTEX. Masson Ed.: Paris (1993)
33. Silberzstein, Max: INTEX: a corpus processing system. In: Proceedings of COLING'94 (Kyoto, August 5-9, 1994).
34. Silberztein, Max: The Lexical Analysis of Natural Language. In: Roche, E., Schabes, Y. (eds.): Finite-State Language Processing, Cambridge, Mass./London: MIT Press (1997) 175–203.
35. Thompson, H., Armstrong-Warwick, S., McKelvie, D., et al.: Data in your language: The ECI Multilingual Corpus 1. In: Proceedings of the International Workshop on Shareable Natural Language Resources. Nara (1994)