Anotação de predicados complexos num *corpus* de português

Amália Mendes e Sílvia Pereira
Centro de Linguística da Universidade de Lisboa

Pretende-se, neste trabalho, dar conta do desenvolvimento de um novo recurso para o português que fornece informação sobre predicados complexos. Trata-se do *corpus* CINTIL-PREPLEXOS[1], um *corpus* com um milhão de palavras desenvolvido pelo Centro de Linguística da Universidade de Lisboa (CLUL)[2], etiquetado com categorias morfo-sintácticas, lematizado e revisto manualmente, que inclui ainda um nível de anotação com informação sobre alguns tipos de predicados complexos.

1. CONSTITUIÇÃO E ANOTAÇÃO DO *CORPUS*

O *corpus* CINTIL-PREPLEXOS, com um milhão de palavras, foi compilado com base em diferentes recursos previamente desenvolvidos pelo CLUL: o *corpus* PAROLE, escrito e etiquetado com informação morfo-sintáctica, com 250.000 palavras, que resultou de um projecto europeu que permitiu compilar *corpora* comparáveis para um grande conjunto de línguas (Bacelar do Nascimento *et al.* 2008); o *corpus* oral C-ORAL-ROM, com 300.000 palavras etiquetadas e lematizadas, e com alinhamento som-texto (Bacelar do Nascimento *et al.* 2005); textos escritos do Corpus *de Referência do Português* Contemporâneo – CRPC (Bacelar do Nascimento 2000): um *corpus* de grandes dimensões com mais de 300 milhões de palavras. A partir destes recursos, constituiu-se o *corpus* CINTIL[3], que contém um terço de transcrições de gravações (registos formais e informais), sendo os restantes dois terços constituídos por textos escritos. Este *corpus* foi inicialmente etiquetado e manualmente revisto numa parceria entre o grupo NLX (FCUL) e o CLUL. O *corpus* CINTIL-PREPLEXOS é um desenvolvimento do *corpus* CINTIL, tendo por objectivo acrescentar um novo nível de anotação, de acordo com a tipologia de construção de predicados complexos.

Este *corpus* constitui um recurso inovador no âmbito dos trabalhos sobre a língua portuguesa devido ao seu tamanho (1M), aos vários níveis de informação linguística que contém

[1] O *corpus* CINTIL-PREPLEXOS foi desenvolvido no âmbito do projecto Predicados Complexos: tipologia e anotação de *corpus* (PREPLEXOS), financiado pela Fundação para a Ciência e Tecnologia (PTDC/LIN/68241/2006). O trabalho de anotação morfo-sintáctica do *corpus* CINTIL foi desenvolvido no âmbito do projecto TagShare, pelo grupo NLX da Faculdade de Ciências da Universidade de Lisboa e pelo grupo RePort do CLUL, com financiamento da Fundação para a Ciência e Tecnologia (POSI/PLP/47058/2002).
[2] Os recursos aqui descritos estão disponíveis na página do CLUL (www.clul.ul.pt) ou na página da ELDA (www.elda.org).
[3] O *corpus* CINTIL está disponível para consulta de concordâncias em //cintil.ul.pt).

e à variedade dos tipos de textos que inclui (vários géneros de texto escrito e uma alargada cobertura de diferentes situações e registos de oralidade).

O primeiro nível de anotação do *corpus* inclui informação sobre a classe morfo-sintáctica, a flexão das classes abertas, expressões multi-lexicais pertencentes à classe dos advérbios e às classes fechadas e expressões com função de nome próprio. O conjunto de etiquetas foi desenhado para poder dar conta de informação específica dos textos orais, sobretudo informais, e contém portanto etiquetas para ocorrências de fragmentos de palavras e para elementos extra-linguísticos, entre outras. A este nível de anotação acresce ainda informação sobre o lema de cada unidade do *corpus*. Apresenta-se em (1) um pequeno extracto do *corpus* anotado com lema (em maiúsculas), classe de palavra (V, CN, inf, etc) traços flexionais (3s, ms, etc) e entidades que designam lugares, obras, pessoas ou eventos (como, por exemplo, B-LOC, I-LOC, no exemplo abaixo, em que B e I identificam respectivamente o início e o fim da expressão e LOC identifica o tipo de entidade, neste caso, um lugar):

(1) pretende/PRETENDER/vpi-3s[O] reconverter/RECONVERTER/inf#nifl[O] o/O/da#ms:O] centro/CENTRO/cn-ms[B-LOC] de/DE/prep[I-LOC] Matosinhos/MATOSINHOS/pnm[I-LOC]

2. TIPOLOGIA DE PREDICADOS COMPLEXOS

Consideramos como predicados complexos construções que partilham certas propriedades, descritas em Butt (1995): terem vários núcleos e uma estrutura argumental complexa; serem constituídos por mais do que um elemento, sendo que cada um deles fornece parte da informação geralmente associada ao núcleo; a estrutura desta expressão multi-lexical é idêntica à de um predicado simples. Para a anotação do *corpus*, centrámo-nos, sobretudo, em dois tipos específicos de predicados complexos, envolvendo:

- uma estrutura com dois verbos (por exemplo, *fazer rir*)

Assume-se, frequentemente, que estes predicados complexos incluem pelo menos dois verbos que se comportam como um único constituinte quando sujeitos a fenómenos de Subida de Clítico ou de Passiva Pronominal (cf. Kayne, 1975, Gonçalves, 2002, 2003) e que preservam ambos a sua estrutura argumental. Distinguem-se neste grupo as construções de reestruturação (*não o queriam ver*) e as construções causativas (*fazer rir*).

- um verbo leve (Jespersen, 1909/1949) seguido de um nome deverbal (*dar um passeio*) ou de um nome que expressa uma emoção ou sentimento, e que designaremos como nome psicológico (*ter medo*). Uma das propriedades mais referidas deste tipo de construção é a capacidade de ser parafraseada pelo verbo pleno correspondente, do qual o nome deriva (*ter um desmaio=desmaiar; dar um contributo=contribuir*). Nestas construções, tanto o verbo leve

como o nome deverbal contribuem para a predicação, pelo que a estrutura argumental e a distribuição de papéis temáticos é determinada, simultaneamente, pelos dois constituintes.

Tendo em consideração estes dois tipos de predicados complexos e partindo da observação dos dados disponíveis no *corpus*, pretendemos dar conta das suas propriedades sintáctico-lexicais e da sua interpretação numa abordagem integrada e aplicar esses resultados no novo nível de anotação.

3. ANOTAÇÃO DOS PREDICADOS COMPLEXOS

A anotação dos predicados complexos divide-se em categorias principais, que, posteriormente, definem as subcategorias. As categorias principais para os dois tipos de construção mais observados são construções que envolvem dois verbos (etiqueta [CV]) e construções com um verbo e um nome (etiqueta [CN]). No que diz respeito a construções verbo-verbo, estabeleceram-se duas subetiquetas, para as construções de reestruturação [CVR] (cf. 2a) e para as causativas [CVC] (cf. 2b). Quanto às construções verbo-nome distinguimos contextos em que o nome ocorre isolado [CNB] (cf. 3a) ou em que é acompanhado de um determinante [CN] (cf.3b). Cada elemento do predicado complexo é etiquetado com esta informação, como nos exemplos seguintes:

(2) a. Não o queriam[CVR] ver[CVR].[4]

b. Fazendo[CVC] traduzir[CVC] ao rapaz "Pucelle" de Voltaire

(3) a. dar[CNB] contribuições[CNB]

b. dar[CN] uma[CN] contribuição[CN]

A anotação também informa sobre a posição em que os elementos do predicado complexo ocorrem na ordem canónica (posição 1, 2, etc.), bem como sobre a posição em que os elementos ocorrem no *corpus*, posição inicial, intermédia ou final (B=Beginning, I=Intermedium, E=End). Tratando-se de predicados que não estão totalmente cristalizados, aceitam alterações à ordem em construções de negação, relativas, participais ou passivas. Apresentamos em (4) e (5) dois exemplos retirados do *corpus*, com todos os níveis de anotação: classe de palavras, flexão, entidades nomeadas, tipo de predicado complexo, posição canónica de cada elemento no predicado complexo e posição na qual ocorre no contexto específico:

(4) não/ADV[O] o/CL#ms3[O] queriam/QUERER/V#ii-3p[O] [CVR1_B] ver/VER/INF#nifl[O] [CVR2_E]

[4] Os exemplos apresentados são retirados do *corpus*, podendo nalguns casos ter sido encurtados.

(5) depois/LPREP1[O] de/LPREP2[O] um/UM#ms[O][CN2_B] aviso/AVISO/CN#ms[O][CN3_I] dado/DAR,DADO/PPA#ms[O][CN1_E]

A análise dos dados do *corpus* revelou determinados contextos para os quais foi necessário criar novas categorias e etiquetas:

a) muitas construções de predicado complexo podem ter uma interpretação ambígua. É o caso das construções com clítico *SE*, como no exemplo (6), em que o sintagma nominal *justiça* pode ser interpretado como sujeito do verbo mais alto (construção Passiva Pronominal) ou como objecto directo do segundo verbo (construção impessoal).

(6) Pretende-se cometer justiça.

Estes casos de ambiguidade são assinalados com uma etiqueta específica, que indica poder tratar-se de uma construção de reestruturação ou de uma construção com subordinada infinitiva [CVR_VINF], como exemplificado em (7):

(7) Pretende/PRETENDER/V#pi-3s[O][CVR_VINF1_B]-se/CL#gn3[O] cometer/COMETER/INF#nifl [O][CVR_VINF2_E] (…) justiça/JUSTIÇA/CN#fs[O]

b) construções em que determinado elemento pode ser parte de um predicado complexo de reestruturação e simultaneamente de um predicado complexo causativo, como no exemplo (8), em que *querer deixar* é um predicado de reestruturação e *deixar fugir* é ambíguo entre um predicado complexo causativo e um predicado seguido de subordinada infinitiva:

(8) não/ADV[O] o/CL#ms3[O] queriam/QUERER/V#ii-3p[O][CVR1_B] deixar/DEIXAR/INF#nifl[O][CVR2_E][CVC_VINF1_B] fugir/FUGIR/INF#nifl[O][CVC_VINF2_E]

c) construções de reestruturação em que o segundo verbo do predicado complexo entra numa estrutura de coordenação, como no exemplo (9) com a sequência *querer ouvir e registar*:

(9) repetiu/REPETIR/V#ppi-3s[O] profusamente/ADV[O] para/PREP[O] quem/REL[O] o/CL#ms3[O] quis/QUERER/V#ppi-3s[O][CVR1_B] ouvir/OUVIR/INF#nifl[O][CVR2_1_E] e/CJ[O] eventualmente/ADV[O] registar/REGISTAR/INF#nifl[O][CVR2_2_E]

O processo de anotação baseou-se na extracção de concordâncias dos contextos que são possíveis candidatos a estes tipos de predicado complexo. As construções do tipo verbo-verbo foram identificadas e anotadas tendo em conta os verbos que iniciam construções de predicados complexos de reestruturação (*querer*, *desejar, costumar, tentar, pretender, tencionar, conseguir*), bem como predicados complexos causativos (*mandar, deixar, fazer*). Assumindo que as construções deste tipo se comportam como um único constituinte quando sujeitas a

fenómenos de Subida de Clítico e Passiva Pronominal, procurámos contextos em que ocorrem estes dois fenómenos.

Além disso, em construções causativas, procurámos construções em que o sujeito do segundo verbo ocorre como objecto directo do verbo mais alto (10a), (10b) ou como objecto indirecto do verbo mais alto (11a-b).

(10) a. Esse perfume faz espirrar a Ana.

b. Esse perfume não a faz espirrar.

(11) a. A Maria mandou comer a sopa aos meninos.

b. A Maria não lhes mandou comer a sopa.

Tendo em conta a vasta lista de possíveis candidatos a predicados complexos com verbos leves, centrámo-nos em construções com os verbos *ter*, *dar* e *fazer* seguidos de um nome (deverbal ou psicológico). Apesar disso, o facto de os predicados complexos poderem ser descontínuos tornou o processo de pesquisa e anotação largamente complexo.

4. CONCLUSÕES

O *corpus* CINTIL-PREPLEXOS inclui a anotação de dois tipos principais de predicados complexos e vai ser disponibilizado *online*, na página do CLUL, para pesquisa de concordâncias. Pretende-se, com este novo recurso, apresentar dados de *corpora* que dêem uma perspectiva das propriedades globais destas construções, fornecendo, igualmente, informação importante para uma perspectiva contrastiva entre línguas. Novos dados poderão surgir, com esta análise, para a interpretação da interface Sintaxe-Semântica. O *corpus* CINTIL-PREPLEXOS é, também, um importante contributo para um futuro nível de anotação a nível sintáctico e semântico, pois apresenta uma anotação de predicados multi-lexicais, que têm de ser analisados como um único predicado e que se enquadram no esforço geral de anotação de diversos tipos de expressões multi-lexicais no nosso *corpus*: locuções prepositivas, conjuncionais e denotadoras de diversos tipos de entidades. O próximo passo será integrar na nossa anotação informação sobre expressões multi-lexicais de diversas categorias, mas com significado idiomático, cuja identificação foi feita a partir do *corpus* CRPC, no âmbito de investigações em curso no CLUL sobre fraseologia e combinatórias do português.

5. REFERÊNCIAS

Bacelar do Nascimento, Maria Fernanda, Palmira Marrafa, Luísa Alice S. Pereira, Ricardo Ribeiro, Rita Veloso & Luisa Wittmann (1998): "LE-PAROLE - Do *corpus* à modelização da informação lexical num sistema-multifunção". En *Actas do XIII Encontro da Associação Portuguesa de Linguística*. Lisboa:APL, Setembro de 1998, pp. 115-134.

Bacelar do Nascimento, Maria Fernanda, José Bettencourt Gonçalves, Rita Veloso, Sandra Antunes, Florbela Barreto & Raquel Amaro (2005): "The Portuguese Corpus". En Cresti, Emanuela & Massimo Monegnia (eds.): *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam/Philadelphia: John Benjamins Publishing Company, Studies in Corpus Linguistics nº15: 163-207.

Bacelar do Nascimento, Maria Fernanda (2000): "Corpus de Référence du Portugais Contemporain". En Bilger, Mireille (ed.) : *Corpus, Méthodologie et Applications Linguistiques*. Champion et Presses Universitaires de Perpignan : Paris. 25-30.

Barreto, Florbela, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes & João Silva (2006): "Open Resources and Tools for the Shallow Processing of Portuguese". En *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC2006), Genoa, Italy, XXXX.

Bowern, Claire (2006): "Inter theorical approaches to complex verb constructions: position paper". *The Eleventh Biennal Rice University Linguistics Symposium.*

Butt, Miriam (1995): *The structure of complex predicates.* Stanford, Califórnia: CSLI Publications.

Gonçalves, Anabela (2002): "The causee in the *faire*-Inf construction of Portuguese". *Journal of Portuguese Linguistics*, 1-2**.**

Gonçalves, Anabela (2003): "Defectividade funcional e predicados complexos em estruturas de Controlo do Português". En Castro, Ivo & Inês Duarte (orgs.): *Razões e Emoção. Miscelânea de estudos em homenagem a Maria Helena Mira Mateus*. Lisboa: Imprensa Nacional-Casa da Moeda, Vol. I: XXX.

Guasti, Maria Teresa (1993): *Causative and Perception Verbs*. Rosenberg & Sellier: Turim

Jespersen, Otto (1909/1949): *A Modern English Grammar on Historical Principles*. Londres: George Allen & Unwin; Copenhaga: Ejnar Munksgaard.

Kayne, Richard (1975): *French Syntax: the Transformational Cycle*. Cambridge, Mass.: The MIT Press.

Wurmbrand, Susi (1997): "Restructuring Infinitives". En *Proceedings of ConSOLE V*. Leiden: SOLE.