COMBINATÓRIAS LEXICAIS NUM CORPUS LINGUÍSTICO ESPECIALIZADO

Paula Marquez Neto
CLUL

1. Introdução

Esta comunicação destina-se a ilustrar um processo de análise lexical global, isto é, um conjunto de acções que levarão a uma compreensão do modo de funcionamento de um determinado tipo de linguagem – neste caso, o discurso científico –, representado pelo discurso da Astronomia.

Cada uma destas acções, ou opções de trabalho, relaciona-se estreitamente com a seguinte, de maneira que pode parecer que cada uma sempre implicou as outras, quando, de facto, é por um alto grau de adaptabilidade de umas às outras que tantas vezes aparecem juntas.

O objectivo imediato que serve de pretexto para pôr em marcha todo este processo de análise linguística é a «descoberta» de combinatórias lexicais no discurso da Astronomia. Quando falo em *combinatórias*, não me refiro necessariamente a combinatórias fixas – entendo por *combinatória* um *grupo de palavras* que apareça sistematicamente no discurso. Posteriormente desenvolverei mais este ponto.

Para perseguir o meu objectivo era necessário dispor de uma base de dados linguísticos de uma dimensão, actualidade e variedade confortáveis, e, sobretudo, de dados em contexto, ou seja, em que cada palavra a ser analisada estivesse rodeada de um contexto atestado na realidade, de forma a poder dar conta da maneira segundo a qual cada palavra se combinaria com as palavras à sua volta, de modo a formar as combinatórias pretendidas.

Dada esta necessidade, o recurso a um corpus linguístico impunha-se.

2. Corpus ASTRO. Ao longo de dois anos e meio recolhi material linguístico variado, proveniente de fontes variadas, desde língua falada a língua escrita, desde um nível mais especializado a um nível menos especializado, e que abarcasse as diferentes classes temáticas em que a disciplina da Astronomia se pode repartir. Obtive, assim, o corpus ASTRO, um corpus de cerca de 560 000 palavras, repartidas por 54 textos.

Os textos têm dimensões distintas, que vão desde as 225 até às 90 991 palavras. A distribuição percentual entre classes de diferentes dimensões é equilibrada. O corpus ASTRO inclui obras em português, língua original, e português, língua proveniente de tradução, reflectindo, deste modo, a realidade documental e editorial da Astronomia em Portugal nos nossos dias.

Trata-se de um corpus contemporâneo, não incluindo, portanto, obras para lá da década de 60. Uma vez o corpus *on-line*, num PC Pentium, estava pronto para o tratamento informático que se lhe deu em seguida, e que é necessário à consecução das próximas etapas de análise.

3. Tratamento informático. Este tratamento informático consistiu na indexação automática de cada palavra no corpus. Esta indexação foi feita através do programa Visual C++, e consistiu, portanto, na aplicação de uma etiqueta a cada palavra no corpus, etiqueta essa que situava cada palavra no mesmo corpus, ou seja, que lhe atribuía um lugar, de modo a que, nas operações posteriores, se pudesse sempre identificar cada palavra, e apreciar a sua proveniência.

Já indexado, o corpus ASTRO passou a ser gerido pelo programa de gestão de bases de dados ACCESS (versão 2.0). Foi neste programa, que é extremamente flexível e abrangente, que se efectuaram as operações de estatística lexical sobre o corpus ASTRO, consistindo esse tratamento estatístico na próxima fase da análise dos dados.

4. Tratamento estatístico. A estatística linguística e, mais propriamente, a estatística lexical, que é a área que mais nos interessa neste caso, é uma área da estatística em plena expansão, principalmente devido ao

boom dos computadores pessoais e à acessibilidade dos computadores em geral, que tem possibilitado o acesso permanente a grandes quantidades de dados linguísticos, ou seja, a amostragens suficientemente extensas para permitirem uma análise estatística significativa.

A técnica estatística que utilizei para analisar combinatórias lexicais é uma de entre muitas actualmente disponíveis. Trata-se da técnica da *informação mútua* – ou *mutual information* – uma das técnicas de utilização possível dentro do ramo da estatística lexical que lida com a *associação de palavras* – ou *word-association* – refiro por vezes a designação inglesa, pois trata-se de técnicas que se vão ancorar na línguística anglo-saxónica.

A informação mútua é uma técnica que analisa pares de palavras. É tomada em conta uma determinada palavra – o **nó** – e analisada a relação que essa palavra estabelece com palavras à sua volta, a dada distância (a **amplitude**). A probabilidade de ocorrência das duas palavras a essa determinada distância é, então, comparada com a probabilidade de ocorrência independente de cada uma das duas palavras no corpus. Obtém-se, assim, o que se designa por *Índice de Combinação*, índice este cuja interpretação poderá ilustrar a força da combinação das duas palavras.

Para recolher combinatórias lexicais, normalmente formadas por duas ou três palavras, esta técnica tem funcionado com bastante eficácia. Trata-se, também, de uma técnica que exprime na prática aquilo que é um conceito de pressupostos teóricos – a Abordagem Lexical.

Resumindo muito brevemente, pois esta comunicação não pretende tratar a questão da Abordagem Lexical, esta última é uma aproximação à língua e ao léxico que parte do uso para chegar às regras, que encara a língua não como uma estrutura-base preenchida pelo léxico, mas como *pedaços* de léxico, junções sistemáticas de palavras – ou **combinatórias** – que se relacionam entre si e se entreligam, formando, portanto, o tecido linguístico (a língua).

A técnica estatística da informação mútua, na medida em que contribui para isolar tais *pedaços* de léxico – as combinatórias – vai ao encontro da Abordagem Lexical, sendo esta, aliás, uma abordagem linguística que muito tem beneficiado das disciplinas de Corpus Linguístico e Estatística Lexical. Assim nos deparamos, portanto, com a intercomunicação entre as várias componentes deste processo de análise,

referida no ponto 1. A união destas três disciplinas tem contribuído para que se venha a encarar o léxico de um ponto de vista mais organizacional, e menos linear.

Segue-se um exemplo do tipo de análise que se pode fazer tirando do partido destas três disciplinas – Corpus Linguístico, Abordagem Lexical e Estatística Lexical, através da técnica da Informação Mútua.

5. Análise. Observe-se a Figura I. Nela se encontra uma **lista de concordâncias** da palavra ***nebulosa***[1]. Esta última é, portanto, o **nó**, sendo as palavras que a cercam os **co-ocorrentes**. Cada linha da lista designa-se por **janela**, tendo neste caso cada janela a **amplitude** 10, ou seja, há 5 palavras de cada lado do nó. Esta lista está ordenada alfabeticamente pela palavra +1, ou seja, a primeira palavra à direita do nó. É neste lugar que se encontrará a maior parte dos co-ocorrentes privilegiados do nó que formarão com ele combinatórias lexicais significativas de duas palavras.[2] Na coluna à esquerda surge o nome do texto do corpus ASTRO de onde foi extraída cada janela, o que constitui uma informação interessante do ponto de vista da *repartição*, de modo a que não se considere como facto linguístico uma observação que apareça apenas num só texto, sem uma maior distribuição no corpus.

Como se pode ver, apenas com a consideração da lista de concordâncias devidamente ordenada há observações linguísticas interessantes a fazer. Assim, poder-se ia apontar como possíveis pares combinatórios para a palavra **nebulosa** as palavras **anular**, **brilhante**, **esférica**, **original**, **planetária**, **primitiva**, **protoplanetária**. Estas palavras formariam, portanto, com **nebulosa**, uma *combinatória lexical*. A Figura I mostra-nos, também, que as preposições **de** e **do** parecem introduzir mais elementos combinatórios, dando origem a combinatórias lexicais de três elementos.

Atente-se agora na Figura II. Esta dá-nos uma visão quantitativa e estatística do que observámos aleatoriamente na Figura I. Na 1ª coluna (**Palavra:**) temos as palavras co-ocorrentes do nó (**nebulosa**), ordenadas pela frequência de aparição na posição +1; a 2ª coluna (**f(x, y)**) mostra a frequência conjunta no corpus ASTRO, numa janela de amplitude 10, da combinação da palavra na primeira coluna (designada quantitativamente por **y**) com o nó (designado quantitativamente por **x**); as dez colunas seguintes ilustram a distribuição da palavra y nas várias posições da janela; a 13ª coluna (**m**) revela a média estatística[3]; a 14ª

(v), a variância[4], a 15ª (**fx**) a frequência do nó no corpus; a 16ª (**fy**) a frequência da palavra co-ocorrente no corpus, a 17ª (**IC**) o valor do Índice de Combinação; a 18ª (**IF**) o valor do Índice de Fixidez[5].

As palavras que maior frequência exibem na posição +1 são, respectivamente, **do** e **de**. No entanto, distribuem-se também em quase todas as outras posições da janela (embora com valores menores), daí a sua alta variância. Por outro lado, embora se conjuguem frequentemente com a palavra x (por ex., **do** f(x, y)=45), também têm uma alta frequência individual no corpus (por ex., **do** fy=9575), o que lhes confere um IC relativamente baixo. De facto, análises prévias revelam-nos que, como elementos de relação que são, estas preposições apresentam normalmente ICs baixos, entre o 3 e o 5.

Com altos ICs encontramos, por exemplo, as palavras **protoplanetária, planetária, primitiva** e **anular**. Haveria que analisar os dados da repartição, a lista de concordâncias completa, bem como considerar a distribuição total destas palavras em torno do nó **nebulosa**, mas, com um IC característico e com o padrão combinatório exibido (média quase sempre de 1, variância próxima do 0, etc.), tudo parece apontar para combinatórias lexicais privilegiadas. Quanto à sua fixidez, **planetária** e **anular** destacam-se das restantes pelo seu alto nível de IF. Poderiam, portanto, ser apontadas como combinatórias fixas.

Observe-se, no entanto, o caso de **nebulosa protoplanetária** – porque apresenta esta combinatória um IF de apenas 4.76? E, se a frequência da palavra y **protoplanetária** no corpus é de 13, como pode a f(x, y) ser igual a 14? O que se passa é que, das 13 vezes em que **protoplanetária** aparece no corpus, numa delas esta palavra combina-se com duas palavras **nebulosa** diferentes. Para uma delas, **protoplanetária** coloca-se na posição +1 (como nas outras 12 ocorrências), para a outra, a mesma ocorrência aparece também na posição -5... Eis a razão pela qual alguns dos valores estatísticos apresentados pela combinatória **nebulosa protoplanetária** pareciam tão estranhos em relação àqueles exibidos pelas outras combinatórias privilegiadas.

6. Conclusão. Não julgo ser necessário explicar de que modo este tipo de análise vem auxiliar o trabalho do lexicólogo em particular, do linguista em geral. Acresce dizer que, assim como alguns níveis de IC revelam diferentes comportamentos lexicais, também outros revelarão escolhas de índole sintáctica e semântica.

A conclusão que parece sobressair deste tipo de análise é a de que ela é certamente uma das maneiras mais eficazes e produtivas de explorar um corpus linguístico extenso. De qualquer forma, resta lembrar a importância de uma interpretação perspicaz dos dados quantitativos, apoiada em sólidos conhecimentos linguísticos, sob pena de o exercício estatístico não revelar todas as suas possibilidades e/ou limites, não se conseguindo mais do que leituras superficiais dos dados.

Palavra nó: nebulosa

max	com um telescópio a famosa	nebulosa	anular m 57 a via
sovietic	é também muito conhecida a	nebulosa	anular na constelação da lira
nebulosa	instrumento de média potência esta	nebulosa	apresenta-se com o aspecto de
nebulosa	na constelação da lira esta	nebulosa	apresenta-se como um anel de
sovietic	termo de comparação que a	nebulosa	brilhante conhecida na constelação
nebulosa	notam bastante bem sobre uma	nebulosa	brilhante e às quais j
sovietic	a nossa galáxia e a	nebulosa	de andrómeda diferentemente das gal
nebulosa	nebulosas à vista desarmada a	nebulosa	de andrómeda a nebulosa do
nebulosa	de um estudo exaustivo da	nebulosa	de andrómeda e de outras
nebulosa	é o facto de a	nebulosa	do caranguejo ser constituída por
sovietic	fulguração de uma supernova a	nebulosa	do caranguejo tem uma série
max	dessa explosão na região denominada	nebulosa	do caranguejo termo atribuído por
sovietic	de movimento desta porção de	nebulosa	esférica <eq> permanecer constante
sovietic	massa m de uma dada	nebulosa	esférica está animada de movimento
sovietic	de 1 km/s portanto a	nebulosa	original a partir da qual
sovietic	esta razão de facto na	nebulosa	original antes do início da
sovietic	de energia irradiada por esta	nebulosa	planetária é dezenas de vezes
ligacoes	luminosas que se conhecem uma	nebulosa	planetária é uma nuvem em
sovietic	se encontra na fase de	nebulosa	planetária liberta-se dum invólucro
nebulosa	a rodeia nasce assim a	nebulosa	planetária propriamente dita na jan
nebulosa	uma parte do gás da	nebulosa	planetária que continua a sua
meteor	poderia ser explicada considerando	nebulosa	primitiva com uma estrutura heterog
meteor	acordo com estas hipóteses a	nebulosa	primitiva não seria isotropicamente
jupiter	da que se atribui à	nebulosa	primitiva que teria dado origem
jupiter	ao de uma estrela na	nebulosa	primitiva que teria dado origem
sovietic	transmitem o momento dentro da	nebulosa	protoplanetária 4 o período seguint
sovietic	entre <num> o disco da	nebulosa	protoplanetária adquire talvez uma
sovietic	gigantes a poeira que constituía	nebulosa	protoplanetária concentra-se cada v

Figura 1 – Lista de concordâncias.

Palavra Nó: nebulosa

Palavra:	f(x, y)	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	m	v	fx:	fy:	IC:	IF:
do	45	1	1	1	0	0	30	0	3	2	7	1.556	4.558	162	9575	4.00	0.86
de	100	10	10	9	10	3	29	2	8	9	10	0.030	10.369	162	27924	3.60	0.34
protoplanetária	14	1	0	0	0	0	13	0	0	0	0	0.571	2.388	162	13	11.84	4.76
planetária	7	0	0	0	0	0	7	0	0	0	0	1.000	0.000	162	25	9.89	98.93
primitiva	5	0	0	0	0	0	4	1	0	0	0	1.200	0.160	162	24	9.47	36.41
é	27	2	4	2	5	0	4	2	5	2	1	-0.222	9.802	162	5474	4.07	0.41
que	47	5	9	3	7	0	4	6	4	6	3	-0.362	11.763	162	14469	3.46	0.29
original	3	0	0	0	0	0	3	0	0	0	0	1.000	0.000	162	42	7.92	79.22
da	80	5	2	4	1	57	3	1	0	3	4	-0.838	4.411	162	9850	4.79	1.06
em	23	2	3	2	0	0	3	1	6	3	3	0.957	12.129	162	7039	3.47	0.28
se	15	2	0	2	0	0	3	2	5	0	1	0.733	9.396	162	5492	3.21	0.34
anular	2	0	0	0	0	0	2	0	0	0	0	1.000	0.000	162	8	9.73	97.29
apresenta-se	2	0	0	0	0	0	2	0	0	0	0	1.000	0.000	162	25	8.09	80.85
esférica	2	0	0	0	0	0	2	0	0	0	0	1.000	0.000	162	41	7.37	73.72
brilhante	3	1	0	0	0	0	2	0	0	0	0	-1.000	8.000	162	208	5.61	0.69

Figura II – Padrão combinatório.

Notas

[1] Note-se que se trata apenas de uma amostragem não contínua da lista de frequências completa.

[2] Entenda-se que a lista de concordâncias pode ser ordenada por qualquer outra posição da janela (+2, +3, -1, -2, etc...), consoante os objectivos imediatos da análise.

[3] Medida estatística de Localização; quanto mais próxima de 1, mais indica que a palavra y se organiza regularmente em torno da palavra x.

[4] Medida estatística de Dispersão; quanto mais próxima de 0, mais indica que a palavra y tem tendência para se fixar sempre na mesma posição em relação à palavra x.

[5] Índice introduzido por SANTOS PEREIRA (1994), em que se compara a mobilidade da palavra y no corpus com o seu IC com a palavra x. O IF só deve considerar-se na sua relação com o IC, e isto para auxiliar na destrinça entre altos ICs muito similares.

Referências

CHURCH, Kenneth W. e Patrick HANKS. 1990. Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, 16--1, Cambridge, MA: ACL, pp.22-29.

CHURCH, Kenneth, W. GALE, P. HANKS e D. HINDLE. 1991. Using Statistics in Lexical Analysis, *in* Ed. Uri ZERNIK, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, New Jersey, Hove and London: Lawrence Erlbaum Associates.

GALVÃO DE MELLO, F. 1971. *Introdução aos métodos estatísticos*, Vol. I, Lisboa: Cadernos do Instituto de Orientação Profissional.

LEWIS, Michael. 1993. *The Lexical Approach*, London: LTP.

MARQUEZ NETO, Paula. 1995. *Combinatórias Lexicais no Discurso da Astronomia – Um Estudo em Estatística Lexical*, Dissertação de Mestrado em Linguística Portuguesa Descritiva apresentada à FLUL, Lisboa: n/publ.

MULLER, Charles. 1968. *Initiation à la statistique linguistique*, Paris: Larousse.

SANTOS PEREIRA, Luísa Alice. 1994. *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, Dissertação de Mestrado em Linguística Portuguesa Descritiva apresentada à FLUL, Lisboa: n/publ.

SINCLAIR, Jonh. 1991. *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.