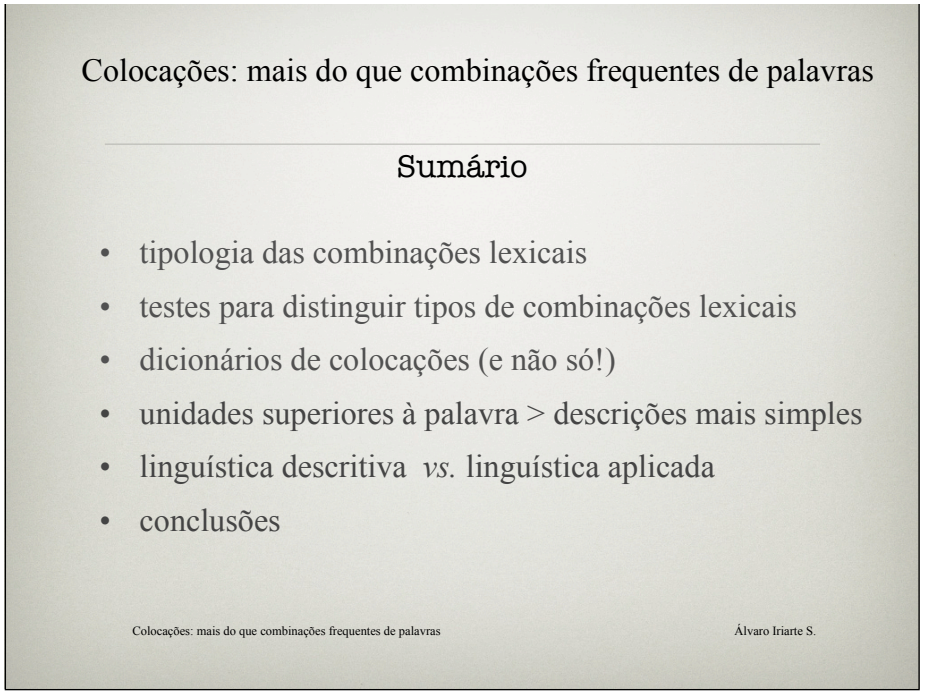
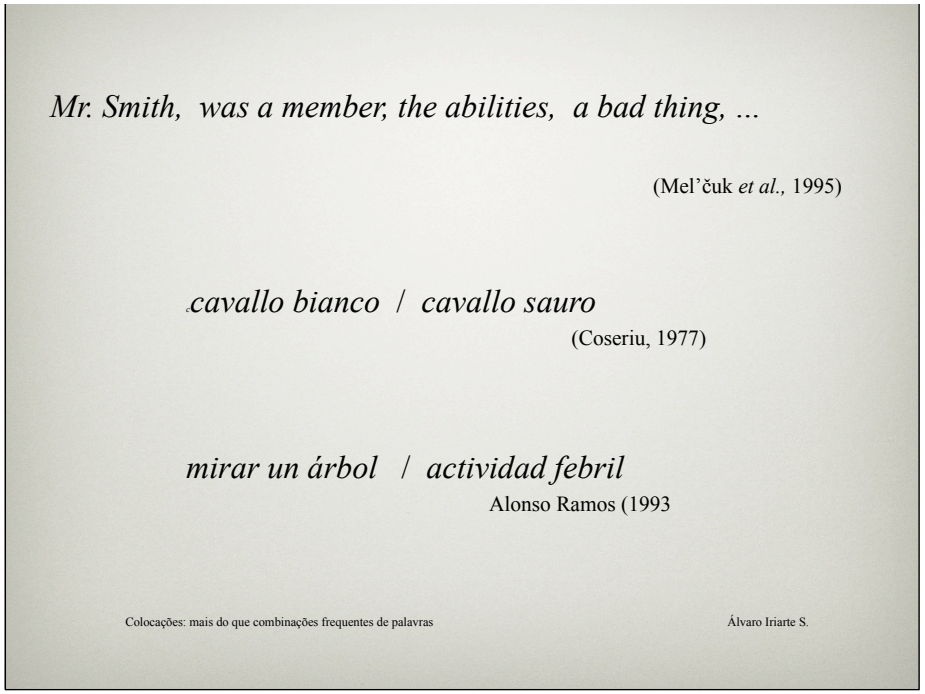
Colocações: mais do que combinações frequentes de palavras

Álvaro Iriarte S.

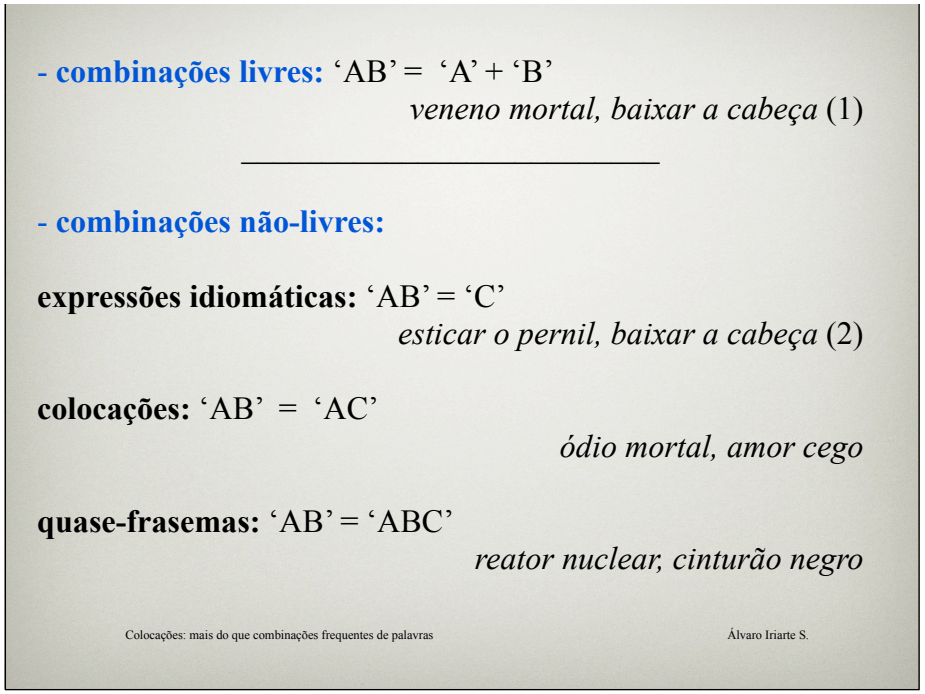


conclusões

Durante anos aborreci os meus amigos informáticos sobre as limitações dos modelos estatísticos de análise e descrição da combinatória lexical ...



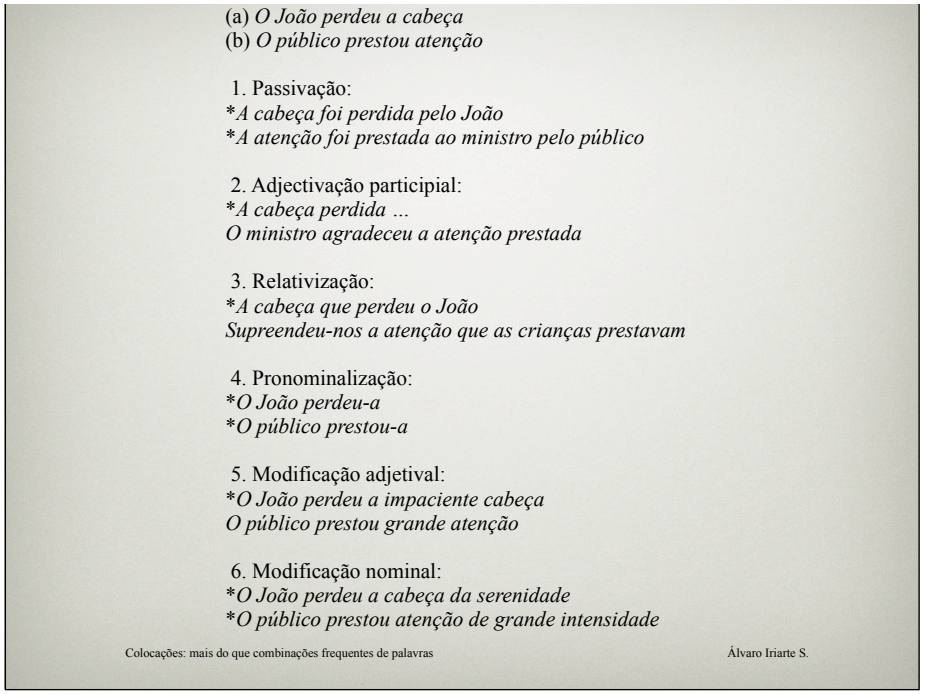
Os linguistas sempre contestaram a concepção das colocações como sendo meras combinações frequentes de palavras
Mr. Smith , etc. são combinações frequentes, mas não colocações (nem qualquer outro tipo de combinação lexical restrita (ou "não livre")
A combinação em italiano "cavallo bianco" é uma combinação frequente porque refere coisas frequentes.
O mesmo acontece com o 1º exemplo em espanhol.
Mas "cavallo sauro" (cavalo baio, alazão) e "actividad febril" são combinações lexicais frequentes porque os elementos que as conformam aparecem juntos frequentemente, independentemente de a realidade que referem ser frequente ou não.



Já há alguns anos que uso este quadro para explicar aos meus alunos a diferença entre combinações lexicais livres e combinações lexicais restritas.

AB = A+C:
não são totalmente livres porque algum tipo de restrição atua na altura de se combinarem os seus componentes: ódio seleciona mortal para significar “em alto grau” e amor, por exemplo, seleciona cego para significar a mesma coisa.
não são totalmente composicionais no sentido de que, de alguma maneira, o seu significado não é igual à soma do significado dos seus componentes:
em ódio mortal, mortal não significa “letal” (como em ferida mortal ou arma mortal) mas significa “muito; intensamente; em grande medida”
AB = ABC
em cinturão negro, encontramos o sentido de 'cinto' e de 'negro' mais um sentido aproximado de 'grau de conhecimento ou habilidade em artes marciais'.

Tipologia que Igor Mel’ čuk (1995) estabelece, dentro da “Teoria Sentido-Texto”, para o *Dictionnaire Explicatif et Combinatoire du Français Contemporain* (DEC),



A linguística teórica (e alguma prática lexicográfica e terminográfica) foi estabelecendo uma série de testes, baseados em critérios morfo-sintácticos, para ajudar a delimitar o segmento de enunciado que corresponde a uma unidade pluriverbal (a uma combinação “não livre” de palavras) face a outras combinações livres de palavras.
Ou para distinguir tipos de combinações “não livres” de palavras. Por exemplo, o frasema perder a cabeça (a) apresenta maiores restrições sintáticas do que a colocação prestar atenção (b) que já admite algumas transformações.

7. Modificação adverbial:
**O João perdeu a cabeça intensamente*
O público prestou atenção ininterruptamente

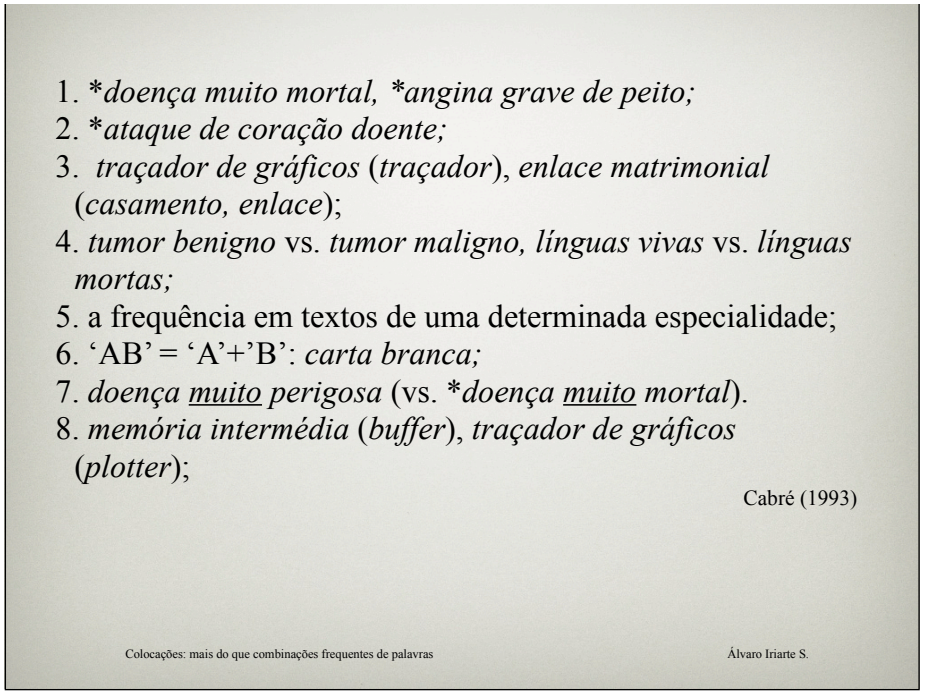
8. Determinação:
**O João perdeu aquela cabeça*
**O público prestou aquela atenção*

9. Quantificação:
**O João perdeu muito a cabeça*
O público prestou muita atenção

10. Indefinição:
**O João perdeu uma cabeça*
**O público prestou uma atenção*

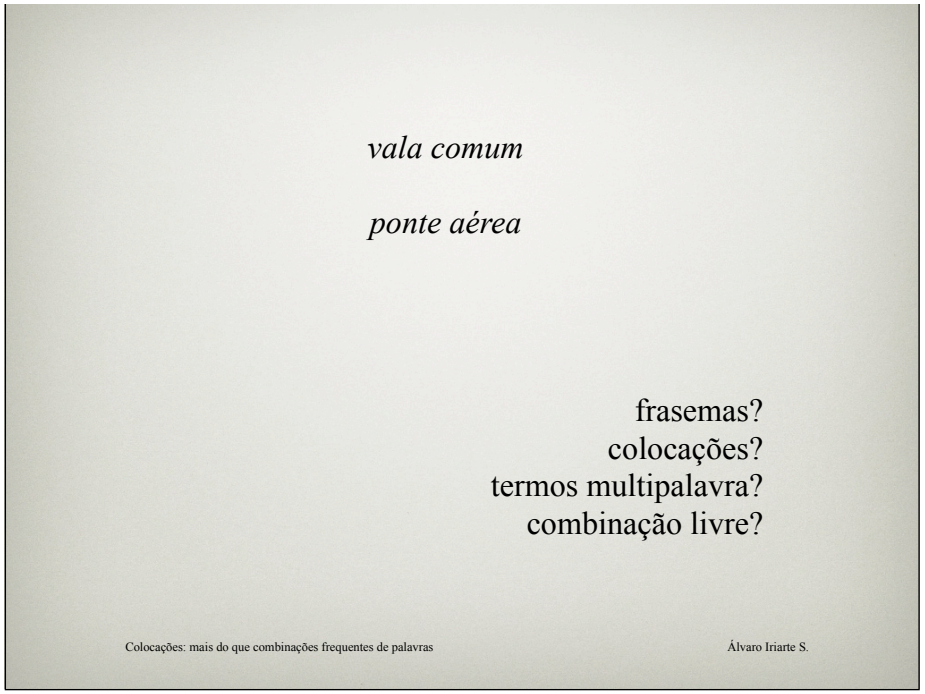
11 Pluralização:
**O João perdeu as cabeças*
**O público prestou atenções*

12. Presença/ausência de artigo:
**O João perdeu cabeça*
**O público prestou a atenção*
(embora: *O público prestou a devida atenção*).



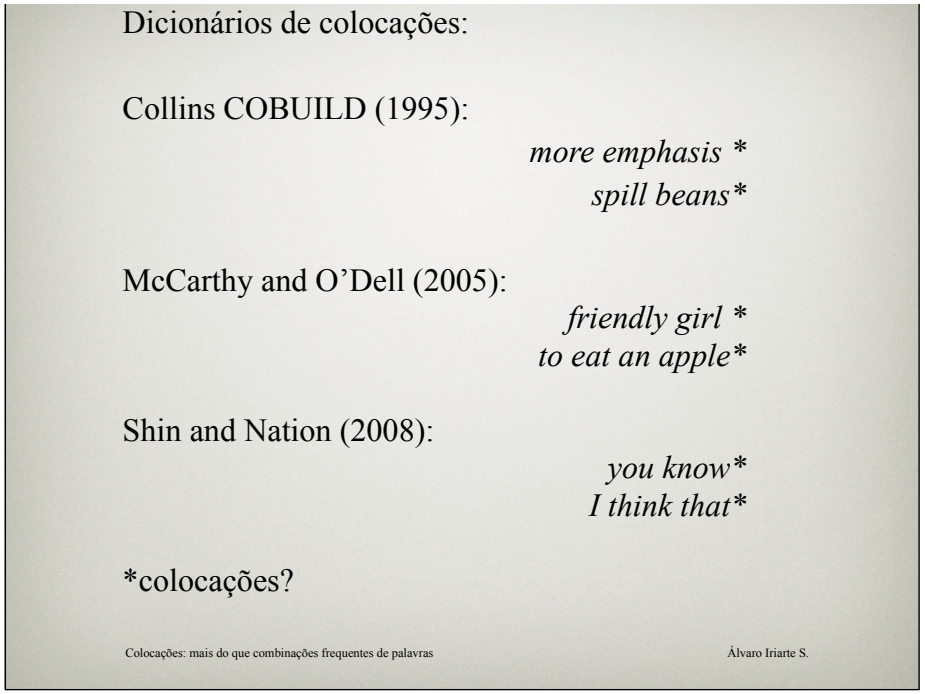
Também a prática terminográfica foi estabelecendo uma série de testes para ajudar a delimitar o segmento de enunciado que corresponde a um termo (unidade terminológica pluriverbal) face a outras combinações livres de termos.

1 a impossibilidade de inserir outros elementos linguísticos no interior do sintagma terminológico: **doença muito mortal, *angina grave de peito;*
2 o facto de não se poder complementar separadamente nenhuma das partes do conjunto: **ataque de coração doente;*
3 o facto de poder substituir o conjunto por um sinónimo: *traçador de gráficos* (*traçador*), *enlace matrimonial* (*casamento, enlace*);
4 o facto de possuir um antónimo na mesma especialidade: *tumor benigno* vs. *tumor maligno, línguas vivas* vs. *línguas mortas;*
5 a frequência de aparição do mesmo sintagma terminológico em textos de uma determinada especialidade;
6 o facto de o significado do conjunto não poder ser deduzido do significado dos elementos que o formam: *carta branca;*
7 de modo complementar, a presença de determinadas unidades linguísticas no interior do sintagma revela que muito provavelmente se trata de uma combinação livre: *doença muito perigosa* (vs. **doença muito mortal*).
8 o facto de que noutras línguas o sintagma corresponda a uma única unidade lexical: *memória intermédia* (*buffer*), *traçador de gráficos* (*plotter*);
Cabré (1993)



Mas a própria autora acaba por reconhecer que o grau de rendimento de cada um destes testes é variável, (Cabré, 1993: 304)

A fronteira entre colocações, expressões idiomáticas, combinações livres de palavras, nomes compostos ou termos pluriverbais de linguagens de especialidade é difusa. E na prática, isto é, nos produtos que encontramos no mercado, os resultados fiquem muito aquém da teoria.



Alguns exemplos:
Dicionários com uma concepção de colocação exclusivamente ou principalmente estatística, em que o critério único ou principal é a frequência (e não a arbitrariedade combinatória) apresentam resultados como estes:

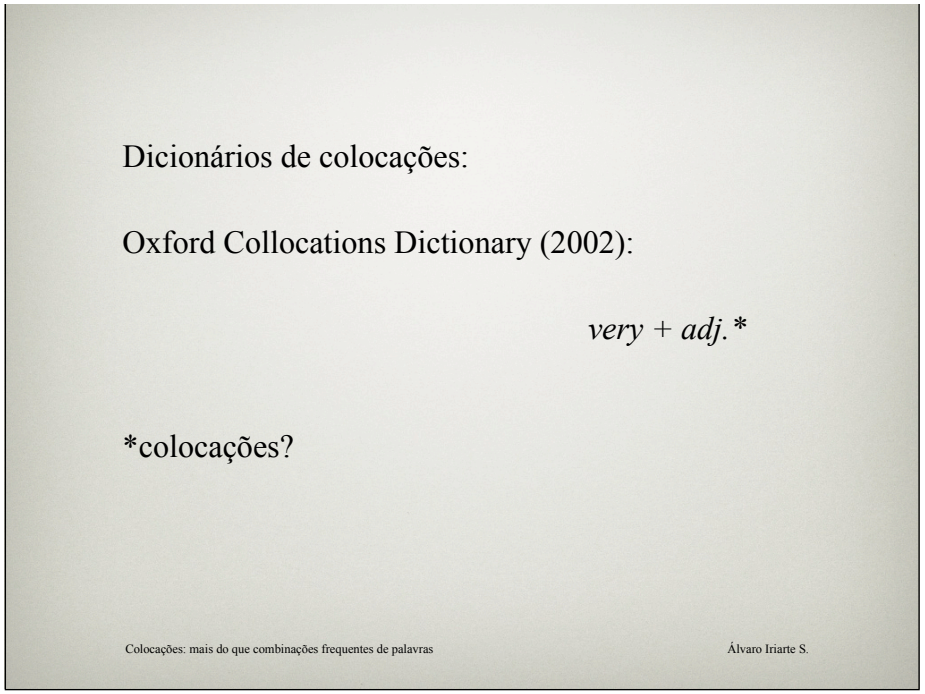
Collins COBUILD (1995):
more emphasis (combinção livre)
spill beans (expr. idiomática)

McCarthy and O'Dell (2005):
friendly girl (combinção livre)
to eat an apple (combinção livre)

Shin and Nation (2008)
you know (combinção livre) *
I think that (combinção livre)

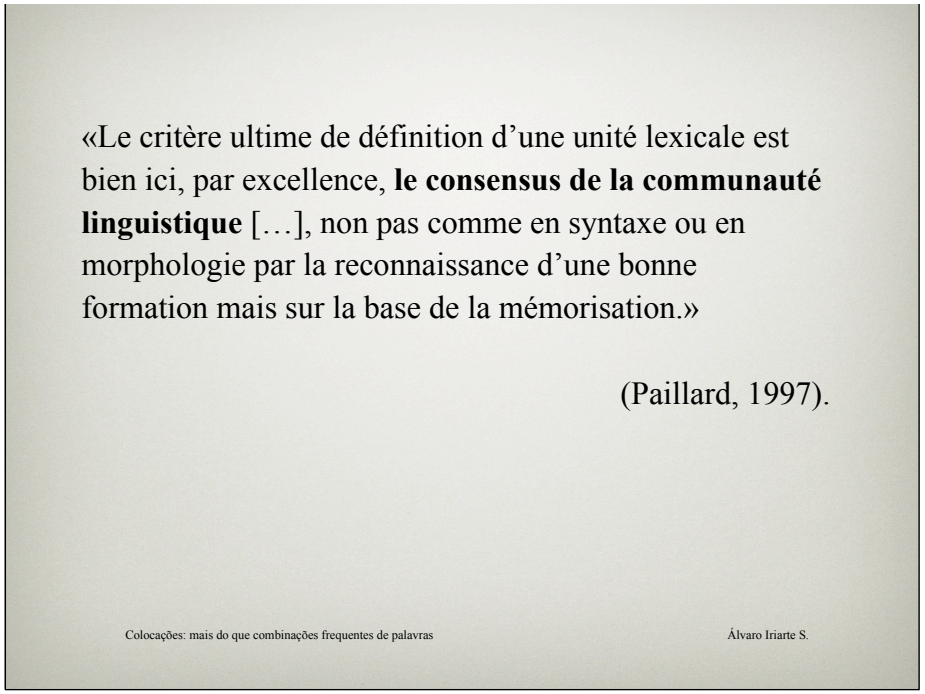
- Mas pense-se em formas como I Know, you know, I see, you see, estou a ver, eu sei, ya sé, ya veo, que claramente não são usos livres (pragmatemas)

(Moreno, 2009)



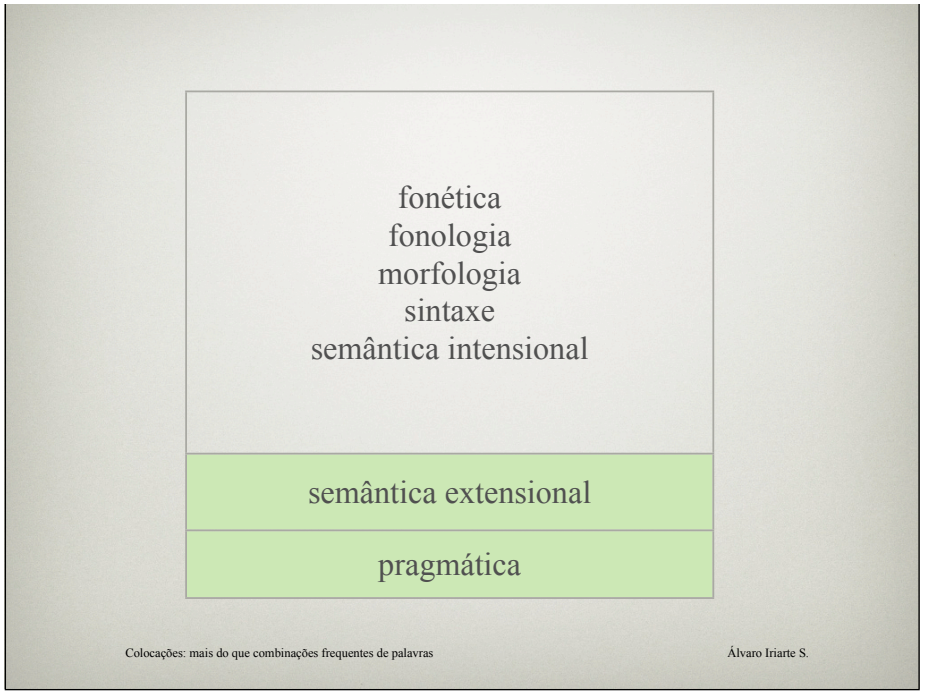
também dicionários com uma concepção de colocação não apenas estatística (frequência e institucionalização), em que também foram considerados critérios de tipo semântico e fraseológico (idiosincrasia, arbitrariedade da língua, restrições combinatórias arbitrárias)

(Moreno, 2009)



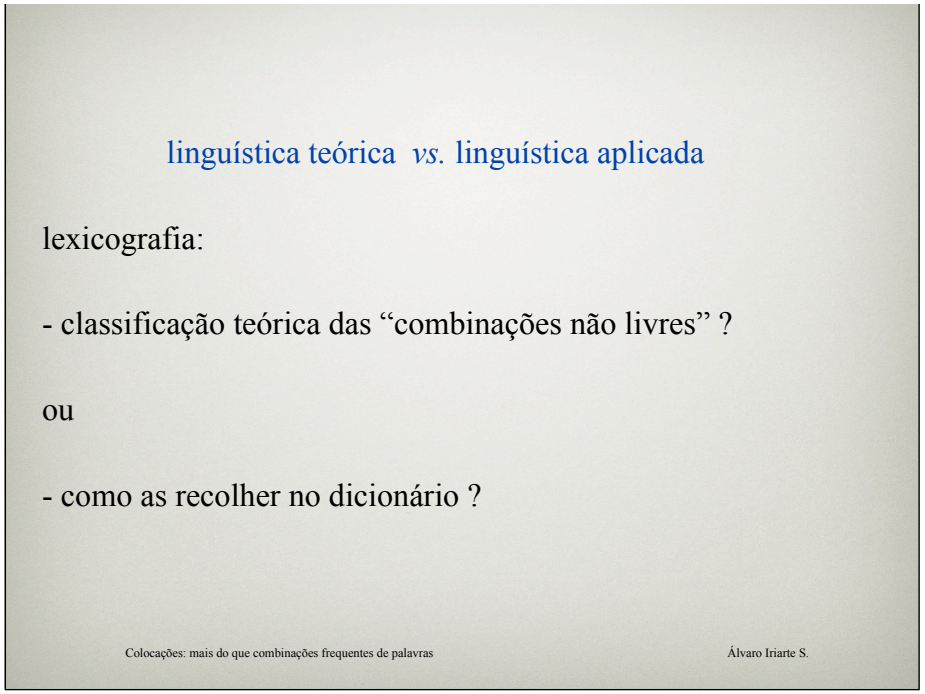
Não podemos deixar de duvidar da economia e até da eficácia das tentativas de identificação e classificação dos diferentes tipos de "unidades linguísticas compostas" a partir de possíveis marcas formais
Do ponto de vista lexicográfico e terminológico, assumimos sem reservas que os critérios que nos permitirão considerar se um termo pluriverbal foi lexicalizado não poderão ser de tipo morfo-sintáctico, tendo mais a ver com o consenso e a com memória da comunidade linguística que o utiliza (Paillard, 1997: 66).

Por falar em consenso: lembro que um aspeto importante na elaboração de **ontologias** é a questão da "informação usada e validada por uma determinada comunidade"



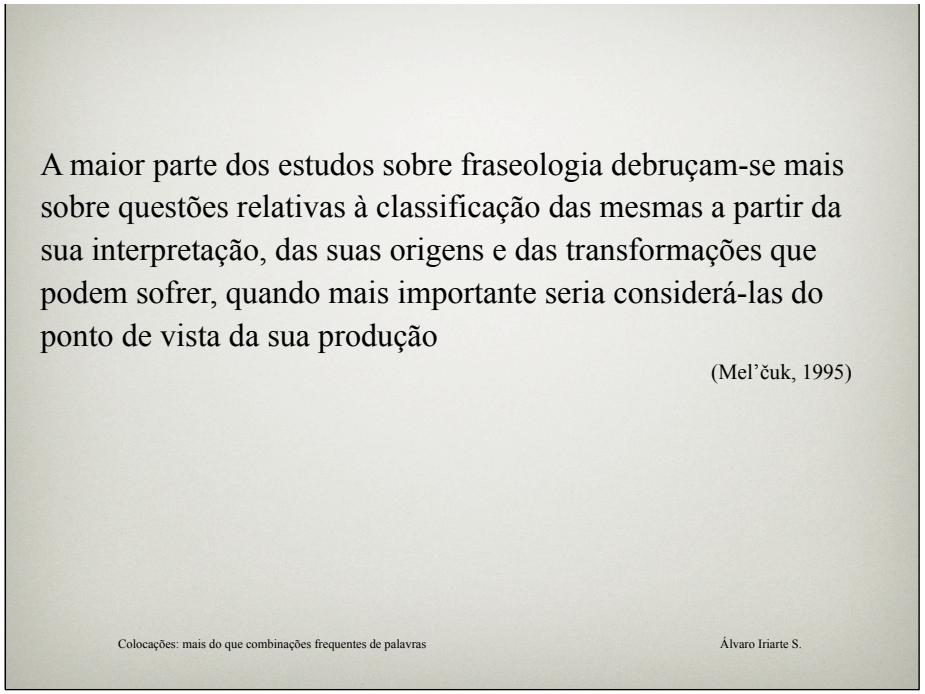
Recupero aqui um diapositivo que utilizei no workshop do Per-Fide do ano passado
Esse consenso da comunidade, essa memorização situa-se claramente dentro da semântica-extensional y da pragmática dos estudos semióticos (Morris, 1985), “aquela parte da língua que está “profundamente ligada ao conhecimento do mundo “ (e que uma parte da linguística excluiu durante muito tempo do seu objeto de estudo).

(no domínio sintático as estruturas são imanentes, enquanto no léxico é impossível um estudo exclusivamente linguístico porque está ligado aio conhecimento do mundo e dessa maneira a outras disciplinas como a psicologia, a história, a etnologia, etc.)

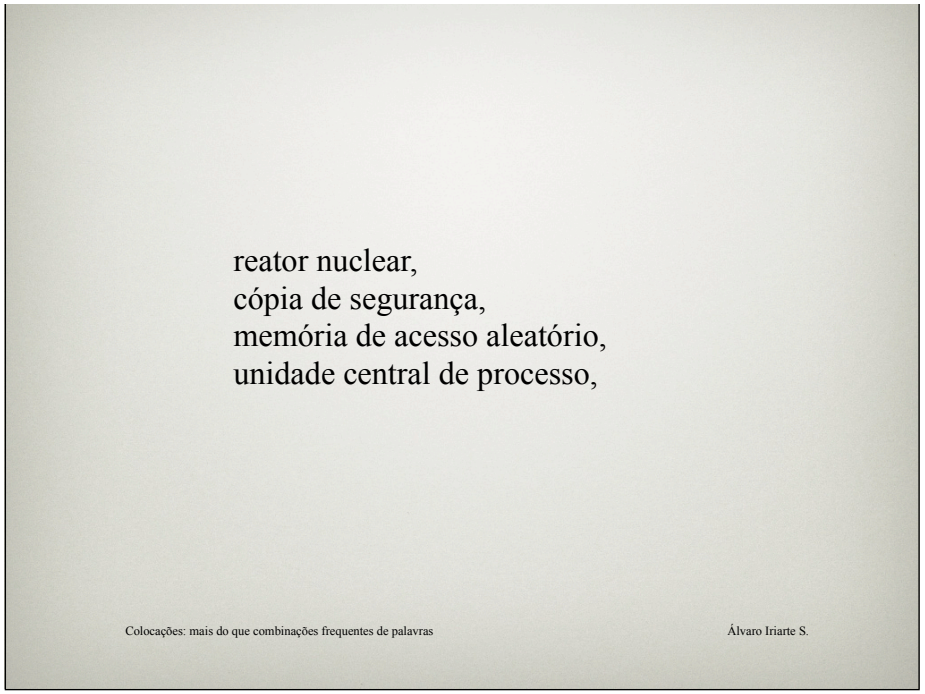


Um exemplo: a lexicografia
O objetivo da lexicografia não é tanto tentar uma classificação teórica das "combinações não livres" como encontrar a melhor maneira de as consignar no dicionário
A classificação dos fenómenos fraseológicos é um problema para a lexicologia e a linguística teórica em geral. O problema para a lexicografia será o da seleção das combinações lexicais que devem aparecer no dicionário, isto é, que tipo de combinações lexicais devem ser consideradas como unidades lexicográficas ou terminológicas.

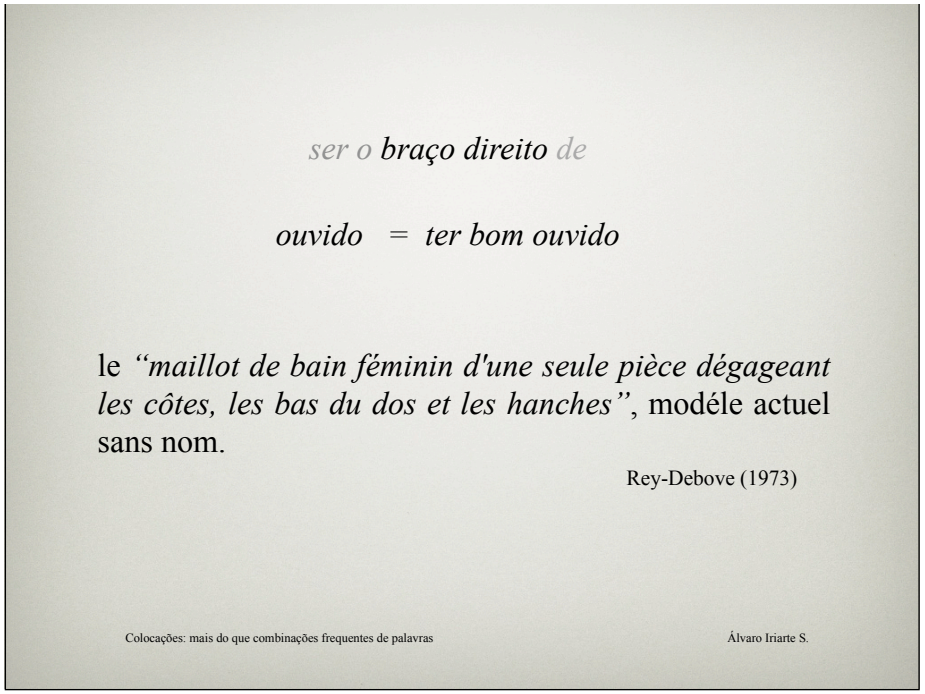
Noutras áreas da linguística aplicada a preocupação será: como ensinar estas combinações , quais são os equivalentes destas combinações noutras línguas, etc.



Do ponto de vista da linguística aplicada (lexicografia, terminologia, ensino de LE, tradução, etc.), o que interessa, mais do que as classificações que a lexicologia possa fazer de determinadas combinações lexicais, é como fazer a inventariação, o tratamento e a recuperação de toda a informação relativa às combinações lexicais que não possam ser traduzidas palavra por palavra, de tal modo que o utilizador saiba como utilizá-las no discurso

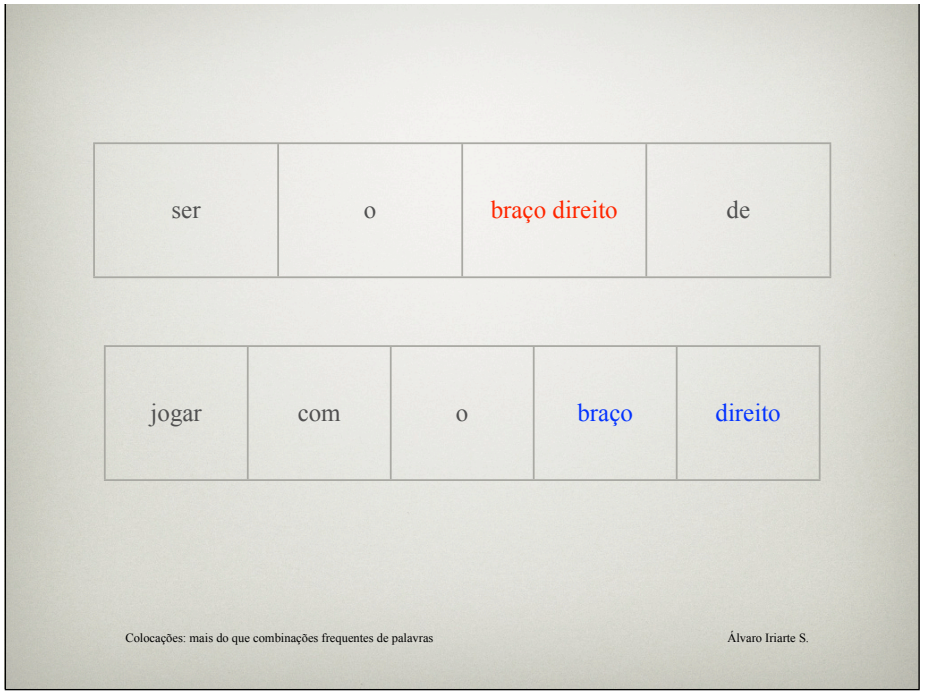


o desafio agora é saber quais os limites superiores destas unidades.
Em vez de apontar para unidades como a palavra ou até inferiores à palavra, deveremos prestar especial atenção às combinações de palavras e tentar resolver o problema de se devem ser consideradas como entradas lexicais independentes ou não.
Ou, doutra maneira, quando é que se pode considerar que uma combinação lexical foi lexicalizada ou habitualizada?
Porque a dificuldade está em que estruturas como as colocações obedecem muito frequentemente às mesmas regras combinatórias que regem as combinações totalmente livres.



Na lexicografia e na terminologia, o maior problema que se nos coloca ao trabalhar com unidades pluriverbais, com combinações lexicias, é o de estabelecer os limites superiores destas unidades, que, por vezes, pode chegar inclusive a coincidir com a sua própria descrição ou é difícil distingui-las
Já falei muitas vezes do exemplo de "ter bom ouvido" e de como os dicionários apresentam como aceção de uma palavra o que, em rigor, é o significado dessa palavra quando combinada com outras ("ter bom ouvido", como aceção 3 de ouvido, no Dic. Aurélio, por exemplo).

"maillot d..." Quando ontem se falava do problema da falta de equivalentes e da necessidade de utilizar uma descrição, uma paráfrase, do termo em questão lembrei-me deste exemplo dos anos 70 de como, em último extremo, o termo poderia chegar a coincidir com a sua definição.
(hoje às 9 mostraram-nos outro exemplo)



Mas o problema à hora de trabalhar como unidade superiores à palavra é que a unidade de análise e descrição lexical não poderá ser qualquer fragmento delimitado aleatoriamente. Deverá ter, na medida do possível, um mínimo de autonomia estrutural que permita distingui-la como unidade. Assim, por exemplo, num enunciado como ser o braço direito de (‘ser o principal auxiliar de’) a unidade a considerar não deveria ser, contrariamente à prática lexicográfica habitual, a totalidade do sintagma, mas sim (o) braço direito, uma vez que ser + o + braço direito + de é uma estrutura gramatical perfeitamente regular e transparente, construída segundo as regras da gramática portuguesa.
Mas já vimos como os testes, de tipo formal, que supostamente permitiriam delimitar o segmento de enunciado que corresponde a uma unidade pluriverbal, face a outras combinações livres de palavras, não funcionam

granularidade

(JJ: " escala *vs.* precisão" ?)

O meu gato de **estimação** O meu **gato de estimação**

Mi gato de **compañía** Mi **gato de compañía**

My **pet** cat My **pet cat**

aborto:
1. miscarriage (espontâneo). 2. abortion (provocado)

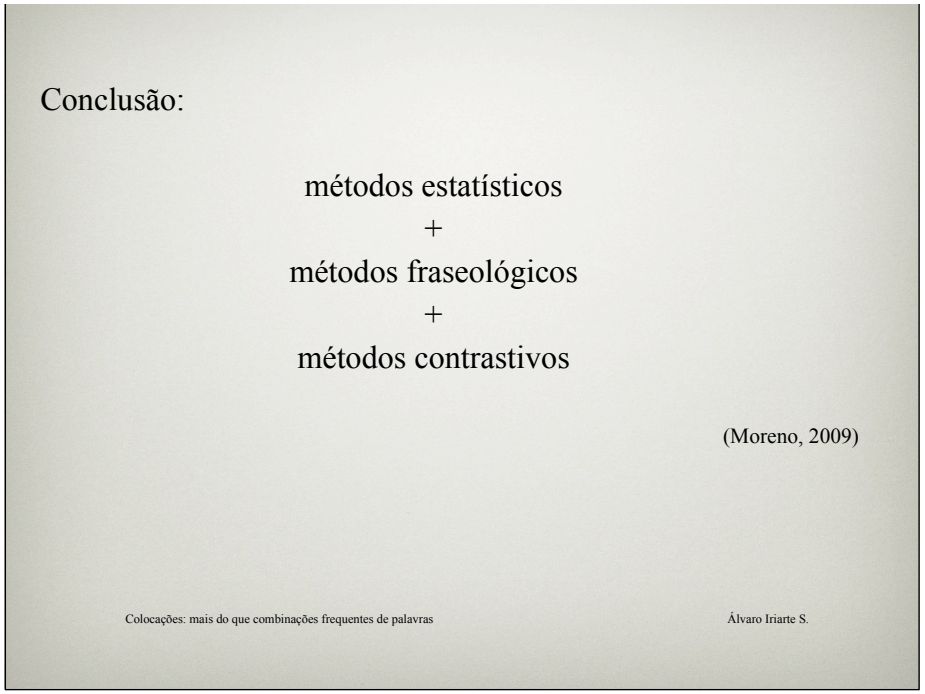
ter um aborto: to have a miscarriage
fazer um aborto: to have an abortion

Colocações: mais do que combinações frequentes de palavras Álvaro Iriarte S.

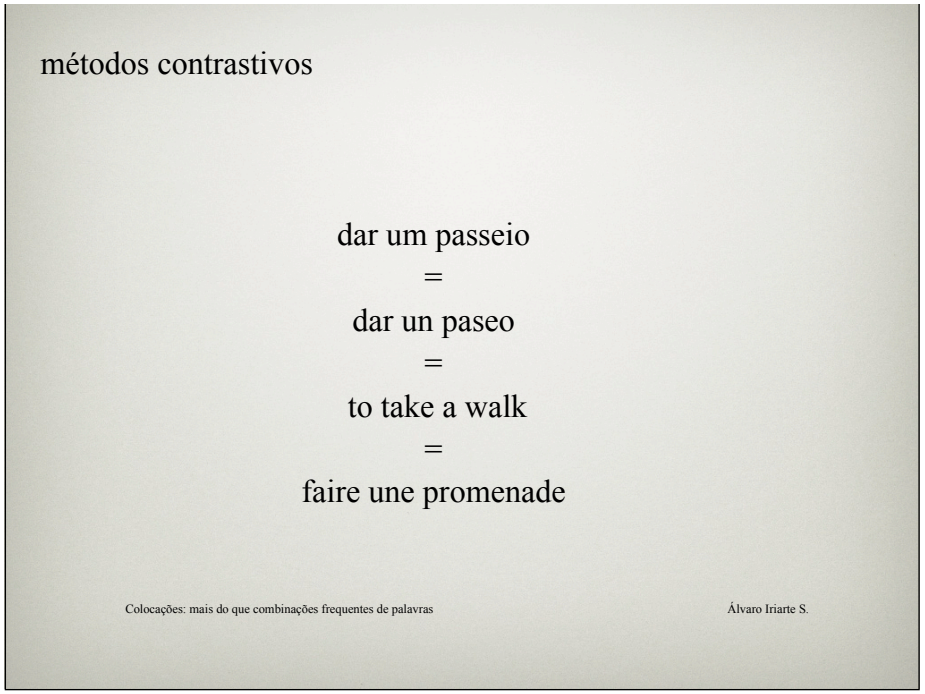
recolho novamente um exemplo do ano Workshop Per-fide do ano passado
para chegar a atenção para uma questão que levantou o Eng. José João ontem de manhã:
"ser muito minucioso, ir muito ao detalhe, pode ser contraproducente"
> optar por unidades superiores à "palavra" => descrições mais simples

Seria mais difícil uma descrição lexicográfica (em termos de aceções e subaceções) dos valores de *estimação* e *companhia* nestes exemplos do que se tomarmos a combinação "gato de estimação" ou "gato de compañía"

O mesmo acontece com as duas aceções da palavra aborto num dic. bilingue de português inglês

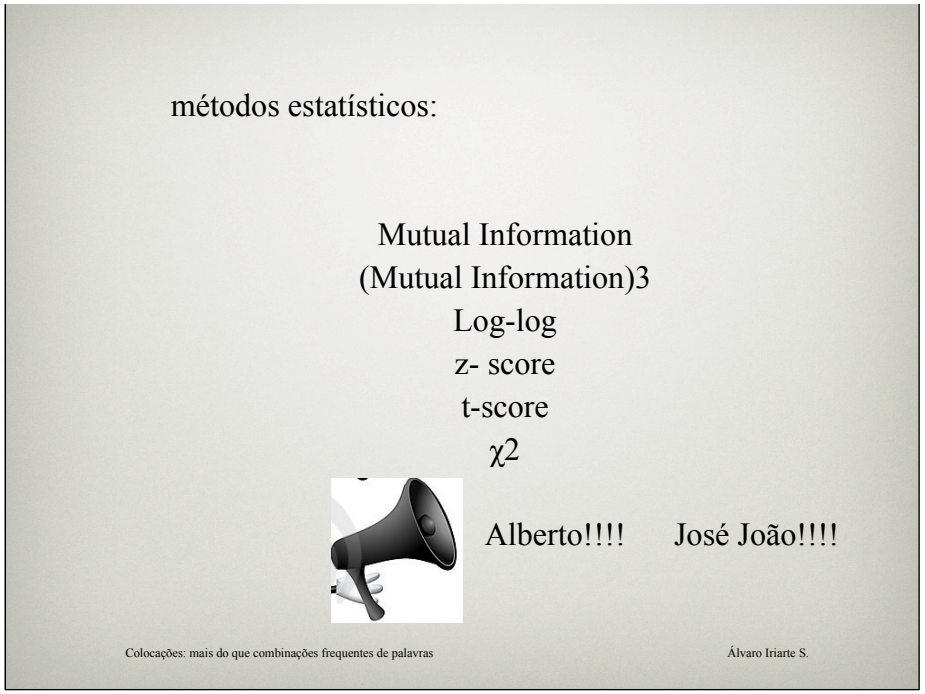


Conclusão: Métodos estatísticos + métodos fraseológicos + métodos contrastivos
(Moreno, 2009)

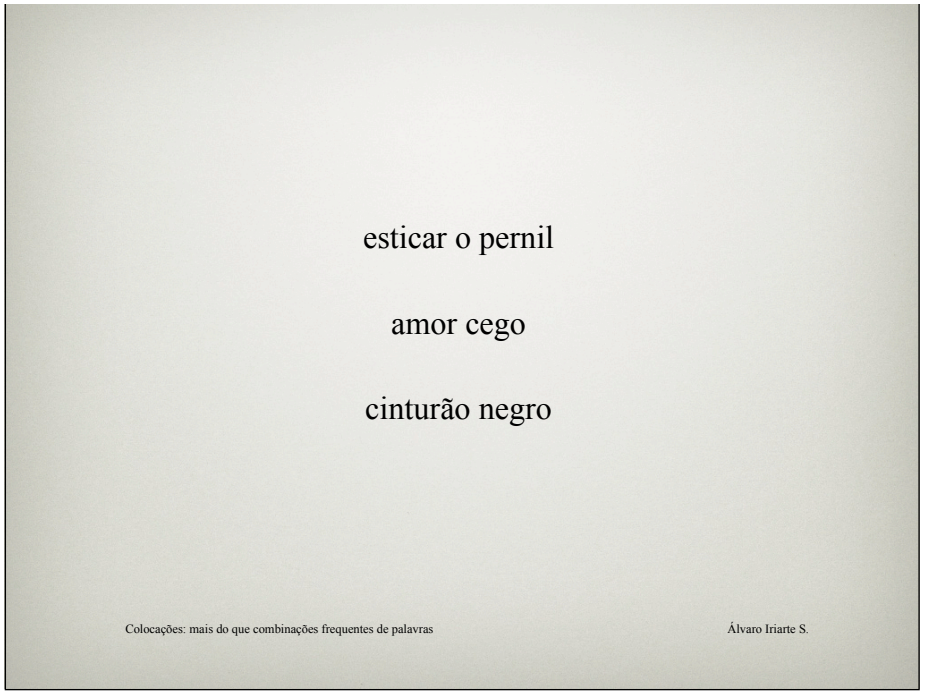


1. Em termos de lexicografia bilingue, as unidades de análise e descrição lexicográficas ou terminológicas poderão ser estabelecidas por contraste ou comparação das duas línguas, o que acarretará necessariamente o estabelecimento de uma unidade lexicográfica de carácter variável, que vai da palavra até à oração. A seleção da unidade de tratamento lexicográfico ou terminológico virá imposta pelo equivalente da L2.

Aqui é onde entram ferramentas como o Per-fide que, aproveitando o carácter irregular destas estruturas (as transformações admitidas numa língua e não na outra) poderão ser muito úteis para a identificação e para a extração automática deste tipo de unidades pluriverbais



Se não estiver enganado, e a classificação teórica destas “combinações lexicais” for irrelevante (ou quase) para a linguística aplicada, os métodos estatísticos, combinados com as estratégias contrastivas de que falei, poderão ser suficientes para a lexicografia, a terminologia, a tradução automática, etc.



Porque, como linguistas aplicados, o que é que queremos?
Continuar com esforços infecundos sobre que tipo de combinações são estas ou encontrar os equivalentes correspondentes e, muito importante, recolhe-los, num dic., numa base de dados, etc. de tal maneira que quem não conheça estas combinações possa encontrá-las facilmente?

Referências bibliográficas

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Barcelona: Antártida/Empúries.
Everaert, M., Linden, E. J. Van der, Scheak, A. and Schzender, R. (eds.) (1995). *Idioms: Structural and Psychological Perspectives.* Hillsdale-New Jersey Hove-U.K.: Lawrence Erlbaum Associates.
Guimier, C. (ed.) (1997): *Co-texte et calcul du sens.* Actes de la table ronde tenue à Caen les 2 et 3 février 1996. Caen: Presses Universitaires de Caen.
Mel'čuk, I. (1995). "Phrasemes in Language and Phraseology in Linguistics", em Everaert *et al* (ed.) (1995), 167-232.
Moreno Jaén, M. (2009). *Recopilación, desarrollo pedagógico y evaluación de un banco de colocaciones frecuentes de la lengua inglesa a través de la lingüística de corpus y computacional.* Granada: Editorial de la Universidad de Granada.
Morris, C. (1985). *Fundamentos de la teoría de los signos*. Barcelona: Paidós.
Rey-Debove, J. (1973). *Lexique et dictionnnaire*. Paris: Denoel.

Muito obrigado