Exemplos atestados e exemplos construídos na prática do léxico-gramática

Éric LAPORTE
Université Paris-Est
Institut Gaspard-Monge
5, bd Descartes
77454 Marne-la-Vallée CEDEX 2
eric.laporte@univ-mlv.fr

Tradução do francês: Francisco Antonio P. Léllis

Uma das grandes controvérsias epistemológicas em Lingüística é a oposição entre lingüística introspectiva e lingüística de corpus. Além do interesse teórico dessa questão, estão em jogo importantes conseqüências potenciais no tratamento automático das linguas. Croft (1993, 1998) contribui nesse debate quando propõe uma oposição entre método experimental e método observacional. Nesse artigo[1] queremos examinar essa diferença metodológica apoiando-no nos resultados da aplicação em grande escala de um método de descrição de línguas vivas por locutores nativos, o léxico-gramática.

1. Manipulação e observação em Lingüística

Opõem-se, frequentemente, duas abordagens metodológicas em Lingüística, todas as duas caricaturadas num célebre artigo de Fillmore (1992) que as denominou « lingüística de poltrona » e « lingüística de corpus ». Em um texto onde trata da tipologia das línguas, Croft (1993, 1998) propõe uma oposição entre método experimental e método observacional, que abarca quase perfeitamente a oposição precedente.

O « método experimental », que talvez fosse mais próprio qualificar de « manipulador », consiste em emitir uma hipótese lingüística, forjar exemplos fazendo variar sistemática e independentemente os parâmetros pertinentes, submeter essas formas a julgamentos introspectivos de aceitabilidade, e deduzir regras. Esse método pode teóricamente ser aplicado na ausência de qualquer corpus de exemplos preexistente ao estudo.

Já o « método observacional » consiste em observar as formas que constam num corpus preexistente e, em seguida, formular generalizações. Esse método é para Croft uma « alternativa legítima ao método experimental ».

Tomemos como exemplo um problema análogo ao utilizado por Boons *et al.* (1976) para discutir sobre esses dois métodos. Trata-se da preposição regida pelo nome *perícia* : *de* ou *em* ?

Apliquemos o « método experimental » e formulemos, por exemplo, a hipótese que só uma das duas preposições é empregada, ou que se trata de variantes livres. Os exemplos construidos serão os seguintes :

(1) *A polícia fez uma perícia nos pneus do carro*
(2) *A polícia fez uma perícia dos pneus do carro*
(3) *Entregaram-se perícias em vários computadores*
(4) *Entregaram-se perícias de vários computadores*

[1] Este artigo foi publicado em francês com o título original "Exemples attestés et exemples construits dans la pratique du lexique-grammaire" nos *Mémoires de la Société de linguistique de Paris*, nouvelle série, vol. 16, 2008, Leuven : Peeters.

etc., de maneira a variar os valores lexicais dos argumentos do verbo, dos determinantes e de qualquer outro parâmetro julgado pertinente pelo experimentador. Observando que todas essas formas são aceitáveis, validaremos a hipótese de uma regra segundo a qual trata-se de variantes livres.

Apliquemos agora o « método observacional » : nos munimos de um corpus e contamos, por exemplo, o número de ocorrências de cada uma das seqüências *perícia em* e *perícia de.* Com um corpus de 1.700.000 palavras extraído do Lácio-Web[2] (Aluísio et al. 2003), obtivemos uma só ocorrência de *perícia em.* Podemos deduzir uma só coisa : o sub-corpus é pequeno demais para oferecer uma visão fiel. Porém, tomando como corpus a coleção das páginas Web do Brasil escritas em português, obtemos como resultado 26.000 ocorrências de *perícia em* e 9.600 de *perícia de*. Podemos deduzir que as duas construções são usuais.

Independentemente das questões de fundo, é preciso reconhecer que alguns dos termos utilizados pelos diferentes autores são pouco apropriados. Não aludimos aqui à expressão "lingüística de poltrona", que é abertamente polêmica e estigmatiza os excessos de lingüistas pouco apressados em verificar se suas teorias estão em acordo com o uso efetivo, mas aludimos aos termos "método experimental" e "método observacional", ambos contestáveis. Efetivamente, cada um dos dois métodos que acabamos de ilustrar funda-se sobre observações empíricas. Além disso, nos dois casos, ao aplicá-los, respondemos a uma questão previamente formulada, usando uma estratégia que lança mão de uma confrontação com a realidade, o que faz da operação uma experiência. Toda experimentação implica, aliás, uma parte de observação.

Nesse artigo, pretendemos examinar vários aspectos dessa oposição, graças, de um lado, aos argumentos avançados por Croft, e de outro lado, graças à experiência ganha ao aplicar em grande escala um método de descrição sintático-semântica, o léxico-gramática. Limitaremos a discussão ao estudo de uma língua viva por locutores nativos. Nesse quadro, argumentaremos em favor de uma exploração combinada dos dois tipos de método.

2. Paralelo com as ciências experimentais

O paralelo com as ciências experimentais é muitas vezes utilizado nesse debate. Esse paralelo pode, efetivamente, ser bastante esclarecedor. Examinaremos em particular as práticas da pesquisa médica, da Biologia e da Física.

2.1. Experiência e observação em pesquisa médica

Na realidade, Croft (*ibid.*) utilisa a oposição existente entre experimental e observacional na Medicina. Mais uma vez, esses termos são pouco adaptados, mas são tradicionais. Em Medicina, o elemento de distinção entre estudo observacional e estudo experimental é uma diferença de grau no controle dos parâmetros. Nos dois casos, trata-se de avaliar as correlações entre sintomas e seus fatores eventuais tais como hábitos de vida, configurações genéticas ou tratamentos médicos.

Um estudo experimental é um estudo no qual o experimentador controla todos os parâmetros que podem revelar-se como fatores potenciais dos sintomas estudados. Para cada um desses parâmetros conhecemos seu valor (por exemplo, sabemos sobre cada sujeito se é vegetariano ou se come carne), ou sabemos seguramente que os resultados estatísticos do estudo serão independentes graças à constituição cuidadosa de grupos de sujeitos selecionados para este efeito. Por exemplo, para tornar o estudo independente da inclinação pessoal dos sujeitos em relação ao consumo de carne, constitui-se um grupo de vegetarianos e um grupo

[2] http://www.nilc.icmc.usp.br/lacioweb/

de comedores de carne, pedindo-lhes que sigam o regime alimentar que corresponde a cada grupo e não que sigam livremente seu apetite. Assim, a escolha espontânea, mesmo se não está registrada formalmente como parâmetro, não terá nenhuma influência sobre os resultados estatísticos, com a condição que os grupos de indivíduos sejam suficientemente grandes.

Em um estudo de observação, não se exige que o experimentador controle todos os parâmetros. Por exemplo, podemos comparar os sintomas de um grupo de vegetarianos e de um grupo de comedores de carne, sem se preocupar sobre a maneira como determinaram seus regimes alimentares, o que é bem mais fácil de ser realizado. Entretanto, os resultados de tal estudo poderão ser falseados pela confusão entre dois parâmetros *a priori* distintos: o regime alimentar efetivo e o regime que o sujeito teria adotado espontaneamente. Por exemplo, este estudo não permitirá de predizer os sintomas de um sujeito preferencialmente vegetariano, mas para o qual teria sido prescrito de comer carne.

O protocolo experimental é considerado como o mais adequado para permitir uma demonstração estatística de causa a efeito, enquanto que o estudo observacional fornece apenas indicações. Assim, contrariamente à sugestão de Croft no campo da tipologia das línguas, o método observacional em pesquisa médica não é, em geral, uma "alternativa legítima ao método experimental".

2.2. Experiência e observação em Física e em Biologia

As correspondências evocadas por Croft (*ibid.*) entre a Física e o "método experimental", e entre a Biologia e o "método observacional" parecem-nos ainda menos convincentes do que o paralelo feito com a pesquisa médica. A Física utiliza uma parte de observação: por exemplo, a observação sistemática do espectro dos corpos. Mesmo se essas observações são efetuadas mediante experiências, um de seus objetivos é a observação pura e simples do mundo que nos envolve. Já a Biologia pode recorrer a atividades de tipo observacional: um naturalista observa a presença de espécies de animais ou vegetais encontrados num meio natural; mas isso não exclui a realização de experiências, por exemplo, sobre a influência de fatores internos ou externos sobre o comportamento de um animal.

Nota-se, de fato, nas ciências experimentais en geral, uma combinação de práticas comparáveis ao que Croft chama de método experimental e de método observacional. Esses dois tipos de prática correspondem a objetivos distintos, são consideradas como complementares pelos científicos, e fornecem resultados uma à outra, mutuamente. Por exemplo, é natural que as observações, mesmo realizadas fora de um protocolo experimental estrito e custoso, suscitem hipóteses que são em seguida testadas graças à experiências rigorosas; inversamente, resultados obtidos por algumas experiências, em particular medidas de grandezas, podem ser considerados como observações empíricas. Essas situações são amplamente ilustradas pela história das ciências experimentais.

Pensamos que, na realidade, a mesma situação é válida em Lingüística, e que a disputa entre o linguísta introspectivo e o linguísta de corpus, tais como foram caricaturados por Fillmore (1992), decorre apenas das suas resistências na utilização de abordagens metodológicas que são, porém, compatíveis. É o que tentaremos de demonstrar em seguida.

3. Manipulação e observação na prática do Léxico-gramática

O Léxico-gramática (Gross, 1975, 1981, 1994) designa ao mesmo tempo uma metodologia e uma prática efetiva de descrição manual sintático-semântica. Essa metodologia e essa prática desenvolveram-se simultaneamente a partir do final dos anos 1960, enriquecendo-se mutuamente. Vamos delimitar o quanto devem aos métodos "experimental" e "observacional".

3.1. O Léxico-gramática: princípios e resultados

A base teórica sobre a qual se funda o Léxico-gramática é o distribucionalismo de Harris (1964, 1976). Os princípios metodológicos que se desenvolveram em seguida (Gross, 1975, 1981, 1994) podem ser considerados como a adoção de prioridades num programa de descrição sintático-semântica das línguas.

A interação entre o léxico e a sintaxe é assim considerada como uma chave imprescindível. A pesquisa exclusiva de regras de sintaxe geral, independentes do material lexical que utilizam, é denunciada como um impasse. Inversamente, a descrição do vocabulário de uma língua é vista como o estudo das maneiras como cada elemento lexical insera-se nas frases. Em outros termos, a unidade mínima tomada como contexto para a descrição de uma palavra é a frase elementar.

O léxico-gramática coloca igualmente uma exigência de formalização. Os resultados da descrição devem ser suficientemente formais para permitir:
- uma verificação pela confrontação com a realidade do uso,
- uma aplicação ao tratamento automático das línguas.

Essa obrigação de formalização manifesta-se pela adoção de um modelo discretizado da sintaxe. Assim, a aceitabilidade é modelizada por uma noção binária: para as necessidades da descrição, uma frase é considerada como aceitável ou não. Da mesma maneira, a ambigüidade lexical é representada pela separação de uma palavra em um número inteiro de entradas lexicais, que são distintas umas das outras da mesma forma que duas entradas de palavras morfologicamente diferentes. As propriedades sintático-semânticas são identificadas por fórmulas que representam estruturas de frases, fórmulas simples como $N_0\ V\ N_1\ W = N_1\ V\ W$ (ver secção 3.2), que formam uma lista, sistematicamente confrontadas com todas as entradas. Enfim, somente são estudadas as propriedades para as quais se encontrar um procedimento que permita determinar de maneira suficientemente confiável se uma entrada a possui ou não: as propriedades são pois modelizadas como binárias e não como um *continuum.*

Os resultados obtidos pela aplicação desses princípios metodológicos por algumas dezenas de lingüistas durante algumas dezenas de anos fazem do léxico-gramática uma empresa sem precedentes. Limitando-se ao francês, foram estabelecidas em torno de 13.000 entradas verbais, 10.000 entradas nominais, 12.000 entradas de frases fixas, 11.000 entradas adverbiais. Desse total de 75.000 entradas, 98% foram cruzadas e confrontadas com centenas de propriedades sintático-semânticas. Mais da metade dessas entradas estão à disposição gratuitamente no endereço http://infolingu.univ-mlv.fr e constituem uma base de informações sintático-semânticas para o tratamento das línguas sem equivalente no mundo pelo seu volume, pela riqueza dos fenômenos lingüísticos que abarca e pelo seu grau de formalização. Dado os princípios acima colocados, os resultados da descrição têm naturalmente a forma de tabelas de dupla entrada, que cruzam as entradas lexicais com as propriedades sintático-semânticas (figura 1). Alguns milhares de outras entradas foram publicadas somente sobre suporte impresso, mas seguem o mesmo modelo. Enfim, descrições mais ou menos substanciais, seguindo sempre o mesmo modelo, existem sobre uma dezena de outras línguas, sendo as mais representadas o italiano, o portugês, o grego moderno e o coreano.

N0=:Nhum	N0=:N-hum	Neg.obrigat.	V	DET- obrigat. DET opcional	C1	C1=:Nplural	C1=:Npc	Modificador	com N	Apassivação	ação	ação-processo	processo	estativo	Lit. Romanesca	Lit.Dramática	Lit. Técnica	Lit. Oratória	Lit. Jornalística
+	-	-	<saber>	o	caminho das pedras	-	-	-	-	+	+	-	+	-	-	-	-	-	+
+	-	-	<saber>	a	lição	-	-	-	-	-	-	-	+	+	+	-	-	-	-
+	-	-	<sacar>		orelhas	+	+	-	-	-	-	-	+	-	-	-	-	-	-
+	-	-	<sacudir>	as	cadeiras	+	+	-	-	-	+	-	-	-	-	-	-	-	+
+	-	-	<sacudir>	o	esqueleto	-	+	-	-	-	+	-	-	-	-	-	-	-	+
+	-	-	<sacudir>		poeira	-	-	-	-	-	+	-	+	-	-	-	-	-	+
+	-	-	<safar>	a	onça	-	-	-	-	-	+	-	-	-	+	-	-	-	-
+	+	-	<saldar>	as	contas	+	-	-	-	-	+	-	-	-	-	-	-	-	-
+	-	-	<salgar>	o	galo	-	+	-	-	-	+	-	-	-	-	-	-	-	-

Figura 1. Extrato de uma tabela de expressões verbais fixas (Vale, 2001)

3.2. Precauções metodológicas para a reprodutibilidade

Para explorar a interação entre o léxico e a sintaxe, é preciso, naturalmente, combinar sistematicamente as entradas lexicais com todas as estruturas de frases observadas, e também analisar as seqüências assim geradas: são aceitáveis? Quais são suas particularidades distribucionais e semânticas?

A qualidade dos resultados depende, pois, das capacidades dos lingüistas em analisar os exemplos construídos. A experiência mostra que a maneira mais eficaz de efetuar essa exploração e essa análise é recorrer massivamente à introspecção. Entretanto, se é então confrontado a três riscos de erros.

O primeiro risco é o de que o lingüista tenha uma capacidade insuficiente para analizar as seqüências e, principalmente, de julgar da sua aceitabilidade. No léxico-gramática excluímos que um lingüista aplique o método a um outro idioma senão sua própria língua materna, mesmo com a ajuda de um informante. Mas, mesmo assim, a capacide de julgamento da aceitabilidade é um talento do qual não somos todos igualmente dotados, como acontece aliás, de uma certa forma, em quase todos os campos da atividade humana.

O segundo risco é a diferença que existe entre a língua descrita e o idioleto do descritor. Tivemos um exemplo concreto desse risco quando o avaliador de um artigo de nossa autoria reparou, durante o processo editorial, que a expressão adverbial francesa *au petit bonheur la chance* era citada sob a forma de *au petit bonheur de la chance*, a única forma que conhecia nosso idioleto pessoal.

O terceiro risco é a existência de um preconceito inconsciente do lingüista, influenciado pelo desejo que seja averiguada uma de suas hipóteses. Temos, por exemplo, uma tendência natural a regularizar um fenômeno. Durante o estudo da relação de nominalização estabelecida entre as duas frases seguintes:

Zé aterrissa = *Zé efetua uma aterrissagem*

podemos assim ser tentados de superestimar a aceitabilidade da seqüência (5):

Zé embreia = (5) ? *Zé efetua uma embreagem*

Todos esses problemas são bem conhecidos dos lingüistas que praticam uma atividade descritiva regular. Esses problemas apareceram desde o início da construção do léxico-gramática, no fim da década de 1960. Têm equivalentes em todas as ciências experimentais:

trata-se de dificuldades práticas susceptíveis de serem um obstáculo à reprodutibilidade de uma experiência ou da medida de uma grandeza. Uma experiência, uma medida, só têm um interesse científico se são reprodutíveis, quer dizer, se um experimentador que as praticar de novo obtem os mesmos resultados.

Essa exigência de reprodutibilidade é tão fundamental em matéria de descrição lingüística quanto nas (outras) ciências experimentais. Os três riscos expostos acima são três causas sistemáticas de não-reprodutibilidade das observações ou das experiências necessárias à construção de um léxico-gramática. Tornam irrealista a exigência de uma reprodutibilidade absoluta, mas se os analizamos, chegamos à conclusão que são superáveis na medida em que existe uma comunidade lingüística que fala a língua estudada. Com efeito, se tal comunidade existe, não vemos porque não produziria locutores com talentos variados, inclusive o de julgar aceitabilidades. O problema dos idioletos pode, da mesma maneira, ser resolvido pela confrontação dos julgamentos emitidos por diferentes locutores. Enfim, os preconceitos que podem falsear nossos julgamentos podem igualmente ser detectados e combatidos graças a um controle entre pares, se feito de forma correta.

Um dos méritos do léxico-gramática é o de ter-se dotado desde o início de um arsenal de precauções metodológicas contra os riscos próprios a atividade de construção de exemplos (Gross, 1984), e de tê-los feito evoluir na medida das necessidades e do aparecimento de novos meios técnicos.

a) Uma das precauções consiste em organizar sessões coletivas regulares durante as quais os lingüistas controlam mutuamente seus julgamentos e suas análises. Assim, o léxico-gramática dos verbos distribucionais[3] do francês (Gross, 1975, Boons *et al.*, 1976, Guillet e Leclère, 1992) foi construído durante reuniões nas quais participavam pelo menos cinco lingüistas: Jean-Paul Boons, Jean Dubois, Maurice Gross, Alain Guillet e Christian Leclère, de 1969 a 1984. Atualmente, o projeto Bélgica-França-Quebec-Suíça (BFQS) sobre as diferenças entre expressões verbais fixas em quatro variantes do francês (Labelle, 1990, Lamiroy et al., 2003) realiza-se durante reuniões com 4 a 6 lingüistas. O principal incoveniente dessa prática é o seu custo.

b) A segunda precaução metodológica consiste em se interrogar sistematicamente sobre os critérios de verificação das propriedades sintático-semânticas estudadas, avaliando a reprodutibilidade da aplicação desses critérios. Por exemplo, um dos critérios adotados para determinar se um verbo distribucional transitivo direto admite a transformação chamada média, anotada $N_0\ V\ N_1\ W = N_1\ V\ W$:

(6) *Zé dobrou o seu salário* = (7) *O seu salário dobrou*

consiste em aplicar formalmente a transformação e em julgar da aceitabilidade do resultado:

Zé olhou o cardápio * *O cardápio olhou*

Trata-se de um critério formal. A experiência mostra que os critérios formais são, em geral, de uma reprodutibilidade claramente superior à maior parte dos outros tipos de critério aplicáveis. Afim de beneficiar desse efeito, sempre que possível, a definição das propriedades sintático-semânticas representadas no léxico-gramática se apoia em critérios formais, detalhados nos livros, artigos e teses publicados com as tabelas.

Entretanto, os critérios formais não são suficientes para caracterizar as propriedades e devem às vezes ser acompanhados de critérios semânticos. Privilegiamos nesse caso aqueles que se utilizam de uma avaliação semântica diferencial, claramente mais confiável

[3] Os verbos distribucionais são aqueles que podem ser analisados como predicados. São reconhecidos pelo fato de que a distribuição dos argumentos depende do verbo (exemplo: *Zé bateu uma frase no teclado*). Esse conceito se opõe aos de verbo suporte (*Zé bateu uma foto da Ana*) e de expressão verbal fixa (*Zé bateu um fio para a Ana*), nas quais a distribuição dos argumentos depende do nome predicativo ou de toda a parte fixa da expressão.

que a avaliação semântica absoluta (Gross, 1975). A avaliação semântica diferencial consiste em comparar duas diferenças semânticas. Por exemplo, a diferença entre (6) e (7):

(6) *Zé dobrou o seu salário* = (7) *O seu salário dobrou*

pode ser comparada com a entre (8) e (9):

(8) *Zé pesa a sacola* = (9) *A sacola pesa*

O leitor perceberá provavelmente, como nós, que a diferença entre (6) e (7), que parece ter uma relação com a causalidade do processo, não se reencontra de maneira alguma entre (8) e (9). Esta observação é muito mais reprodutível do que a que consistiria, por exemplo, em caracterizar em que (6) e (7) diferem semanticamente, ou, quanto mais, em caracterizar o sentido de (6). Além disso, a noção de avaliação semântica diferencial está no centro da noção harrissiana de transformação: a relação $N_0\ V\ N_1\ W = N_1\ V\ W$ será considerada como uma transformação somente se a diferença semântica entre (6) e (7) se reencontrar num número suficiente de pares reproduzindo as duas mesmas estruturas com outro material lexical.

O léxico-gramática utiliza critérios semânticos absolutos em casos exceptionais, onde fornecem resultados julgados suficientemente reprodutíveis. Assim, certas propriedades sintático-semânticas dos verbos distribucionais formalizadas no léxico-gramática pressupõem que uma frase exprime um deslocamento de uma entidade, denotada por um dos argumentos, em relação a um lugar, denotado por outro. Por exemplo, a propriedade $N_0\ V\ N_1\ Loc\ N_2$ pode ser verificada só se houver um deslocamento da entidade denotada por N_1 em relação ao lugar denotado por N_2 :

(10) *Zé enfia o envelope na gaveta*

Para tornar suficientemente reprodutível a aplicação desse critério, em particular no caso de um deslocamento mais ou menos abstrato, ou de frases nas quais a interpretação envolve outro processo paralelo ao deslocamento, foi necessário enquadrá-lo elaborando o seguinte processo (Guillet e Leclére, 1992): forjam-se duas frases locativas, sendo uma a negação da outra, por exemplo:

(11) *O envelope não está na gaveta*

(12) *O envelope está na gaveta*

e verifica-se que a interpretação de (10) supõe que a de (11) é verificada antes do processo, e a de (12) depois.

Este é o preço pago para podermos considerar as propriedades sintático-semânticas como definidas com suficiente precisão para que faça sentido confrontá-las ao léxico inteiro.

As tabelas de léxico-gramática realizadas nas décadas de 1970 e 1980 o foram graças às precauções metodológicas que acabamos de expor aqui. Como o leitor reparou, não se utilizam em nenhum momento de um corpus, limitando-se assim a uma lingüística exclusivamente introspectiva e manipuladora.

Com efeito, nessa época, as coleções de textos disponíveis em formato eletrônico eram pequenas demais para poder melhorar o processo. Os concordanceadores disponíveis não eram suficientemente elaborados para permitir a produção de concordâncias lematizadas a partir de um texto não anotado (aliás, ainda é o caso dos concordanceadores utilizados pela maioria dos lingüistas). Enfim, não existia praticamente nenhum corpus de textos anotados nem lematizados.

c) No começo da década de 1990 essa situação mudou, o que permitiu aos construtores do léxico-gramática de recorrer cada vez mais facilmente a uma terceira precaução metodológica: a utilização de exemplos atestados nos corpus. Com efeito, por um lado, com a criação do sistema Intex (Silberztein, 1993), foi possível pesquisar nas grandes coleções de textos estruturas lingüísticas especificadas pelo seu conteúdo lexical e morfossintático (lemas, categorias gramaticais, traços flexionais) e produzir as

concordâncias correspondentes[4], muito mais úteis que aquelas produzidas por concordanceadores sem léxico. Por outro lado, com a criação do Web, e em seguida do motor de pesquisa Google (1999) e do sistema Webcorp (Renouf, 2003), imensas coleções de textos tornaram-se acessíveis. Assim, o projeto BFQS recorre freqüentemente à formas atestadas no Web (figura 2). Esse controle suplementar pela observação de corpus substitui-se parcialmente ao controle mútuo evocado mais acima em *a)*, com a vantagem de que coloca em ação mais locutores. Entretanto, não pode substituir pura e simplesmente o conjunto de precauções metodológicas que expusemos. As razões dessa impossibilidade são bem conhecidas e foram apresentadas muitas vezes no debate sobre os méritos respectivos da lingüística introspectiva e da lingüística de corpus. Contentemo-nos de lembrá-los brevemente:

- A observação de corpus não fornece análises das diferenças de sentidos ou das diferenças entre as variantes de uma língua.
- Não fornece, por si só, a formalização dos fatos observados.
- Não atesta as inaceitabilidades: por exemplo, a ausência da expressão *perícia de* num corpus de 1.700.000 palavras não prova em nada que essa expressão seja inusitada.

Expressão	B	F	Q	S	Paráfrase	Exemplo
Amuser à des riens (s')	+	+	+	+	Se distraire avec des futilités	Il est comme un petit enfant, il s'amuse à des riens.
Amuser à un rien (s')	+	!	-	+	Se distraire avec des futilités	
Amuser bien (s')	+	-	-	-	Se plaire quelque part	Est-ce que tu t'amuses bien dans ton nouvel appartement ?
Amuser la galerie	+	+	+	+	Distraire l'assistance	"Lorsqu'il était petit, il amusait la galerie avec ses mimiques, ses blagues : un acteur était né. " (www)
Amuser le tapis	-	+	-	+	Distraire l'assistance	"Raffarin veut-il amuser le tapis ? Après tout, pourquoi pas, mais la situation dramatique de la France mérite mieux." (www)
Amuser le temps	-	-	+	-	Faire passer le temps	Pierre n'a rien fait de la journée. De plus en plus, j'ai l'impression qu'il amuse le temps.

Figura 2. Um extrato do dicionário BFQS

Por todas essas razões, e embora o trabalho descritivo do léxico-gramática tenha recorrido de maneira crescente a um controle pela observação de corpus durante as décadas de 1990 e 2000, isso não levou de maneira alguma a abandonar as precauções elaboradas no

[4] O sistema Unitex (Paumier, 2006), livremente disponível (http://univ-mlv.fr/~unitex), propõe a mesma funcionalidade. O sistema Glossanet (Fairon e Singler, 2006) faz a mesma coisa explorando páginas web renovadas cotidianamente, também de graça.

período anterior; essa nova precaução apenas juntou-se às precedentes, fazendo do léxico-gramática um método que pertence ao mesmo tempo à lingüística introspectiva e à lingüística de corpus, um pouco como preconizava Fillmore (1992). Os projetos americanos FrameNet (Baker *et al.*, 2003) e VerbNet (Kipper-Schuler *et al.*, 2006) testemunham de uma relativa convergência em direção dos objetivos do léxico-gramática.

Examinemos, por exemplo, os procedimentos utilizados no projeto BFQS para detectar as expressões verbais fixas cujo emprego não é uniforme nas quatro variantes do francês. Os representantes de cada variante estabelecem primeiro quatro listas separadas. Em seguida, comparam-nas. Ora, para descobrir, por exemplo, que uma expressão da lista B (para Bélgica) é inusitada na variante F (para França), é preciso um contato entre um representante da variante B e um representante da variante F. Com efeito, se uma expressão da lista B está ausente da lista F, o autor dessa última pode simplesmente não tê-la reparado. Ademais, se uma expressão está presente ao mesmo tempo nas lista B e F, isso não significa necessariamente que seja comum às duas variantes: é preciso, nesse caso, confrontar as interpretações; se são diferentes, trata-se, de um ponto de vista lexicológico, de duas expressões distintas, cada uma sendo usual em uma variante e inusitada na outra.

Como vemos, este procedimento precisa de análises de natureza introspectiva. Está longe de se limitar à comparação informática das duas bases de dados, e ainda menos de dois corpus.

d) As tabelas de léxico-gramática estando, no essencial, publicadas, é possivel a todos e a cada um julgar se as diferentes precauções tomadas desempenharam seu papel, verificando se os resultados obtidos estão de acordo com os julgamentos de aceitabilidade que podem emitir os locutores do francês. Tal exame mostra que uma certa proporção das marcas indicando se as entradas possuem ou não as propriedades são errôneas. Entretanto, as duas principais causas aparentes desses erros são:

- por um lado, a existência de colunas correspondendo a propriedades mal definidas, introduzidas a título experimental pelos lingüistas à espera da opinião de seus pares, mas que seria melhor não considerar como soluções satisfatórias dos problemas descritivos aos quais correspondem; por exemplo, as propriedades dos verbos distribucionais que se referem aos nomes das partes do corpo (*Npc*);
- por outro lado, a presença de erros de informática durante a transferência de dados de um sistema para outro[5].

Assim, as informações lingüísticas formalizadas nas tabelas do léxico-gramática em seu estado atual possuem um interesse científico e técnico de primeira ordem, mesmo não sendo totalmente isentas de erros. Mas o método de construção dessas tabelas apresentam um interesse mais importante ainda do ponto de vista lingüístico; além disso, permite corrigir os erros e construir as tabelas que faltam para outras partes do léxico ou então para outras línguas.

3.3. Verificação da pertinência das hipóteses

Na seção precedente evocamos essencialmente técnicas de análise de exemplos construídos. Ocupemo-nos agora da arte de construir exemplos. Para poder retirar conclusões válidas da análise de exemplos, é preciso construí-los de maneira rigorosa e organizada. Consideremos, por exemplo, o problema da relação entre três construções do verbo *abundar*:

(13) *As obras para cello abundam no período romântico*

(14) *O período romântico abunda de obras para cello*

[5] Durante os vinte primeiros anos de existência do léxico-gramática, os instrumentos informáticos de manipulação de bases de dados contendo textos eram deficientes, e as normas de representação do texto eram caóticas.

(15) *O período romântico abunda em obras para cello*

A identidade do material lexical e a semelhança semântica fazem pensar a uma transformação que intervertiria o sujeito e o complemento. Coloquemos, pois, à prova a hipótese de uma transformação N_0 *abundar Loc* $N_1 = N_1$ *abundar (de + em)* N_0 (cf. Boons *et al.*, 1976; Salkoff, 1983). Nessa notação informal não aparecem os determinantes e os modificadores do substantivo, mas o determinante de N_0 não concorda nos exemplos (13) e (14). Em outras transformações intervertindo os argumentos sintáticos de um verbo distribucional (ditas de cruzamento), os determinantes podem ser conservados:

Zé salpica um pouco de açúcar no bolo
Zé salpica o bolo com um pouco de açúcar
Zé vai salpicar esse açúcar todo no bolo
Zé vai salpicar o bolo com esse açúcar todo

Testemos pois a hipótese (generalizante, portanto simplificadora) de uma conservação do determinante na transformação hipotética que nos ocupa agora. Uma experimentação adaptada à verificação dessa hipótese consiste em fazer variar o determinante de N_0 nos exemplos (13) e (14) antes de analisar as seqüências assim forjadas:

(13) *As obras para cello abundam no período romântico*
= (16) * *O período romântico abunda (das + nas) obras para cello*
(17) * *Obras para cello abundam no período romântico*
= (14)-(15) *O período romântico abunda (de + em) obras para cello*

A construção dos exemplos não é trivial: envolve a utilização de contrações (*em as = nas*) e necessita lembrar que o determinante zero pode ser equivalente a um determinante indefinido. As diferentes interdições marcadas nos exemplos (16) e (17), acima citados, invalidam a hipótese que os suscitou e sugerem que a realidade do uso é mais complexa. Todavia, levando em conta outros determinantes, ao contrário, sugere-se que a hipótese seria válida para certos tipos de determinante, por exemplo *este tipo de* e *algum tipo de*:

(18) *(Este + Algum) tipo de obra abunda no período romântico*
= (19) *O período romântico abunda (de + em) (este + algum) tipo de obra*

Isso sugere ao experimentador que passe em revista as diferentes categorias de determinantes e que realize experiências independentes em função dessa tipologia.

Mesmo sem levar mais longe esse estudo de caso, podemos observar que os exemplos que permitiram encontrar respostas parciais a nossas sucessivas questões têm a particularidade de fazer variar cada parâmetro independentemente. Entre (14) e (15) a preposição é que varia. Entre (13) e (14) variamos dois parâmetros ao mesmo tempo: a posição dos argumentos e o determinante de N_0 . Essa falta de rigor é corrigida pela construção dos exemplos (16) a (17) e depois (18) e (19), destinados a separar os dois parâmetros em questão.

Em outros termos, como em toda ciência experimental, imaginam-se experiências destinadas a colocar em evidência, separadamente, os efeitos ligados aos diferentes parâmetros que podem revelar-se como sendo os fatores dos fenômenos observados. Pode-se, assim, validar as diferentes hipóteses subjacentes a essas experiências, ou então imaginar outras hipóteses.

4. As críticas de Croft

Em seu artigo, Croft (*ibid.*) preconiza implicitamente o emprego do "método observacional" como uma "alternativa legítima ao método experimental", formulando duas críticas sobre esse último. Examinemos essas críticas graças à experiência ganha ao aplicar o léxico-gramática. É preciso, entretanto, lembrar que Croft se refere à tipologia das línguas, um domínio que não pode ser identificado com o da descrição sintático-semântica.

4.1. Objetividade e subjetividade

A primeira crítica de Croft contra a lingüística introspectiva é clássica: como a experiência introspectiva baseia-se no julgamento de aceitabilidade pelo próprio experimentador, esse último é o objeto de sua própria experiência, existindo pois um risco de viés, ou deformação da observação, que já evocamos na seção 3.2. Como diz Croft, "um psicólogo rejeitaria imediatamente tais condições". A lingüística de corpus não corre esse risco, pois o observador é independente dos autores do corpus.

Assinalemos, para começar, que em outras ciências experimentais o risco ligado ao fato de que o experimentador é, em parte, objeto de sua própria experiência, é considerado em certas condições como um risco controlado que não prejudica em nada a validade dos resultados obtidos. Assim, quando um naturalista observa o cheiro de um cogumelo (um dos elementos essenciais para a determinação das espécies, na prática), não se exige que o faça sem saber de onde vem o cheiro, ou sem saber onde estava quando encontrou o cogumelo. Tais condições seriam rejeitadas imediatamente por um biólogo, pois são tão inúteis quanto impraticáveis. Os biólogos possuem, aliás, o bom senso de não procurar a ajuda de um psicólogo para determinar as espécies de cogumelos. Em ciências experimentais, a noção de reprodutibilidade é considerada como mais pertinente que a noção de objetividade.

Retornando à lingüística, enviamos o leitor à seção 3.2 desse artigo para lembrar que os autores do léxico-gramática, cientes do risco de viés, lançaram mão de um conjunto de precauções que os obrigaram a um grande rigor e que se baseiam em noções de ordem não psicológicas, mas sim lingüísticas.

Croft não alude a esses procedimentos, apesar de que foram aplicados a dados de uma larga cobertura lexical e gramatical, e que a maioria dos resultados obtidos pelo léxico-gramática terem sido publicados. Talvez este silêncio explique-se pelo fato que Croft julgue a lingüística introspectiva só através de sua tendência mais conhecida, a gramática gerativa.

É verdade que a crítica sobre a falta de objetividade é suficientemente justa em relação a esse movimento, grande produtor de estruturas abstratas (mas é muito difícil, na prática, verificar a conformidade dessas estruturas com os fatos observados); em outros termos, produtor de hipóteses infalsificáveis, no sentido de Popper (1959). A gramática gerativa não tem a fama de aplicar métodos particularmente elaborados em matéria de observação. A noção de observação é, aliás, considerada como relativamente trivial nas tradições "culturais" da gramática gerativa. Assim, a "adequação observacional" ocupa o nível o mais baixo na hierarquia dos três níveis de adequação de uma representação gramatical: adequação observacional, descritiva e explicativa. Isso é particularmente significativo, dado o papel decisivo assumido pela noção de prestígio nesse movimento. Talvez essa posição coletiva possa ser entendida como uma reação à posição metodológica de Harris, que, pelo contrário, visa resolutamente a "superfície" (diretamente observável) das línguas.

Entretanto, não cabe avaliar a lingüística introspectiva através de representantes que são descuidados com a observação dos fatos, sejam eles os mais conceituados.

Já a lingüística de corpus tem, às vezes, uma exigência de objetividade tão excessiva quanto é menosprezada pela gramática gerativa. O desenvolvimento espetacular da lingüística de corpus apresenta, aliás, também, certos aspectos de uma revolução em relação a uma posição metodológica anterior, vista como a "lingüística de poltrona", especialmente no Reino Unido, onde a lingüística de corpus tem quase sufocado as outras abordagens da lingüística.

Mas não nos percamos em direção da sociologia dos movimentos científicos. Retornemos a argumentos científicos.

4.2. Elaboração de hipóteses

A segunda crítica de Croft aplica-se ainda menos ao domínio que nos interessa. Na aplicação do "método experimental", reclama ele, "nenhuma generalização pré-formulada é testada, como sempre deve ser feita pelo experimentador que conduz uma experiência". Eis

aqui uma razão bem pouco convincente para justificar a renúncia aos métodos da lingüística introspectiva.

Primeiramente, a formulação de hipóteses prévias às experiências está no centro da prática efetiva da lingüística introspectiva (cf. seção 3.3). Os lingüistas empregam-no, aliás, para se proteger contra a complexidade dos fatos que, segundo Croft, impedem de aplicá-lo. É verdade que a tipologia das línguas, sobre a qual falava Croft, acrescenta um nível de complexidade suplementar.

Em segundo lugar, a observação dos fatos, mesmo independentemente da formulação de uma hipótese, pode ser uma atividade científica legítima. As ciências experimentais fornecem numerosos exemplos, desde a pesquisa médica (os estudos observacionais, justamente) até a Física (a observação e o recenseamento sistemáticos dos corpos celestes ou das propriedades dos elementos) ou a Biologia (a observação das espécies que habitam os biótopos).

Em terceiro lugar, o "método observacional" preconizado por Croft não comporta, em geral, a formulação de hipóteses.

Rejeitar a lingüística introspectiva pelo motivo que não formula hipóteses, portanto, seria equivalente a se privar de um instrumento invocando o fato que seria necessário servir-se dele.

5. A regra e o exemplo

Como vimos, o artigo de Croft (*ibid.*) empresta os termos e as noções do "método experimental" e do "método observacional" à pesquisa médica, onde se considera consensualmente, ao contrário do que Croft preconiza implicitamente, que o primeiro método fornece mais informações que o segundo. Aprofundemos o paralelo com a pesquisa médica um pouco mais longe do que simplesmente o empréstimo de termos. O defeito de um estudo "observacional" é o de fornecer apenas uma coleção de casos, enquanto que um estudo "experimental" é concebido para que a coleção de casos tenha as propriedades estatísticas necessárias para que possamos deduzir a existência de relações de causa a efeito, ou seja, de regras. Temos, portanto, dois tipos de estudo: uns menos custosos, os outros capazes de colocar regras em evidência. Ora, essa complementariedade se encontra também entre a lingüística de corpus e a lingüística introspectiva.

Para demonstrá-lo, lembremo-nos em primeiro lugar de algumas noções de bom senso sobre os conceitos de regra e de exemplo. Em Matemática, uma regra tem mais valor que os exemplos, se é mais geral. Assim, comparemos a regra: "nenhum número inteiro par superior a 2 é um número primeiro", com dois exemplos: "6 e 14 não são números primeiros". É bem mais natural se considerar a regra como mais interessante, pois dela podemos deduzir os dois exemplos. Essa preferência tem seus limites. De fato, se a regra é falsa, não tem nenhum valor. Da mesma maneira, se os exemplos enumeram a totalidade das possibilidades cobertas pela regra, o seu interesse não é mais tão óbvio.

Examinaremos, nos parágrafos abaixo, como essas noções contribuem para o debate sobre a lingüística introspectiva e a lingüística de corpus, e em seguida, sobre as principais abordagens do tratamento automático das línguas.

5.1. Em Lingüística

Em princípio, a lingüística introspectiva e manipuladora tem a capacidade de descobrir e formalizar regras, cuja acumulação pode formar uma gramática. Enviamos o leitor à seção 3 para encontrar os exemplos. Pelo contrário, a lingüística de corpus pura limita-se a encontrar exemplos tirados de um corpus e, portanto, produz resultados de um escopo menos geral, a não ser que se arrisque uma generalização mais ou menos temerária. Em outros termos, não resolve o problema da formalização. Essa é a razão principal que nos conduz a defender a

persistência de abordagens lingüísticas que não dependam apenas e exclusivamente da lingüística de corpus. Entretanto, essa reflexão tem seus limites.

Em primeiro lugar, as regras produzidas pela lingüística introspectiva podem ser falsas, principalmente se a confrontação com a realidade lingüística for insuficiente ou pouco rigorosa: é a "lingüística de poltrona".

Em segundo lugar, mesmo se nenhum corpus pode cobrir todas as possibilidades de uma língua, o web (apesar de seus defeitos, pois encontramos nele muitos erros) visto como uma coleção de textos de uma dada língua, se aproxima de uma certa maneira desse ideal pelo seu volume e sua diversidade (pelo menos para línguas como o francês e o inglês).

Justamente, essas duas reservas é que motivaram a elaboração do *savoir-faire* metodológico acumulado ao longo do desenvolvimento do léxico-gramática, paralelo ao da lingüística de corpus. Por um lado, precauções metodológicas apropriadas permitiram encontrar soluções à falta de rigor nas observações que suscitou a revolução dos corpus; por outro lado, a fonte de informações crescente que constituem as coleções de textos foi cada vez mais utilizada.

Os resultados obtidos pela aplicação desse conjunto de métodos surpreenderam mais de um lingüista. Trata-se realmente de regras, mas as diferenças entre entradas lexicais e entre construções compõem um cáos de irregularidades muito mais importante do que podíamos prever. O modelo do léxico-gramática permite evidentemente de representar o fato de que dois elementos lexicais possuem exatamente as mesmas propriedades sintático-semânticas, mas, quando isso acontece, o exame de outras propriedades leva, quase sempre, a encontrar diferenças entre as duas entradas. Ora, nenhuma teoria lingüística tinha previsto até então uma tal diversidade, que é contrária à intuição mais ou menos unânime de lingüistas e locutores. Um vasto empreendimento de coleção de observações era, portanto, tão necessário em Lingüística quanto em Física e em Biologia, e continua sendo.

Ademais, esse resultado espantoso conduz a relativizar a noção de regra. Parece, de fato, que temos uma clara tendência intuitiva a exagerar a generalidade de regras em matéria de sintaxe e de semântica, já que somos tão numerosos a ser surpreendidos ao ver que têm tantas exceções. Isso demonstra mais uma vez a necessidade de considerarmos dados observáveis susceptíveis de contrabalançar essa tendência, colocando em evidência os contra-exemplos às regras que nos parecem prevalecer.

Desse ponto de vista, uma das tradições "culturais" da gramática gerativa constitui um desvantagem: a idealização da noção de regra e a valorização extrema da generalidade das regras. O perigo dessa tendência é o de perder de vista que uma regra geral, mas que não se acha em conformidade com a realidade, é uma generalização apressada, sem valor científico.

5.2. No tratamento automático das línguas

Se bem que o domínio do tratamento automático das línguas não se tenha estruturado sobre o modelo da Lingüística, nele encontramos também uma oposição entre duas abordagens metodológicas que correspondem de uma certa maneira à oposição entre lingüística introspectiva e lingüística de corpus.

Na abordagem dita simbólica, utiliza-se um modelo formal no qual representa-se (por símbolos) as noções lingüísticas ou cognitivas pertinentes, suas relações, e as regras a serem aplicadas durante os tratamentos. O modelo e as regras de manipulação são construidos manualmente Por exemplo, para a tradução automática, essas regras podem descrever as correspondências de língua a língua ou as construções sintáticas em uma das línguas. A lingüística introspectiva, na medida em que produz resultados formais, alimenta a abordagem simbólica.

Na abordagem chamada probabilista, que é majoritária, parte-se de um corpus que serve de fonte de exemplos do tratamento a ser efetuado e produz-se automaticamente, por

análise estatística, dados numéricos com a ajuda dos quais programas simulam um comportamento o mais próximo possível do ilustrado pelo corpus. Por exemplo, para a tradução automática, acha-se um corpus de textos numa língua e sua tradução na outra, e fornece-se-os a um programa de aprendizagem automática. A produção de regras é automatizada sob a forma de um tratamento estatístico do corpus de aprendizagem. Trata-se, pois, de uma "industrialização" daquilo que é artesanal na abordagem simbólica.

A abordagem híbrida, que consiste a combinar as duas precedentes, é pouco desenvolvida, e consiste geralmente em introduzir algumas gotas simbólicas num mar probabilista.

Os argumentos que asseguraram a popularidade da abordagem probabilista junto aos engenheiros estão ligados à automaticidade do processo, que não só traz uma garantia de objetividade como também limita o custo de realização dos produtos. Esses dois argumentos lembram uma parte do debate entre lingüística introspectiva e lingüística de corpus.

O argumento da objetividade corresponde a uma das críticas classicamente feitas pelos linguístas de corpus (cf. seção 4.1). Aqui, ainda, os adeptos da abordagem probabilista falam de objetividade mas jamais de reprodutibilidade; nunca avaliam a eficácia das precauções metodológicas que certos adeptos da lingüística introspectiva ou da abordagem simbólica tomam para ter certeza que seus resultados descrevem bem uma língua e não um experimentador. Talvez ignorem essas precauções; mas a ignorância não justifica uma escolha científica.

O argumento do custo evoca a oposição entre os estudos observacionais e os estudos experimentais em pesquisa médica: essas últimas necessitam maior rigor e são susceptíveis de fornecer resultados mais precisos, mas o preço a pagar por essa diferença de qualidade é que têm um custo maior. Da mesma maneira, elaborar manualmente modelos formais da sintaxe, da semântica e do léxico é mais custoso do que mandar realizar modelos numéricos por programas de análise estatísticas. Entretanto, não se pode abordar a questão do custo sem abordar a da qualidade: a relação custo/qualidade é um critério de avaliação mais pertinente do que o do custo. Ora, os produtos da abordagem probabilista fornecem apenas resultados rudimentares, que são suficientes para aplicações comercialmente rentáveis, mais elementares, como os motores de pesquisa. Com efeito, a abordagem probabilista é tecnicamente incompatível com a manipulação de estruturas complexas, quer dizer, possuindo numerosos parâmetros, cada um dos quais pode ter numerosos valores. Essse é bem o caso dos objetos de base da sintaxe: entradas lexicais, construções sintáticas, contextos. O argumento do custo está, pois, longe de ser decisivo.

Citamos um terceiro e último argumento, freqüentemente citado nas publicações especializadas na abordagem probabilista: o estudo descritivo da língua é qualificado como sendo "longo e aborrecido", razão invocada para excluir a abordagem simbólica. (Essa expressão fai tantas vezes repetida que quase faz parte do jargão da área.) Eis aí outro argumento bem pouco convincente. Se um autor que invoca esse argumento se dedicou à pesquisa em informática, podemos crer sem dificuldade que tem pouco gosto pela descrição lingüística. Mas em quê isso é um obstáculo à sua utilização de dados lingüísticos construídos por outros pesquisadores, que, pelo contrário, apreciam essa atividade? Se o autor em questão não gosta de cozinhar, vai ele praticar o jejum? Da mesma forma, se a realização de um sistema de qualidade o necessita, porque não utilizar os resultados obtidos por outra disciplina? Supondo que nosso autor seja de boa fé, falta-lhe muita imaginação. O fato que revistas e colóquios dentre os mais prestigiados publiquem centenas de artigos nos quais os autores reproduzem este argumento parece-nos vergonhoso para seus comitês de seleção.

Se os méritos científicos e técnicos da abordagem probabilista são assim tão discutíveis, como explicar a sua popularidade? Talvez seja porque permite a seus adeptos economizar o esforço de uma colaboração entre duas disciplinas, a informática aplicada e a

lingüística descritiva. Se essa explicação é válida, o fundo do problema pode ser o mesmo da disputa entre lingüística introspectiva e lingüística de corpus: um receio de colaborar entre duas abordagens metodológicas que, porém, são compatíveis.

6. Conclusão

A oposição de Croft (1993, 1998) entre "método experimental" e "método observacional" renova o velho debate entre lingüística introspectiva e lingüística de corpus, suscitando um paralelo com as ciências experimentais às quais Croft empresta os termos. O exemplo do léxico-gramática, um método de descrição sintático-semântica cujos fundamentos referem-se explicitamente às ciências experimentais, é particularmente esclarecedor nesse debate e nesse paralelo. Lembremos o essencial dos ensinamentos aos quais temos proposto chegar.

- A formulação de regras conformes à realidade do uso de uma língua é uma técnica que não se resume a uma mera observação de exemplos.
- Necessita não só uma observação intensiva de exemplos, como também precauções metodológicas rigorosas nessa atividade de observação.
- As tradições aparentemente opostas da línguística introspectiva e da lingüística de corpus são, pois, complementares e de natureza a se combinar para favorizar o sucesso de tal empresa, sendo contra-produtivo excluir um ou outro.
- A metodologia do léxico-gramática fornece um exemplo concreto e produtivo de resultados.

Essas reflexões convidam os lingüistas a superar seu receio de combinar os dois tipos de método.

Referências

Aluísio, Sandra M., Gisele M. Pinheiro, Marcelo Finger, Maria das Graças V. Nunes, Stella E.O. Tagnin. 2003. The Lácio-Web Project: overview and issues in Brazilian Portuguese corpora creation. *Proceedings of Corpus Linguistics*, Lancaster, UK, pp. 14-21.

Baker, Collin F., Charles J. Fillmore, Beau Cronin. 2003. The Structure of the Framenet Database, *International Journal of Lexicography* 16.3, pp. 281-296.

Boons, Jean-Paul, Alain Guillet, Christian Leclère. 1976. *La structure des phrases simples en français. 1. Constructions intransitives*, Genève : Droz.

Croft, William. 1993. "Functional-typological theory in its historical and intellectual context", *Sprachtypologie und Universalienforschung* 46, pp. 15-26.

Croft, William. 1998, La théorie de la typologie fonctionnelle dans son contexte historique et intellectuel, *Verbum* 1998/3, pp. 289-307.

Fairon, Cédrick, John V. Singler. 2006. "I'm like, 'Hey, it works!': Using GlossaNet to find attestations of the quotative (be) like in English- language newspapers", in A. Renouf and A. Kehoe (eds). *The Changing Face of Corpus Linguistics*. Language and Computers 55. Rodopi, Amsterdam/New York, pp. 325-336.

Fillmore, Charles. 1992. "'Corpus linguistics' vs. 'Computer-aided armchair linguistics'". *Directions in Corpus Linguistics*, Mouton de Gruyter, pp. 35-60. (Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm.)

Gross, Maurice. 1975. *Méthodes en syntaxe*, Paris : Hermann.

Gross, Maurice. 1981. « Les bases empiriques de la notion de prédicat sémantique », *Langages* 53, pp. 7-52, Paris : Larousse.

Gross, Maurice. 1984. "A linguistic environment for comparative Romance syntax". In *Papers from the XIIth Linguistic Symposium on Romance Languages*, P. Baldi (ed.), pp. 373-446, Amsterdam/Philadelphia: John Benjamins.

Gross, Maurice. 1994. “Constructing Lexicon-grammars”. *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford University Press, pp. 213-263.
Guillet, Alain, Christian Leclère. 1992. *La structure des phrases simples en français. 2. Constructions transitives locatives*, Genève : Droz.
Harris, Zellig. 1964. “The Elementary Tranformations”, *Transformations and Discourse Analysis Papers* 54, in Harris, Zellig S. 1970, *Papers in Structural and Transformational Linguistics*, Dordrecht: Reidel, pp. 482-532.
Harris, Zellig. 1976. *Notes du cours de syntaxe*, Paris : Seuil.
Kipper-Schuler, Karin, Anna Korhonen, Neville Ryant, Martha Palmer. 2006. Extending VerbNet with Novel Verb Classes. *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), Genoa.
Labelle, Jacques. 1990. “Norms and variants en French”, *Lingvisticae Investigationes* 13:2, pp. 281-306, Amsterdam/Philadelphia: John Benjamins.
Lamiroy, Béatrice, Christian Leclère, Jean René Klein, Jacques Labelle. 2003. “Expressions verbales figées et variation en français: le projet BFQS”, *Cahiers de lexicologie* 83-2, pp. 153-172.
Paumier, Sébastien. 2006. *Unitex Manual.* http://univ-mlv.fr/~unitex.
Popper, Karl. 1959. *The Logic of Scientific Discovery*, Basic Books, New York.
Renouf, Antoinette. 2003. 'WebCorp: providing a renewable data source for corpus linguists', in S. Granger & S. Petch-Tyson (eds.), *Extending the scope of corpus-based research: new applications, new challenges.* Amsterdam: Rodopi.
Salkoff, Morris. 1983. Bees are swarming in the garden: a systematic synchronic study of productivity. *Language* 59, pp. 288-346.
Vale, Oto A. 2001. *Expressões Cristalizadas do Português do Brasil: uma proposta de tipologia*, Tese de doutorado, Universidade Estadual Paulista, Araraquara.

Resumo

A oposição colocada por Croft (1993) entre “método experimental” e “método observacional” renova o velho debate entre lingüística introspectiva e lingüística de corpus, suscitando um paralelo com as ciências experimentais, às quais Croft empresta os termos. O exemplo do léxico-gramática, um método de descrição sintático-semântica cujos fundamentos referem-se explicitamente às ciências experimentais, confirma, se fosse necessário, que a formulação de regras conformes à realidade do uso de uma língua não se resume a uma simples observação de exemplos, e que necessita não só uma observação intensiva de exemplos, como também o uso de precauções metodológicas rigorosas nessa atividade de observação. As tradições aparentemente opostas da lingüística introspectiva e da lingüística de corpus são, portanto, complementares e de natureza a se combinar para favorecer o sucesso de uma tal empresa. Essas reflexões convidam os lingüistas a superar seu receio histórico de combinar os dois tipos de método. Da mesma maneira, no tratamento automático das línguas, a maior parte da comunidade limita-se à abordagem probabilista, renunciando a uma colaboração protencialmente fecunda entre a informática aplicada e a lingüística descritiva.