CÓRPUS, INTROSPECÇÃO E O OBJETO DA DESCRIÇÃO GRAMATICAL

Mário A. Perini[1]
Gabriel de Ávila Othero[2]

RESUMO

Neste texto, discutimos brevemente o problema metodológico que o linguista enfrenta ao analisar fatos da língua com base em suas próprias intuições (ou intuições de falantes nativos) ou com base em córpus da língua em estudo. Apontamos para o fato de que o uso dos julgamentos intuitivos apresenta alguns perigos; porém, o uso exclusivo de córpus linguístico igualmente traz suas complicações e limitações.

Palavras-chave: Julgamentos de aceitabilidade. Intuição. Córpus linguísticos. Objeto da Linguística.

O OBJETO DE ESTUDO DA LINGUÍSTICA

Segundo Itkonen (1974), o objeto da Linguística são os julgamentos de uma comunidade a respeito de determinadas sequências sonoras. Itkonen defende a ideia de que as ciências naturais não são um modelo adequado para o estudo científico da linguagem. Em particular, não é possível refutar ou confirmar hipóteses linguísticas através da simples observação de eventos; as hipóteses só podem ser avaliadas em confronto com julgamentos de correção por parte de membros da comunidade linguística. Desse modo, o problema metodológico de refutar ou confirmar hipóteses é qualitativamente distinto para a Linguística e para as ciências naturais. E, como julgamentos podem ser equivocados ou mentirosos (ao contrário de eventos físicos, por exemplo), cada um deles não pode ser mais do que uma *indicação probabilística*, i. e., um julgamento nunca pode ser ou servir como uma prova ou uma refutação cabal.

Isso, parece-nos, significa que a descrição linguística não se dirige a regularidades observadas em um córpus de ocorrências; essas só podem ser entendidas como indicações dos julgamentos da comunidade[14]. O mesmo vale, evidentemente, para os julgamentos explicitados pelos falantes: não passam de indicações. Nenhum desses recursos pode ser tomado como decisivo; podem apenas ter um poder aditivo, aumentando a probabilidade de que a hipótese em exame sobre determinado *fato* da língua seja correta (ou incorreta). Obviamente, em muitos casos, essa probabilidade é muito alta, a ponto de se confundir, na prática, com uma certeza. É o que acontece, por exemplo, quando afirmamos que a frase [1] é correta em português, enquanto a frase [2] não é:

[1] Esse menino gosta de pizza.

[2] * Essa menino gostávamos pizza.

É claro que nem sempre as coisas são assim tão claras, e, em muitos casos, o grau de probabilidade é relativamente baixo. Vejamos, por exemplo, dois exemplos clássicos de Chomsky (1957: 15):

[3] Colorless green ideas sleep furiously.

[4] Furiously sleep ideas green colorless.

Para Chomsky (1957: 15),

> a noção de “gramatical em inglês” não poderá ser identificada, de maneira alguma, com a de “alta ordem de aproximação estatística em inglês”. Parece razoável aceitar que nem a sentença [3] nem a sentença [4] (e nenhuma parte dessas sentenças) tenha ocorrido em inglês. Logo, em qualquer modelo estatístico voltado para a gramaticalidade, essas sentenças seriam excluídas com base nos mesmos motivos, como igualmente “remotas” em inglês. Ainda assim, [3], embora sem sentido, é gramatical, enquanto [4] não é. Vendo essas sentenças, um falante do inglês irá ler [3] com uma entonação normal de sentença, mas irá ler [4] com uma entonação falha em cada palavra; na verdade, com o mesmo padrão de entonação dado a qualquer sequência de palavras que não apresentem relação entre si[3].

Para Chomsky, a não-aceitabilidade de [3] pode se dar por causa da estranheza semântica ou pragmática da sentença na língua; ainda que uma frase possa ser considerada gramatical – em sentido gramaticalmente estrito –, ela pode ser inaceitável, caso de [3] (ver boa discussão sobre o assunto no primeiro capítulo de Radford 1981 e 1988).

Um ponto em que o problema é bastante agudo é justamente quando se marca uma frase ou uma construção como inaceitável na língua. O problema vem da dificuldade de corroborar essa inaceitabilidade. Digamos que os falantes tendem a rejeitar uma determinada frase: isso não pode ser decisivo, porque o julgamento do falante é condicionado por diversos fatores, dos quais um é simplesmente a dificuldade de imaginar uma situação estrutural em que determinada sequência seja aceitável. Recorremos novamente a um trecho de Chomsky (1977: 4):

> Podemos fazer um julgamento intuitivo sobre alguma expressão linguística ser estranha ou mal formada. Mas não podemos, em geral, saber, pré-teoricamente, se essa má-formação é uma questão de sintaxe, semântica, pragmática, crença, limitação da memória, estilo, etc., ou mesmo se essas são categorias apropriadas para interpretar o julgamento em questão. É fato óbvio e não controverso que os julgamentos dos informantes e outros dados não se adéquam perfeitamente em categorias claras, como sintaxe, semântica, etc.

Para vermos um exemplo concreto: podemos rejeitar a frase [5] porque não nos ocorre que Zé poderia ser o irmão mencionado – isto é, *meu irmão* seria aposto de *Zé*[4].

[5] Convocaram Zé meu irmão.

Uma frase pode também ser rejeitada, embora seja não só aceitável mas frequente, quando vai contra certos preconceitos linguísticos que, suspeitamos, se originam principalmente do uso da língua escrita. É o caso de frases de tópico do tipo [6]:

[6] Esse cano sai fumaça.

Naturalmente, essa frase parece irregular em termos de escrita, e não tem análise tradicional aplicável. No entanto, como Pontes (1986; 1987) mostrou, tais frases[5] são correntes no português falado e, quando enunciadas, não causam nenhuma estranheza. No entanto, se a submetermos ao julgamento de falantes nativos, especialmente com alto grau de escolarização, a frase pode ser julgada como “estranha” ou “inaceitável”. Ora, esse é o paradoxo do observador, já conhecido desde os primeiros trabalhos de Labov, pelo menos: ao examinar a fala de falantes nativos da língua, o observador influencia a produção e os julgamentos de aceitabilidade de seus informantes. É comum que o falante, confrontado, muitas vezes, com a própria fala, não aceite determinada frase como “modelo” da língua e acabe julgando-a inaceitável ou, até mesmo, agramatical[6].

Outro fator provém da interferência de fatores extralinguísticos na produção do enunciado. Assim, a frase em teste pode ser

[7] O meu menino caiu o livro no chão.

À primeira vista, a frase parece agramatical ou, ao menos, inaceitável. No entanto, se ela ocorrer em algum córpus do português, alguém poderia vir a argumentar que é ela gramatical ou aceitável. Aqui fica claro que a ocorrência por si só não é critério definitivo de decisão da questão. O falante estrutura seu enunciado à medida que fala e, como não pode voltar atrás e redizer o que já disse, pode acontecer que haja uma mudança de plano no meio da sentença, algo como [8]:

[8] O meu menino // caiu o livro no chão.

Ou pode acontecer que o falante enuncie o verbo *caiu* por lapso e, em vez de abandonar essa ocorrência e repetir um verbo mais adequado, como em [9], ele decida simplesmente ir em frente com a frase, já que o significado vai ficar toleravelmente claro; nesse caso, ele enuncia [7].

[9] O meu menino caiu // derrubou o livro no chão.

Em casos como esse, apesar de a frase ocorrer, temos de marcá-la como inaceitável para efeitos de análise gramatical. Esses são apenas alguns dos fatores que podem interferir na produção de um enunciado. A fala é sujeita a pressões de diversos lados: urgência de tempo, disputa de espaço na memória com outras tarefas (como a de planejar logicamente o discurso, por exemplo), a necessidade de manter o turno da fala, etc. Tudo isso contribui para moldar o enunciado, que é, portanto, o produto de muito mais do que o conhecimento do código aprovado pela comunidade linguística.

2. CÓRPUS E INTROSPECÇÃO

Para verificar como um córpus é, por si mesmo, incapaz de fornecer os dados necessários à análise de uma língua, basta observar as repetições, hesitações e outras "irregularidades" que ocorrem em qualquer córpus de língua falada, mas que não podem ser incluídas em uma gramática. Qual é o critério que o pesquisador aplica para excluir da descrição da língua certas repetições (como em *um jogo de... de... de... tênis*), ao mesmo tempo incluindo outras (como em *ela é muito, muito inteligente*), a não ser o recurso a seu conhecimento internalizado da língua, através do que chamamos, no nosso ofício, "intuições"?[7]

O verdadeiro objeto da Linguística (ainda seguindo Itkonen) é impossível de atingir diretamente. Ele reside em uma espécie de contrato social e, em última análise, na mente dos falantes[8]; só pode ser atingido através de indicações indiretas, ou seja, através de seus efeitos no comportamento dos falantes (por exemplo, nos julgamentos de aceitabilidade e gramaticalidade), assim como na observação das formas ocorrentes, por exemplo, em um córpus. Nenhuma dessas fontes de evidência é pura e confiável. No entanto, não dispomos de nenhuma outra; temos de nos virar com o que está aí para trabalhar com os fenômenos da linguagem.

As soluções dadas acima para as várias situações são, portanto, apenas hipóteses. Por exemplo, no caso de [7] formulamos uma hipótese sobre o que acontece na cabeça do falante ao enunciar essa sequência. Mas pode perfeitamente ser outra coisa: por exemplo, pode ser que haja uma tendência, observada em alguns falantes, a enriquecer a valência do verbo *cair*, de modo a

transformar esse verbo em um transitivo-ergativo, do tipo de *engordar*, *esquentar*, etc. Aí temos que nos valer de observações repetidas. Considerando que os fatores de desempenho hipotetizados para explicar a ocorrência de [7] são acidentais, e sua repetição frequente é implausível, podemos refazer a análise caso se note que o uso de *cair* com objeto direto é particularmente frequente – chegaríamos à conclusão de que se trata de um fenômeno gramaticalmente relevante.

Voltando à questão do uso do córpus e da introspecção, gostaríamos de deixar claro que nenhum desses recursos é dispensável em princípio. No entanto, parece que às vezes se defende o uso exclusivo do córpus, partindo da ideia de que os julgamentos de aceitabilidade seriam artificiais. É a velha controvérsia entre o linguista de corpus *versus* o linguista de poltrona, como chamou Fillmore (1992) – que se declara "um linguista de poltrona que se recusa a abandonar seus velhos hábitos, mas que encontra benefícios em ser um consumidor de alguns dos recursos criados pelos linguistas de córpus" (p. 35) – descrição que se aplica também ao nosso atual modo de trabalhar em Linguística. Ainda concordamos com Fillmore quando ele continua:

> eu acho que não existem córpus, por maiores que sejam, que contenham todas as informações sobre todas as áreas do léxico e da gramática do inglês que eu quero explorar; todos os que vi são inadequados. (...) cada córpus que eu já tive a chance de examinar, no entanto, me ensinou alguns fatos que eu não podia imaginar encontrar de qualquer outra maneira. A minha conclusão é que os dois tipos de linguistas precisam um do outro. Ou melhor: que os dois tipos de linguistas, sempre que possível, devam existir na mesma pessoa. (p. 35)

Em outras palavras, precisamos equilibrar as duas "fontes" de dados: a intuição própria – aliada ao julgamento de outros falantes nativos da língua – e um córpus da língua em estudo. Não há dúvida de que o uso dos julgamentos intuitivos apresenta perigos bem grandes; contudo, o uso do córpus apresenta igualmente perigos (alguns dos quais apontamos acima), e, além do mais, o trabalho com córpus nem sempre é praticável.

Gross (1975: 19) define assim o uso dos julgamentos na pesquisa:

> testar a aceitabilidade de uma sequência é levar a efeito um experimento. A construção de exemplos e contra-exemplos constitui a atividade experimental do linguista que verifica a teoria de certos fenômenos.

É claro que a evidência de laboratório tem suas limitações, e o pesquisador precisa ficar consciente delas. Mas uma ciência que trata de fenômenos tão numerosos e variados como a Linguística não pode dispensar os experimentos. Um químico, por exemplo, nunca pensaria em esperar que dois elementos entrassem em contato na natureza para ver se podem se combinar; esse procedimento simplesmente não é prático e inviabilizaria boa parte da atividade de pesquisa. O mesmo vale para o linguista, em particular para o sintaticista[9]. Nossa evidência vem de diversas fontes, e isso é inevitável, dado o caráter dos dados com que temos de lidar. Como se vê, nada vem de graça. Nossa evidência vem de fontes suspeitas, e, por isso, todo cuidado é pouco[10].

CONSIDERAÇÕES FINAIS

Para finalizar – e para ilustrar a necessidade do recurso à introspecção –, vamos tomar um exemplo que tem a ver com a definição dos papéis temáticos – uma área importante, ainda pouco esclarecida no momento. É importante preservar o princípio segundo o qual os papéis temáticos são acessíveis à intuição direta. Ou seja, a diferença de papéis temáticos dos sujeitos dos pares de frases [10] e [11] e [12] e [13] precisa ser clara para o falante comum, sem o intermediário de alguma operação teórica. É indispensável vincular as duas pontas da análise (som e significado) a fatos concretos, observáveis.[11]

[10] O fazendeiro chorou.

[11] O fazendeiro engordou.

[12] Joãozinho quebrou o espelho.

[13] Joãozinho quebrou a perna.

É claro que isso é muito mais difícil no campo semântico do que no campo fonético. Mas a situação é, em princípio, a mesma: trata-se de fazer observações e de condicionar a descrição dos fatos linguísticos a uma fidelidade extrema a essas observações[12]. Daí a necessidade de definir os papéis temáticos em termos operacionais, de maneira a colocá-los ao alcance da intuição. Temos aqui um ponto em que a intuição é indispensável na pesquisa linguística. Essas percepções semânticas não podem ser depreendidas diretamente de um córpus, mas apenas através de algum tipo de introspecção e julgamento linguístico[13]. Parece-nos que Gross estava falando disso quando afirmou que

> alguns gramáticos tentaram tornar as noções de significado mais abstratas, o que lhes teria dado maior generalidade (é o caso da noção de “objeto”), mas nesse caso essas noções já não são operacionais: seu caráter intuitivo desaparece e já não é possível determinar se elas se aplicam ou não a numerosas formas dadas. (GROSS 1975: 30-31)

Achamos que a tentativa de amarrar os papéis temáticos muito estreitamente à sua representação sintática tem exatamente esse efeito de removê-los da área da intuição; por isso, ao discutir os critérios de delimitação dos papéis temáticos, Perini (2008, especialmente no capítulo 7) propôs o **critério de semelhança semântica** como uma das condições para reunir duas relações conceptuais sob um mesmo rótulo temático. Esse critério tem o papel de evitar a generalização excessiva criticada por Gross. Ele se resume à exigência de que as diversas realizações de um mesmo papel temático devem caber dentro do mesmo esquema (ou “frame”, para utilizar a nomenclatura preferida por Fillmore, 2007).

É claro que o critério de semelhança semântica depende da intuição para ser aplicado. Embora isso seja, sem dúvida, um problema, dado o caráter às vezes vacilante e obscuro dos julgamentos intuitivos, não vemos outra maneira de ancorar a ponta semântica dos dados na realidade: no caso, realidade interior, difícil de atingir com precisão, mas nem por isso menos objetiva.

CORPUS, INTROSPECTION AND THE OBJECT OF GRAMMATICAL DESCRIPTION

ABSTRACT

In this paper, we briefly discuss the methodological problem linguists have to face when analyze language facts based on their own intuitions (or intuitions of native speakers) or based on a corpus of the target language. We point to the fact that the use of intuitive judgments presents some risks, but the exclusive use of linguistic corpus also brings its complications and limitations.

Keywords: Judgments of acceptability. Intuition. Corpus. Object of linguistics.

NOTAS

1 Universidade Federal de Minas Gerais – UFMG.

2 Pós-doc, Universidade Federal do Rio Grande do Sul – UFRGS.

3 Contudo, Pereira (2002) contestou essa afirmação de Chomsky: ele afirma que a sentença [4] é 200.000 vezes menos provável de ocorrer em um córpus do inglês do que a sentença [3]. Ou seja, para ele, essas frases não "seriam excluídas com base nos mesmos motivos, como igualmente 'remotas' em inglês".

4 Em mineiro, é muito comum se dizer *Zé meu irmão* sem nenhuma cesura entoacional entre os dois elementos.

5 Tradicionalmente chamadas frases de anacoluto.

6 Cf. Tarallo (1990), Monteiro (2000) e Labov (2008), por exemplo.

7 Daí vem uma grande dificuldade enfrentada por linguistas computacionais que desejam criar programas robustos de análise sintática, por exemplo. É muito difícil – para não dizer impossível – alimentar um programa de análise sintática automática com uma gramática que consiga analisar frases fragmentadas, com repetições, com elementos parentéticos, etc. Cf. discussão em Othero (2008; 2009).

8 Parece que Saussure enfatizou o aspecto social, e Chomsky o aspecto mentalista; a nosso ver, ambos estão corretos, embora olhem o objeto de ângulos diversos.

9 Ou para o semanticista, se é que os dois se distinguem atualmente. A situação de um foneticista ou de um analista do discurso pode ser diferente.

10 O próprio Chomsky reconhece a inevitabilidade de lançar mão de diversas fontes de evidência (Chomsky, 1980: 109).

[11] Isso decorre do princípio da Sintaxe Simples, de Culicover & Jackendoff (2005); ver detalhes em Perini (2008, cap. 2).

[12] Em outras palavras, "atingir a **adequação observacional**", para usar a terminologia de Chomsky (1964).

[13] Por exemplo, a teoria denominada **Lexicase**, de Starosta (1988), embora muito interessante em outros aspectos, falha totalmente nesse particular, pois joga com papéis temáticos ("casos") seriamente contra-intuitivos.

[14] Isso coloca Itkonen em choque com as ideias de alguns estruturalistas, notadamente Harris, para quem "a pesquisa principal da linguística descritiva (...) é a distribuição ou arranjo dentro do fluxo da fala de certas partes ou traços em relação a outras" (1951: 5).

REFERÊNCIAS

CHOMSKY, Noam. *Syntactic Structures*. Berlin: Mouton de Grutyer, 1957. (Edição consultada: 2002).

_____. Three adequacies, conferência no *IX International Congress of Linguists*, 1964.

_____. *Essays on form and interpretation*. North Holland, 1977.

_____. *Reflections on language*. New York: Pantheon Books, 1980.

CULICOVER, Peter W.; JACKENDOFF, Ray S. *Simpler syntax*. Oxford: Oxford University Press, 2005.

FILLMORE, Charles J. "Corpus linguistics" or "Computer-aided armchair linguistics". In: SVARTVIK, Jan (ed.) *Directions in Corpus Linguistics, Proceedings of Nobel Symposium* 82, Stockholm, 4-8 August 1991, Berlin/NY: Mouton de Gruyter, 1992.

FILLMORE, Charles J. Valency issues in FrameNet. In: HERBST, T.; GÖTZ-VOTTELER, K. (orgs.) *Valency: theoretical, descriptive and cognitive issues*. Berlin: Mouton De Gruyter, 2007.

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

HARRIS, Zellig S. *Structural linguistics*. Chicago: Phoenix Books, 1951.

ITKONEN, Esa. *Linguistics and metascience*. Kokemäki: Societas Philosophica et Phæno-menologica Finlandiæ, 1974.

JOHNSON, Stephen B. *The legacy of Zellig Harris - language and information into the 21st century*. Volume 2: Mathematics and computability of language. Amsterdan / Philadelphia: John Benjamins, 2002.

LABOV, William. *Padrões sociolinguísticos*. São Paulo: Parábola, 2008.

LEES, R. *The grammar of English nominalizations*. Mouton, The Hague, 1960.

MONTEIRO, José Lemos. *Para compreender Labov*. Petrópolis: Vozes, 2000.

OTHERO, G. A. Linguista puro vs. linguista computacional: revisitando a distinção entre linguista de poltrona e linguista aplicado. *Domínios de Lingu@gem – Revista Eletrônica de Linguística* v. 3, 2008.

_____. *A gramática da frase em português: algumas reflexões para a formalização da estrutura frasal em português*. Porto Alegre: Edipucrs, 2009.
PEREIRA, F. Formal grammar and information theory: together again?. *Philosophical Transactions of the Royal Society*, 2000.

PERINI, Mário A. *Estudos de gramática descritiva: as valências verbais*. São Paulo: Parábola, 2008.

PONTES, Eunice. *Sujeito: da sintaxe ao discurso*. São Paulo: Ática, 1986.

_____. *O tópico no português do Brasil*. Campinas: Pontes, 1987.

RADFORD, Andrew. *Transformational syntax: a student's guide to Chomsky's extended standard theory*. Cambridge: Cambridge University Press, 1981.

_____. *Transformational grammar*. Cambridge: Cambridge University Press, 1988.

STAROSTA, Stanley. *The case for Lexicase: an outline of Lexicase grammatical theory*. London: Pinter, 1988.
TARALLO, Fernando. *A pesquisa sociolinguística*. São Paulo: Ática, 1990.