

# Nota introdutória do editor

Stefan Th. Gries  
University of California, Santa Barbara

## 1 Introdução

Se alguém perguntar a um linguista de corpus há quanto tempo essa área de estudos existe, duas respostas são ouvidas mais frequentemente. Uma seria que os métodos da linguística de corpus existem há muito tempo, por exemplo, nas concordâncias bíblicas ou no trabalho de Käding (1897); também seriam lembrados os linguistas comparativistas europeus ou os estruturalistas norte-americanos da primeira metade do século XX, dentre outros. A outra resposta para a pergunta inicial seria a de que a linguística de corpus somente começou a ser perfilada, no cenário europeu, com o trabalho de Firth (1951) sobre colocações, ou o trabalho da Pesquisa sobre o Uso do Inglês,<sup>1</sup> além da frente norte-americana com o trabalho de Fries (1952) sobre o inglês americano falado.

Independente da opção que se fizer por um desses dois pontos de vista – ambos são, provavelmente, sob certos ângulos e os linguistas de corpus podem adotar qualquer um dos dois, dependendo dos objetivos retóricos que tenham – não seria exagero afirmar que foi apenas nos últimos vinte anos que a linguística de corpus realmente se consolidou, tornando-se uma dos métodos mais amplamente utilizados na linguística. Isso é visível em diferentes níveis:

- no nível de **recursos**: os desenvolvimentos tecnológicos ocorridos facilitaram a criação dos primeiros megacorpora do tipo exemplificado pelo British National Corpus;
- no nível do papel desempenhado pelos dados de corpora no desenvolvimento e refinamento de **teorias linguísticas** mais abrangentes, i.e., em trabalhos que vão além da mera descrição linguística. Apesar de tais desenvolvimentos ainda sofrerem resistência por alguns – como exposto na visão de que a linguística de corpus é uma “mera” metodologia – (cf. Worlock Pope (2010) na edição especial do *International Journal of*

---

<sup>1</sup> Survey of English Usage.

*Corpus Linguistics* sobre o assim chamado discurso de *bootcamp*<sup>2</sup>), é difícil ignorar as maneiras através das quais a linguística de corpus, por um lado, e a linguística cognitiva e a psicolinguística, por outro, se complementam;

- no nível da **metodologia estatística**: a tendência geral na linguística em direção a métodos quantitativos pode – finalmente! – ser também vista na linguística de corpus. Na verdade, eu já defendi em outras instâncias que, por ser a linguística de corpus essencialmente baseada em dados distribucionais e quantitativos, ela deveria ter liderado a atual revolução quantitativa na linguística, ao invés de haver delegado essa honra à psicolinguística principalmente ...;
- no nível da **competência dos profissionais** da área: muitos profissionais da área têm, há muito tempo, sido constringidos por algumas poucas ferramentas comerciais para a análise de corpora, as quais limitavam a capacidade dos pesquisadores de pensar para além do que era oferecido pelas opções dos softwares em questão. Entretanto, nesse momento, a área está se definindo e muitos pesquisadores têm-se voltado para ferramentas mais versáteis, poderosas e elegantes, como as oferecidas pela Natural Language Toolkit (cf. <<http://www.nltk.org>>) ou para linguagens de programação (cf. Gries (2009) para um exemplo), que finalmente permitem ao campo lidar com tipos complexos de dados de maneiras mais apropriadas que as utilizadas anteriormente.

A linguística de corpus atualmente já está bem estabelecida: a área possui vários periódicos internacionais com comitês científicos, suas próprias séries de livros com editoras internacionais, um circuito de congressos ativo, e métodos de pesquisa baseados em corpora têm contribuído para a maior parte dos subcampos da linguística. Isto também significa que pesquisadores não têm que incluir em seus trabalhos justificativas, ou mesmo defesas, do porquê de estarem utilizando dados de corpora – a linguística de corpus teve sucesso em se projetar como área e muitos de seus métodos hoje são dominantes (em um sentido positivo).

---

<sup>2</sup> NT: *bootcamp* refere-se aos cursos intensivos de formação sobre tratamento quantitativo de dados oriundos de corpora.

## 2 Este número temático internacional da RBLA

Apesar de sua história de sucesso, a linguística de corpus ainda necessita de maturação e de mais evolução, e este número temático da RBLA é devotado a esse tópico. Quando fui convidado a organizar esse número temático da RBLA sobre linguística de corpus, eu rapidamente decidi que *não* organizaria a típica publicação em que artigos de pesquisa “padrão” apresentam bons e significativos resultados – meu objetivo tornou-se organizar um volume especial que delineasse para onde o campo da linguística de corpus está se dirigindo; um volume que, por assim dizer, ofereça direcionamentos para a área, assim como boas sessões plenárias o fariam. Pensei que seria particularmente adequado que esse número especial fosse publicado em um periódico de acesso aberto, o qual torna as contribuições mais acessíveis do que as restrições de direitos de alguns periódicos comerciais frequentemente permitem, de forma que fiquei muito contente que o comitê editorial da RBLA tenha aceitado a minha proposta.

O próximo passo consistiu em identificar a amplitude de variação de campos que eu considero que muito se beneficiaram da e que muito contribuíram para a linguística de corpus, assim como persuadir pesquisadores proeminentes nesses campos a contribuírem com artigos para esse número temático, os quais respondessem à seguinte questão:

Em seu campo de pesquisa e em seu trabalho com corpora – e estou escrevendo para você por causa do seu trabalho sobre ..... – para onde você pensa que a área de linguística de corpus deve ir e/ou onde ela deve amadurecer, e por quê? Quais são os desenvolvimentos em termos de recursos, padrões, tecnologia, métodos, etc, que você crê serem essenciais 3/ou pelo menos desejáveis, e por que, ou o que poderíamos fazer sobre isso?

Tive a sorte de receber respostas afirmativas e encorajadoras de colegas de grande destaque em várias áreas da linguística ou em suas subdisciplinas, os quais estão listados na Tabela 1, abaixo. Cada um dos artigos delimita respostas para as perguntas-guia de forma individualizada, de modo geral oferecendo uma visão sumarizada do estado da arte, seguida por perspectivas, recomendações, listas de desideratos, estudos de caso, e muito mais, que deve oferecer à área elementos para reflexão para o futuro próximo – eles certamente o fizeram por mim.

Como uma observação final, um sincero “muito obrigado” é oferecido à minha organizadora associada, Heliana Ribeiro de Mello, sem quem esse número especial não teria se materializado. E, gostaria também de sinceramente agradecer, naturalmente, aos colegas que submeteram seus trabalhos e concordaram em contribuir com um número temático com um foco de certa forma pouco usual, os quais nos enviaram artigos inspiradores e repletos de reflexões, que claramente delimitam como a linguística de corpus pode se desenvolver ainda mais, de uma maneira que nenhum autor individualmente jamais poderia sugerir. Se esse número especial estimulá-lo a pensar e a planejar, o mérito será todo dos autores aqui representados.

TABELA  
Sumário desse número temático

Área/sub-disciplina: corpora e...	Autores
... métodos/pesquisa quantitativa	R Harald Baayen (Universidade de Tübingen)
... pesquisa em metáforas	Tony Berber Sardinha (PUC-SP)
... sociolinguística	Tyler Kendall (Universidade do Oregon, Eugene)
... dados multimodais	Dawn Knight (Universidade de Nottingham)
... linguística histórica	Merja Kytö (Universidade de Uppsala)
... aquisição/aprendizagem de segunda língua/língua estrangeira	Fanny Meunier (Universidade Católica de Louvain)
... pragmática da fala	Massimo Moneglia (Universidade de Florença)
... linguística cognitiva	John Newman (Universidade de Alberta, Edmonton)
... dialetologia	Benedikt Szmrecsanyi & Christoph Wolk (Instituto de Estudos Avançados de Freiburg)

## Referências

- FIRTH, J.R. *Papers in linguistics*, 1934-1951. Oxford: Oxford University Press, 1951.
- FRIES, C.C. *The structure of English: an introduction to the construction of English sentences*. New York: Harcourt Brace, 1952.
- GRIES, St.Th. *Quantitative corpus linguistics with R: a practical introduction*. London / New York: Routledge, Taylor & Francis Group, 2009.
- KÄDING, F. W. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: no publ., 1897.
- WORLOCK POPE, C. (Ed.). The bootcamp discourse and beyond. Special issue of the *International Journal of Corpus Linguistics*, v. 15, n. 2, 2010.