B. OFICINAS

OFICINA DE INTRODUÇÃO À LINGUÍSTICA DO CORPUS: RELATO DE EXPERIÊNCIA*

Rodrigo Esteves de LIMA-LOPES[1]

Resumo
Este texto tem por objetivo trazer um relato dos tópicos e exercícios realizados durante a oficina "Introdução à Linguística do Corpus" realizada durante a VIII Escola Brasileira de Linguística Computacional e o XIII Encontro de Linguística de Corpus (EBRALC-ELC). Procurou-se, assim, realizar uma síntese da experiência vivenciada a partir de três perspectivas: 1) dos conceitos teóricos, metodológicos e terminológicos; 2) das ferramentas disponíveis, com a exploração de algumas delas e 3) de exercícios práticos, típicos de uma oficina.

Palavras-chave: Oficina, Linguística do Corpus, Introdução, Conceitos-chave

Abstract
This essay aims at providing some insights on the themes discussed and exercicies made during the workshop "Introdução à Linguística do Corpus", which took place during the VIII Escola Brasileira de Linguística Computacional and the XIII Encontro de Linguística de Corpus (EBRALC-ELC). The experiences in such workshop were summarised by the means of three perspectives: 1) the basic theorethical and methodological concepts discussed; 2) the tools available; and 3) the exercices, which are tipical of an workshop of this kind.

Key-words: Workshop; Corpus Linguistics; Introduction, Key Concepts.

1. Introdução

Este artigo relata a experiência de realização da Oficina "Introdução à Linguística do Corpus" realizada durante a VIII Escola Brasileira de Linguística Computacional e o XIII Encontro de Linguística de Corpus (EBRALC-ELC) na Universidade Federal da Paraíba (UFPB) entre 3 e 6 de novembro de 2015. A oficina em questão teve por objetivo introduzir os princípios básicos da Linguística do Corpus, sendo que seu público-alvo foram pesquisadores em seu contato inicial com suas teorias e ferramentas. Inicialmente, discutiram-se conceitos básicos e terminologia, seguindo para algumas ferramentas, seu usos e funções.

A ideia de tal oficina surgiu da constatação de que boa parte das atividades e apresentações realizadas durante o Encontro e a Escola eram destinadas a um público que já navegasse com certa facilidade pela área. Assim, buscou-se construir uma experiência de troca, buscando utilizar o alicerce já existente nos participantes, ao mesmo tempo que se procurou refletir sobre questões que podem parecer óbvias para o pesquisador experiente, mas ainda se colocam como complexas e anuviadas para aqueles que iniciam o seu trabalho.

* Gostaríamos de agradecer à Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo financiamento dos anais da VII Escola Brasileira de Linguística Computacional e do XIII Encontro de Linguística de Corpus, processo nº 3472/2015-87.

[1] Doutor em Linguística Aplicada pelo Instituto de Estudos da Linguagem - IEL/UNICAMP, Mestre em Linguística Aplicada ao Ensino de Línguas (LAEL/PUCSP). Professor da Universidade Federal da Paraíba. Áreas de pesquisa incluem: Análise Crítica do Discurso, Gramática Sistêmico Funcional, Linguagens e suas Tecnologias, Novas Tecnologias em Comunicação, Tecnologias de Educação, Educação a Distância. Contato: talk2me@rll307.net / http://rll307.net.

A oficina contou com um total de 20 inscritos e 17 presentes, dos quais 12 não tinham experiência alguma na área, dois eram pesquisadores já experientes na área e três eram da área de Linguística Computacional e buscavam discussões interdisciplinares. Essa configuração de público rendeu importantes frutos, enriquecendo a discussão em duas perspectivas. A primeira foi a contribuição dos pesquisadores mais experientes, que aproveitaram a oportunidade para ilustrar os conceitos debatidos com sua experiência de trabalho. O segundo estava relacionado à rica discussão sobre como a Linguística Computacional e do Corpus possuem elementos em comum, enriquecendo a discussão terminológica.

O relato que se segue é organizado de forma reproduzir as etapas constantes da oficina. A seção "Reflexões teórico-metodológicas: o início" traz à baila alguns dos conceitos teóricos, metodológicos e terminológicos discutidos na etapa inicial. Parte-se então para seção "Algumas ferramentas importantes: explicando possibilidades", que traz um resumo das ferramentas discutidas durante a oficina. Por seu turno, a seção "Exercícios guiados: sedimentando os conceitos com a prática" traz a lista de exercícios realizada durante a oficina. O trabalho se encerra com uma avaliação da experiência.

Antes de iniciar, gostaria de chamar a atenção para o fato de que este texto não pretende ser um tratado teórico/prático sobre a linguística do corpus. Ele é apenas um relato sistemático das discussões e da experiência de realização da referida oficina. Assim, muitos dos temas aqui mencionados podem (e devem) ser aprofundados pelo leitor, algo possível com a ampla literatura existente na área, uma vez que a dinâmica de uma oficina de quatros horas e o exíguo espaço de publicação fez com que muitos fossem tratados apenas *en passant*.

2. Reflexões teórico-metodológicas: o início

A oficina foi desenhada de forma a propiciar uma introdução rápida ao histórico da Linguística do Corpus, tendo como base principal:

- Sua origem no contextualismo britânico e sua relação de parentesco com a Gramática Sistêmico Funcional (STUBBS, 1996)
- A importância do contexto linguístico e social para a linguística do corpus (BIBER, 1988; STUBBS, 1996)
- A perspectiva de construção dos significados apoiada nos conceitos de colocação (HALLIDAY; HASAN, 1976; SINCLAIR, 1991) e do princípio idiomático (SINCLAIR, 1991)
- Da sua problemática e definição como área do conhecimento (BERBER SARDINHA, 2000; BIBER; CONRAD; REPPEN, 1999; MCENERY; WILSON, 1996)
- Do conceito de Corpus informatizado, incluindo seus critérios para construção e pesquisa (KENNEDY, 1998; MCENERY; HARDIE, 2012)
- Das ferramentas assim como de suas limitações e funções (BIBER; CONRAD; REPPEN, 1999; LEECH, 1991; MCENERY; HARDIE, 2012; MCENERY; WILSON, 1996)

Passado esse momento inicial, partimos para uma discussão de caráter terminológico cujo objetivo principal era propiciar uma discussão sobre alguns dos termos principais comumente utilizados na área. O conhecimento desses termos é importante para que os participantes pudessem navegar de forma mais confortável pelas diferentes mesas redondas, oficinas e seções de comunicação do evento. De forma geral, a discussão versou sobre os seguintes termos, aqui com suas definições pedagogicamente adaptadas ao contexto desta oficina:

- **Corpus**: Texto ou grupo de textos criteriosamente selecionados para um objetivo de pesquisa, legível por computador e em quantidade relativamente grande.
- **Ferramenta**: Software utilizado para análise de dados linguísticos. Existem as comerciais e as de acesso livre.
- **Lista de Palavras (wordlist):** Lista de ocorrência (absoluta/relativa) de palavras no corpus.
- **Concordância (Concordance ou KWIC):** Lista de ocorrências de uma palavra e o material linguístico co-ocorrente (direita e esquerda).
- **Concordanciador (concordancer):** Ferramenta para o levantamento de concordâncias.
- **Nó/nódulo (node):** Palavra central (posicionalmente), pesquisada na concordância e foco da pesquisa.
- **Palavras-chave (keywords):** Lista de palavras estatisticamente relevantes. Compara-se o corpus de estudo a um maior de referência.
- **Colocados (collocates):** Palavras que co-ocorrem à direita ou à esquerda de uma palavra nó (node)
- **Representatividade**: O quão representativo (estatisticamente) um corpus é do universo que ele representa (**sample**).
- **N-gramas (n-grams)**: Expressões co-ocorrentes com uma palavra nódulo.
- **Parsers**: programas de "árvores" em GGT (Gramática Gerativa Transformacional).
- **Prosódia semântica:** Tom geral em que uma palavra se apresenta no universo pesquisado (positivo, negativo ou neutro).
- **Lema (lemma):** É o lexema ou entrada básica de uma palavra.
- **Types (formas):** Número de itens lexicais únicos.
- **Tokens (ocorrências):** É o número efetivo de ocorrências lexicais em um texto.
- Itens lexicais (ou palavras cheias): Palavras do léxico criativo.
- Itens gramaticais (function words/palavras vazias): conjunções, preposições, artigos etc.
- **Etiquetas (tags):** marcações no corpus (e.g. part-of-speech).

Todos os itens definidos acima tiveram foram exemplificados com dados linguísticos ou captura de telas de ferramentas. O objetivo de tal ilustração era criar uma ideia de como esses conceitos contribuem para a análise de um corpus. Durante esse momento uma discussão bastante profícua sobre o significado da palavra "parser" se colocou. Os participantes oriundos da área de ciências da computação colocaram que, na referida área, o termo "parser" representa qualquer programa de computador capaz de processar dados de linguísticos e não apenas construir representações em GGT.

A discussão migrou, então para a constituição de um corpus. Os critérios expostos e discutidos foram baseados no trabalho de Berber Sardinha (2000):

> **A origem**: Os dados devem ser autênticos.
>
> **O propósito**: O corpus deve ter a finalidade de ser um objeto de estudo linguístico.
>
> **A composição**: O conteúdo do corpus deve ser criteriosamente escolhido.
>
> **A formatação**: Os dados do corpus devem ser legíveis por computador.
>
> **A representatividade**: O corpus deve ser representativo de uma língua ou variedade.
>
> **A extensão:** O corpus deve ser vasto para ser representativo.
>
> (BERBER SARDINHA, 2000, p. 338, nossa ênfase)

No caso da origem, discutiu-se o conceito de autenticidade. Apesar de bastante polêmico, em especial na área de ensino de línguas, o objetivo da discussão era colocar a necessidade de os dados serem colhidos de interações que necessariamente ocorrem no mundo real. Não seria possível, assim, construir um corpus dentro desta abordagem com dados produzidos especialmente para a pesquisa. O resultado de uma análise de Corpus deve ser autêntico na medida que reflete padrões de uso da linguagem realizados por falantes reais. Isso leva a uma clara oposição ao ideário da revolução chomskiana e de algumas abordagens teóricas na área de linguística (BERBER SARDINHA, 2000; LEECH, 1991; SINCLAIR, 1991; STUBBS, 1996).

O propósito está relacionado ao objetivo da pesquisa a ser realizada. Em termos gerais, seria tal objetivo o responsável por ditar a natureza dos dados linguísticos coletados, tanto no que tange aos registros (BIBER, 1988), passando pela forma de tratamento que esses dados terão (MCENERY; WILSON, 1996), como também pelo seu tamanho e pelo prazo de coleta (BERBER SARDINHA, 2000). Tais questões impactam diretamente na noção de composição, uma vez que a proporcionalidade dos dados coletados acaba sendo uma consequência clara de seu propósito.

Já o conceito de formatação atenta para uma realidade importante: não é pensável realizar uma pesquisa dentro do universo da linguística do corpus sem pensar que os dados devem ser legíveis por computador e, por conseguinte, pela ferramenta utilizada.

A representatividade talvez seja a caraterística mais discutida. Em termos gerais, representatividade é um conceito estatístico (OAKES, 1998; BERBER SARDINHA, 2000): dados são representativos quando trazem uma amostra capaz de expressar as principais características de uma população. No caso da Linguística do Corpus, esse conceito se aplicaria a variedade, registro e ou gênero estudados. De forma geral, ao aceitarmos esse princípio, estamos pressupondo que a linguagem é um sistema probabilístico de escolhas (HALLIDAY, 2005), no qual determinados traços podem variar de acordo com sua frequência. A questão que se coloca aqui é que um corpus é uma amostra de um conjunto de dados dos quais não temos conhecimento da sua totalidade. Isso faz com que seja necessário que o planejemos para ser o maior possível dentro do universo de pesquisa que estivermos (SINCLAIR, 1991).

Seguiu-se então para uma discussão das possíveis aplicações da Linguística do Corpus. Dado o exíguo tempo da oficina, tal momento se resumiu a:

- Produção de material didático
    - Aqui foram discutidas duas possibilidades de aplicação. A primeira estava em utilizar as ferramentas de corpus na sala de aula, como seria o caso de concordâncias, listas de palavras e n-gramas. A segunda estaria relacionada a utilização da Linguística do Corpus como insumo para produção de materiais didáticos diversos.
- Lexicografia
    - Discutiram-se questões relacionadas a produção de glossários e dicionários, além de questões mais amplas do estudo do léxico.
- Tradução
    - Discutiu-se em especial a relação da terminologia e dos córpora paralelos, com ênfase em suas contribuições para o ato tradutório.
- Descrição linguística e análise do discurso:
    - Áreas que se beneficiam profundamente das possibilidades teóricas e metodológicas da Linguística do Corpus, uma vez ela permite o levantamento preciso de elementos de linguagem responsáveis por trazer um novo olhar a tais disciplinas.

- Córpora de aprendizes
    - Dados coletados da produção real de aprendizes com o objetivo de avaliar suas características e determinar caminhos para a prática pedagógica.

3. Algumas ferramentas importantes: explicando possibilidades

Após a discussão relatada na seção anterior, a oficina se direcionou para demonstração de algumas ferramentas de Linguística do Corpus. A seguir:

- The Sketchengine (https://the.sketSite permitechengine.co.uk)
    - Sítio que disponibiliza uma infinidade de ferramentas e recursos de córpora. São oferecidas coleções em diversas línguas além da possibilidade de se criar um córpora mantido pelo próprio pesquisador.
- Voyant (http://voyant-tools.org)
    - Sítio que permite a criação de nuvens de palavras, concordâncias e dados estatísticos tomando URLs como base
- AntCorc (http://www.laurenceanthony.net)
    - Programa gratuito, realiza concordâncias e listas de palavras, além de fornecer diversos dados sobre o texto
- WordSmith Tools (http://www.lexically.net/wordsmith/version6/index.html)
    - Suite comercial pioneira na área, possui muitos recursos e funciona em plataforma Windows
- KH Coder (http://khc.sourceforge.net/en/)
    - Programa que não apenas realiza concordâncias e listas de palavras, mas também usa tecnologia de data-mining aplicando-a a textos em diversas línguas. Gratuito.
- Texture (http://textexture.com/)
    - Ferramenta que aplica o conceito de grafos aos colocados em um texto. Funciona apenas em inglês.
- Webcorp (http://www.webcorp.org.uk/live/)
    - Site que realiza concordâncias utilizando a Internet como corpus
- Tree Tagger (http://gramatica.usc.es/~gamallo/php/tagger/TaggerPT.php)
    - Etiquetador para o português
- LXT Tagger (http://lxcenter.di.fc.ul.pt/tools/pt/LXTaggerPT.html)
    - Etiquetador da Universidade de Lisboa
- Claws4 (http://ucrel.lancs.ac.uk/claws/)
    - Etiquetador mantido pela universidade de Lancaster, UK. Muito utilizado na área, precisa de licença comercial.

4. Exercícios guiados: sedimentando os conceitos com a prática

Por se tratar de uma oficina, achamos importante que os participantes realizassem exercícios de análise e explorassem algumas ferramentas discutidas no item anterior. O objetivo dessa fase da oficina foi permitir que uma rápida vivência prática levasse ao levantamento de algumas características específicas de alguns córpora pré-selecionados.

Assim, a lista de exercícios a seguir foi distribuída para os alunos:

1) Acesse o sítio: https://the.sketchengine.co.uk/open/
    a. Clique em “open corpora”
    b. Escolha “ACL Anthology Reference Corpus (ARC)”

        i. A partir dos dados oferecidos pelo sistema, como você descreveria o corpus?
    c. Clique em wordlist
        i. Qual sua impressão sobre as listas de palavras oferecida?
        ii. Quais são os 3-gramas mais frequentes? Qual sua impressão?
    d. Clique em "Sketch diff" e escolha dois lemas em inglês que você gostaria de saber a diferença de uso. Quais suas impressões?
    e. Agora vamos realizar uma concordância.
        i. Clique em "concordance"
        ii. Clique em escolha a palavra "happen" e realize a concordância
        iii. Clique em "sort" e realize uma reorganização multinível, 1L e 1R
        iv. Agora extraia a frequência "node tags"
        v. Agora extraia a frequência "node forms"
        vi. Agora vamos extrair os colocados clicando em colocates.
            1. Após a lista aparecer, clique na letra P, ao lado do colocado
            2. Agora selecione uma palavra no mesmo sítio e repita o exercício de concordância com uma palavra de sua escolha
2) Acesse o sítio http://www.webcorp.org.uk/live/
    a. Escolha "Portuguese" na lista de idiomas
    b. Em "Add popular sites" clique em "Brazil"
    c. Inicie a concordância da palavra "saudade"
    d. Agora volte ao sítio inicial do webcorp
    e. Escola a ferramenta "wordlist tool"
    f. Cole o endereço: "https://pt.wikipedia.org/wiki/Futebol"
    g. Clique sobre a palavra "futebol"
    h. Clique sobre a palavra "visitado"
3) Agora repita o exercício com uma palavra e um sítio, ambos da sua escolha
4) Visite o sítio: http://gramatica.usc.es/~gamallo/php/tagger/TaggerPT.php
    a. Cole o texto: "O futebol feminino tem apresentado um crescimento lento atualmente, principalmente devido a obstáculos sociais e culturais que não permitem o ingresso pleno da mulher ao desporto. "
    b. Clique em "envia ao tagger"
5) Acesse o sítio: "http://lxcenter.di.fc.ul.pt/tools/pt/LXTaggerPT.html"
    a. Clique em "Serviço online"
    b. Clique em "aqui"
    c. Cole o mesmo texto acima
    d. Clique em "Annotate"
6) Agora compare, qual a diferença entre as ferramentas?
7) Realize uma pesquisa em ambas as ferramentas com textos de sua escolha.

5. Avaliação da experiência

Apesar de não ter havido a aplicação de um instrumento de pesquisa formal para obtenção dos dados de avaliação pelos participantes, discussões informais no final da oficina e no decorrer do evento apontaram que:

- Pontos positivos:
    - Como a oficina foi realizada em nível iniciante, isso ajudou deveras na compreensão dos seminários que se seguiram no encontro. Fato determinante para inclusão de pesquisadores iniciantes nas discussões.

    - Como a oficina contou com discussões teóricas e com uma lista de obras relevantes, o trabalho possibilitou que os participantes a tomassem como ponto de partida para aprofundamento.
    - A participação de pessoas de diversas área de estudo enriqueceu discussões sobre a aplicação das ferramentas e dos conceitos discutidos
- Pontos negativos:
    - Tempo: boa parte dos participantes concordaram que uma oficina de caráter introdutório deveria ser de oito horas. Isso facilitaria bastante a compreensão e desenvolvimento das tarefas propostas.
    - Momento de realização: os participantes foram unânimes na opinião de que uma oficina com este caráter deveria ser realizada logo no início do encontro. Isso a tornaria ainda mais útil.
    - Infraestrutura: a queda constante do sinal de rede prejudicou a realização dos exercícios propostos.

Tomando agora minha perspectiva de professor responsável pela oficina, gostaria de ressaltar que esta experiência deveria ser repetida em outros eventos similares, sendo as questões apontadas pelos participantes totalmente pertinentes para sua melhoria. A ampliação do tempo de duração da oficina para oito horas, junto com um melhor posicionamento dentro da agenda do evento, talvez uma oficina pré-congresso fosse ainda mais útil, contribuiriam bastante para seu melhor aproveitamento. Além disso, tais oficinas contribuem deveras para a democratização do conhecimento e ampliação do número de pesquisadores interessados na área, graças a oportunidade de contato inicial. Em outras palavras, tais experiências são relevantes por possibilitar que eventos como o EBRALC-ELC não se tornem herméticos e, por conseguinte, acessíveis somente a pesquisadores avançados.

Gostaria de encerrar esse breve relato agradecendo à organização da VIII Escola Brasileira de Linguística Computacional e do XIII Encontro de Linguística de Corpus (EBRALC-ELC), pelo convite e oportunidade de contribuir para a expansão dos estudos baseados em corpus no Brasil.

Referências

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 2000. v. 16, n. 2, p. 323–367.

BIBER, D. *Variation across speech and writing*. London: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics: investigating language structure and use*. Cambridge: Cambridge University Press, 1999.

HALLIDAY, M. A. K. *Computational and quantitative studies*. London ; New York: Continuum, 2005.

HALLIDAY, M. A. K.; HASAN, R. *Cohesion in English*. London: Longman, 1976.

KENNEDY, G. *An introduction to corpus linguistics*. London: Longman, 1998.

LEECH, D. The state of the art in corpus linguistics. *In*: AIJMER, K.; ALTENBERG, B. (Org.). *English corpus linguistics: studies in Honour of J. Svartvik*. London: Longman, 1991.

MCENERY, T.; HARDIE, A. *Corpus linguistics: method, theory and practice*. Cambridge ; New York: Cambridge University Press, 2012.

MCENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

OAKES, M. P. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press, 1998.

SINCLAIR, J. *Corpus, Concordance, Colocation*. Oxford: Oxford Universiy Press, 1991.

STUBBS, M. *Text and corpus analysis*. [S.l.]: Blackwell, 1996.