A LINGUAGEM DA CIÊNCIA:

PROSPECÇÃO DE DADOS BASEADOS EM CORPORA

Heliana Mello (UFMG, FGV1[1])

hmello@letras.ufmg.br

Renato Souza (FGV)

rsouzaufmg@gmail.com

Resumo: Neste artigo, discutimos brevemente a trajetória da linguística de corpus e suas aplicações aos estudos da linguagem científica. Oferecemos algumas exemplificações de ferramentas de exploração de corpora, aplicações de corpora ao estudo da linguagem científica e empregos ligado-os à garimpagem de dados e processamento da linguagem natural.

Palavras-chave: linguística de corpus, ferramentas computacionais, garimpagem de dados, processamento da linguagem natural.

Abstract: In this paper we briefly discuss the history of corpus linguistics and its applications to the study of scientific language. We provide some exemplification for corpora exploitation tools, corpora applications to scientific language studies, data mining and natural language processing.

Keywords: corpus linguistics, computational tools, data mining, natural language

1Em estágio pós-doutoral junto à EMAp-FGV, 2012-2013.

processing.

Introdução

A ciência, qualquer que seja a sua conceitualização, não existe fora da linguagem. As práticas científicas são variadas e dependem da linguagem em sua configuração. Pensamentos, teorias e dados são organizados e expressos através de sistemas semióticos distintos, de linguagens formais, porém as linguagens escrita e oral estão sempre presentes em sua articulação e expressão.

Ciência e linguagem formam um par convencionalizado – quase um bigrama, nos termos da linguística de corpus. Neste artigo, evidenciaremos a utilização de corpora para a prospecção da linguagem científica. Tal ação pode ter vários objetivos: identificação de palavras-chave de uma dada área, extração de dados para análise e organização de conceitos e entidades de uma área específica (garimpagem de dados), criação de perfis de pesquisa, comparação de práticas profissionais entre distintas áreas do saber, auxílio na escrita de textos científicos, dentre outros.

Primeiramente, discutiremos a definição de corpora, sua caracterização e alguns exemplos. Passaremos, a seguir, à apresentação de algumas ferramentas úteis na sua compilação e tratamento. Exemplificaremos, então, estudos que utilizam corpora como

objeto para discussão da linguagem científica. Na última parte deste artigo, apresentaremos alguns exemplos do trabalho de garimpagem de dados que vem sendo desenvolvido na Fundação Getúlio Vargas por um dos autores e seu grupo de trabalho.

A Linguística de Corpus

A Linguística de Corpus se ocupa da coleta e análise de corpora eletrônicos, um conjunto de dados linguísticos, coletados através de arquiteturas específicas, que funciona como objeto de pesquisa . A linguística de corpus surgiu da necessidade que estudiosos da língua sentiram de se apoiar em usos reais para fazerem generalizações ou esboçarem teorias a respeito do funcionamento linguístico. Atualmente, a Linguística de Corpus está intimamente ligada ao uso do computador, visto que os corpora são conjuntos de dados eletrônicos eletrônicos e manipuláveis através de softwares.

Muito antes do computador, já se fazia uso de compilações de dados que podem ser vistas como precursoras de corpora. Na Grécia Antiga, foi criado o Corpus Helenístico (SARDINHA, 2000). Na Antiguidade e Idade Média, produziam-se coleções de citações da bíblia. Durante boa parte do século XX, foi feito o uso de dados compilados para a descrição de várias línguas e seus dialetos. Os grupos de dados dessas épocas eram coletados, armazenados e analisados manualmente. A dificuldade de se realizarem

estudos desse tipo era enorme. Mesmo assim, havia grande interesse na coleta e exploração de dados sistemáticos. É importante ressaltar o papel dos estudos baseados em dados realizados manualmente, devido à complexidade e ao pioneirismo na época.

O período crítico para os estudos baseados em corpus se deu com a "mudança" de paradigma na linguística, com as ideias de Chomsky, por volta de 1950. Houve uma preferência muito forte por estudos baseados em teorias racionalistas da linguagem que utilizavam a metodologia introspectiva para seus propósitos. Os estudos empíricos receberam muitas críticas nessa época. relacionadas à necessidade de se coletarem dados empíricos e ao modo pelo qual se realizava a coleta e a análise dos dados. Um dos argumentos era a falta de confiabilidade das análises manuais de grandes quantidades de dados linguísticos e o universo parcial e tendencioso que qualquer conjunto de dados linguísticos representa. Embora o cenário fosse desfavorável, os estudos baseados em corpora não pararam. Muitos pesquisadores continuaram seus estudos por meio de corpora. Firth (1957) e os neo- firthianos defendiam a descrição da linguagem por meio de dados reais. O corpus SEU (Survey of English Usage), por exemplo, foi compilado e etiquetado manualmente em 1959. O SEU influenciou a criação de corpora eletrônicos e serviu para o desenvolvimento de etiquetadores computadorizados contemporâneos.

Nos anos de 1960, com o advento do computador e com a queda de prestígio das pesquisas puramente racionalistas, o cenário começou a mudar. Em 1964, o lançamento do Corpus Brown , com 1 milhão de palavras, foi considerado como o gatilho propulsor do

desenvolvimento da Linguística de Corpus. O Corpus Brown é o pioneiro dos corpora eletrônicos por ter nascido em um período ainda desfavorável para os estudos empiristas e, também, pela dificuldade de compilação em computadores mainframe.

Nos anos de 1980, com o aparecimento dos computadores pessoais, ocorreu a popularização dos estudos com corpora . Pesquisadores individuais puderam compilar seus corpora, o que antes somente poderia ser realizado por equipes, com grande custo financeiro. Com o desenvolvimento dos computadores, especificamente com o aumento da capacidade de armazenar e processar dados, maiores números de corpora e ferramentas foram disponibilizadas para pesquisas, contribuindo para a consolidação da Linguística de Corpus.

Um corpus deve ser constituído de dados autênticos, legíveis por computador e representativos de uma língua ou da variedade da língua a qual se deseja estudar. Como já dito, o computador desempenha um papel importante para os estudos na área. As ferramentas computacionais são geralmente utilizadas para reorganização e extração de informações do corpus, para observação e interpretação de dados, fornecendo novas perspectivas para a análise linguística.

Algumas ferramentas computacionais utilizadas na linguística de corpus comumente são:

- Programas para listar palavras (frequenciadores) - fazem a contagem das palavras em um corpus – oferecem listas de frequência de formas. As formas individuais são conhecidas como tipos - types e suas ocorrências, como tokens.

- Concordanciadores – são programas que permitem que o usuário procure por palavras específicas em um corpus, fornecendo listas para as ocorrências da palavra em contexto, e que podem, normalmente ser expandidos para a citação completa em que a forma procurada ocorre.
- Etiquetadores - fazem análises automáticas do corpus e inserem etiquetas (códigos) de ordem morfossintática, sintática, semântica, prosódica ou discursiva.
- Ferramentas de Engenharia Textual- são pacotes de software que buscam modularizar as várias atividades de processamento de linguagem natural, permitindo a montagem de pipelines específicos de tarefas, englobando muitas das anteriormente citadas.

A Linguística de Corpus faz uso de uma abordagem empirista, contrária à abordagem racionalista, do ponto de vista linguístico, e tem como central a noção de linguagem enquanto sistema probabilístico. De acordo com essa noção, os traços linguísticos não ocorrem de forma aleatória, sendo possível evidenciar e quantificar regularidades (padrões). Na área, é comum afirmar que a linguagem é padronizada (patterned), isto é, existe uma correlação entre os traços linguísticos e os contextos situacionais de uso da linguagem. Na Linguística de Corpus, a padronização se evidencia por colocações, coligações ou estruturas que se repetem significativamente. Nesta área, os principais conceitos de padronização são: colocação, coligação e prosódia semântica.

Para muitos pesquisadores, a Linguística de Corpus revolucionou o modo como a linguagem é estudada (MCENERY & WILSON, 1996). Seus achados contribuem para

diversas áreas de pesquisa linguística, dentre elas, comumente, mencionam-se a lexicografia, o de línguas, a tradução, dentre outras.

As principais áreas da Linguística de Corpus são:

- Compilação de corpora;
- Desenvolvimento de ferramentas para análise de corpora;
- Descrição da linguagem;
- Exploração do uso de descrições baseadas em corpora para várias aplicações -como ensino-aprendizagem de línguas e gêneros linguísticos, processamento da linguagem natural por máquinas, reconhecimento de voz, construção de gramáticas e dicionários, etc.

Tipologia de Corpora

Há diversos tipos de corpora que servem também a propósitos distintos (SINCLAIR, 1991). Aliás, uma das máximas da Linguística de Corpus é que um corpus vale tanto quanto a sua adequação ao propósito a que se destina. Assim, quanto à diamesia, os corpora podem ser: escritos (que são os mais comuns e mais facilmente compiláveis), orais (trabalhosos e custosos; dependem de alta qualidade acústica, transcrição e alinhamento do sinal sonoro à sua transcrição através de programas específicos) e corpora multimodais (normalmente incluem imagens, som e texto transcrito.

São, ainda, raros e muito custosos. Há muitas questões éticas relacionadas à exposição de imagens ainda sendo debatidas. Dependem de programas específicos para o seu alinhamento).

Há corpora diacrônicos e sincrônicos. Um exemplo de um corpus diacrônico do português é o Corpus do Português, de Mark Davies e Michael Ferreira, que pode ser acessado gratuitamente via link: http://www.corpusdoportugues.org/. O Corpus do Português não é um corpus totalmente balanceado, mas tem uma cobertura de textos do século XIV ao XX, obviamente, cobrindo textos portugueses antigos exclusivamente, e nos períodos possíveis, também textos brasileiros. Há vários corpora sincrônicos do português acessíveis através do portal Linguateca: http://www.linguateca.pt/.

Há corpora monitores que, servindo para documentar uma dada língua com o passar do tempo, são alimentados com frequência e têm um tamanho gigantesco. Para o inglês, há, por exemplo, o Bank of English (http://www.titania.bham.ac.uk) e o Corpus of Contemporary American English – COCA (http://corpus.byu.edu/coca/). Para o português, contamos com alguns corpora eletrônicos disponíveis à comunidade em geral. O Banco de Português (http://www2.lael.pucsp.br/corpora/bp/) tem parte de seu acervo na Web, assim como o Corpus Brasileiro, com 1 bilhão de palavras (http://corpusbrasileiro.pucsp.br/x/). O Lácio Web já se encontra na Web e tende a crescer (http://www.nilc.icmc.usp.br/lacioweb/). O Tycho-Brahe, de português histórico (http://www.tycho.iel.unicamp.br/~tycho/), também está na Web há muitos anos. Fora do Brasil, como mencionado, a Linguateca (http://www.linguateca.pt/) já disponibiliza vários

corpora em português.

Ferramentas para exploração de corpora

Como mencionado anteriormente, por serem objetos de tratamento computacional, os corpora eletrônicos necessitam de ferramentas computacionais para sua compilação e tratamento. Há diferentes parâmetros que podem instruir a compilação e anotação de um corpus, entretanto, busca-se hoje a adoção de diretrizes que facilitem a padronização dos critérios adotados. Essas podem ser encontradas, por exemplo, nos documentos do Text Encoding Initiative – TEI (http://www.tei-c.org/index.xml) que buscam a padronização da representação de textos em formato digital. Outro exemplo de tentativa de padronização de critérios relacionados ao tratamento de corpora é o Expert Advisory Group on Language Engineering Standards – EAGLES (http://www.ilc.cnr.it/EAGLES/browse.html).

Ferramentas podem estar associadas à compilação de corpora ou ao tratamento de corpora. Alguns programas executam ambas as funções. Esse é o caso do software livre TextSTAT (www.niederlandistik.fu-berlin.de/textstat/). Cf. slide 3: http://www.textolivre.pro.br/chatslide/apresentacoes/melloSouza2011/img2.jpg O TextSTAT é um programa leve, que serve para compilar corpora buscando textos na web ou em pastas específicas, e produz listas de formas e sua frequência e concordâncias.

Outro programa livre para compilação de corpora através da web é o Bootcat

(http://bootcat.sslmit.unibo.it/). O Bootcat dispõem de scripts que, a partir de palavras-chave (seeds), busca páginas específicas na web. Sua utilidade é ilimitada para a compilação de corpora especializados (corpora científicos, por exemplo), corpora paralelos para tradução, corpora para fins lexicográficos, etc (cf. http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf).

Uma terceira ferramenta gratuita interessante é o concordanceador AntConc (http://www.antlab.sci.waseda.ac.jp/software.html). O AntConc oferece um conjunto de sete ferramentas que servem para listar as linhas de concordância de uma dada forma, exibir a linha de concordância em contexto, visualizar o arquivo de texto, listar clusters/n-gramas, listar frequências, palavras-chave e colocados.

Temos software para análise de corpus em português, alguns disponíveis livremente (http://www2.lael.pucsp.br/corpora/ e http://beta.visl.sdu.dk/visl/pt/).

Como exemplo de ferramenta de Engenharia Textual, podemos citar o GATE, o UIMA e o VISUALTEXT. Estas ferramentas permitem realizar a marcação morfossintática, a extração de entidades nomeadas baseando-se em glossários ou ontologias, análise de frequência, concordâncias, entre outras.

Um último exemplo de ferramentas são módulos especializados de linguagens de programação para facilitar o tratamento de corpora. Nesta seara, podemos citar o NTLK (desenvolvido em Python) e os vários módulos para NLP desenvolvidos para a linguagem R.

Bases importantes

Existem muitas bases de divulgação e disponibilização de corpora, algumas de livre acesso e outras pagas. Uma importante base internacional, sobretudo pelos corpora orais que disponibiliza, é a Linguistic Data Consortium (http://www.ldc.upenn.edu/). Para o português, está disponível gratuitamente a Linguateca: (http://www.linguateca.pt/).

Aplicações à linguagem científica

A Linguística de Corpus explora diferentes possibilidades analíticas, com fins específicos. O nosso foco neste artigo é a prospecção de dados da linguagem científica. Para tal, é necessário que se compilem corpora científicos especializados. Esses corpora normalmente não são disponibilizados para a comunidade e são propriedade dos grupos que os utilizam em seus estudos. Há algumas exceções, que, normalmente, voltam-se para o estudo da linguagem científica de um modo geral. Um exemplo é o projeto Scientext (http://scientext.msh-alpes.fr./scientext-site/spip.php?article19). A seguir, listaremos algumas das principais aplicações da Linguística de Corpus aos estudos científicos.

1. Estudos lexicográfico-terminológicos

A partir da exploração de corpora específicos, constroem-se dicionários, glossários, ontologias, implementam-se traduções, etc. Um exemplo de grupo de trabalho nessa área é o projeto TERMISUL (http://www6.ufrgs.br/termisul/index.php). O grupo explicita como seu objetivo "Avançar na pesquisa teórica e aplicada da Terminologia é seu objetivo primeiro. Sua opção teórico-metodológica coloca a Terminologia na perspectiva da linguagem especializada, manifestada no texto especializado."

Outro grupo que trabalha nessa perspectiva desenvolve o projeto TEXTQUIM (http://www6.ufrgs.br/textquim/index.php) , enfocando a linguagem da química. Seus objetivos são: "Fazemos estudos das linguagens técnico-científicas , não restritos às terminologias, considerando o todo dos textos, os modos de dizer, as convencionalidades e as combinatórias de palavras, a enunciação específica de cada gênero textual em diferentes áreas de conhecimento e em diferentes línguas. Nosso usuário principal é o estudante de tradução, o interessado em conhecer os usos da língua através das abordagens da Linguística de Corpus. "

Um terceiro grupo de trabalho na área, com foco em várias áreas científicas, como biocombustíveis, nonotecnologia e fisioterapia, etc, é o GETERM, da UFSCAR (http://www.geterm.ufscar.br/). Seus objetivos, como listados no site do projeto, são:"estudar conteúdos pertinentes à Terminologia/Terminografia; desenvolver pesquisas que gerem produtos terminológicos em língua portuguesa, tais como: glossários,

dicionários, enciclopédias e assemelhados, que satisfaçam demandas reais."

Alguns exemplos de produções nessa área podem ser vistos em:

- Terminologia Verde:

  http://www.lume.ufrgs.br/bitstream/handle/10183/565/000507515.pdf?sequence=1

- Terminologia jurídica:

  http://projeto.lexml.gov.br/arqs/MACIEL.pdf

- TEXTQUIM:

  http://www6.ufrgs.br/textquim/arquivos/perspectivas.pdf

- Artigos Científicos Tutorial:

  http://www6.ufrgs.br/textquim/tutorial.php

2. Estilística, palavras-chave, textualização

Estudos relacionados a esses aspectos, ligados à composição do texto científico propriamente dito, são muito explorados através de corpora de textos acadêmico-científicos orais e escritos e são grandemente utilizados para o ensino de escrita e expressão acadêmico-científica. Exemplos de corpora desta natureza são o Scientext (já visto no slide 1) e o famoso Micase (http://quod.lib.umich.edu/m/micase/), desenvolvido por John Swales na Universidade de Michigan. Swales é um dos precursores do estudo

da linguagem da ciência (http://www.elicorpora.info/).

3. Filosofia da ciência: comparação de marcos epistemológicos

A comparação entre o fazer científico de diversas áreas vem sendo estudado por filósofos através de instrumentos computacionais e processamento estatístico de textos. Nesse campo de estudos, comparam-se concepções científicas, que levam a distintas metodologias exploratórias, através de padrões da articulação discursiva, especificada via índices lexicais, construções e colocados, e palavras funcionais. Um exemplo desse tipo de abordagem é a comparação entre textos científicos de áreas científicas experimentais e históricas (http://lingcog.iit.edu/doc/scientometrics2007.pdf , www.abdn.ac.uk/~csc323/lrecAZCoreSCfinal.pdf)

4. Alimentação de ferramentas de sistemas de informação e bancos de dados

O tratamento computacional de dados de corpora científicos tem servido à pesquisa na Ciência da Informação como um manancial para testagem de ferramentas desenvolvidas na área, via mensuração de sua eficácia, acurácia e validade, além de alimentar o desenvolvimento de novas ferramentas voltadas para a determinação de domínios e suas tarefas co-associadas. Um exemplo de trabalho abordando essa visão foi

desenvolvido por John McMullen, da Universidade da Carolina do Norte (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.9537&rep=rep1&type=pdf).

Um outro exemplo da aplicabilidade da prospecção de dados de corpora para o tratamento da linguagem da ciência e criação de recursos para gerenciamento de informação é o trabalho na área médica (http://www.clt.gu.se/swedish-scientific-medical-corpus-terminology-management-and-linguistic-exploration), que vem sendo desenvolvido no Centro de Tecnologia da Linguagem na Universidade de Gotemburgo (http://www.clt.gu.se/research), na Suécia. Trata-se do projeto MEDLEX, que compilou um corpus científico médico (em expansão), com anotações em vários níveis, i.e., morfossintático, informação semântica (entidade, Medical Subject Heading, terminologia), analisadas por um parser (http://spraakbanken.gu.se/swe/publikationer/collection-encoding-and-linguistic-processing-swedish-medical-corpus-medlex-experience).

5. Detecção de tendências emergentes (Emergent Trend Detection)

Essa é uma nova área na garimpagem de dados, para a qual existem métodos específicos. Sua aplicabilidade se dá tanto no campo científico (identificação de temas e áreas de pesquisa emergentes) quanto no comercial (criação de produtos que atinjam novas faixas e interesses de mercado). No campo científico, a detecção de novas tendências é feitas a partir do tratamento de corpora científicos, normalmente, compostos por artigos publicados em revistas de grande reconhecimento. Estudos nessa área são,

usualmente, desenvolvidos por cientistas da computação ou da informação. Um exemplo de artigo que trata o assunto é: www.jaist.ac.jp/~bao/papers/HoangKSS.pdf

Algumas destas aplicações têm sido desenvolvidas em projetos específicos da Escola de Matemática Aplicada, na Fundação Getúlio Vargas, onde os autores desenvolvem projetos em conjunto. Tendo como objetivo desenvolver pesquisas em campos específicos, ligadas à projetos de desenvolvimento, estão sendo coletados corpora para estes propósitos. Alguns destes projetos serão delineados a seguir.

Inicialmente, está-se desenvolvendo uma ferramenta distribuída e escalável para processamento de linguagem natural. A experiência dos autores com os arcabouços modulares de software para NLP disponíveis – alguns deles apresentados anteriormente neste artigo, como o GATE – foi insatisfatória. A baixa performance de suas implementações e a dificuldade de lidar com grandes massas textuais em tempo real motivou o desenvolvimento da ferramenta PyPLN, baseada em Python. Além do desenvolvimento da ferramenta, deseja-se coletar vocabulários específicos para extração inteligente de informações, também chamadas de ontologias leves. Estas ontologias apresentam compilações de conceitos de determinadas áreas de conhecimento e permitem a identificação dos conceitos no processo de mineração. Atualmente, estão sendo engendradas ontologias no campo do Direito (nomes de advogados, de leis, de juízes, etc.); da História Contemporânea (personalidades, eventos, processos, lugares, etc.), da Mídia (nomes de jornalistas, de veículos, de empresas, de assuntos) e algumas ontologias genéricas, como a coleção de “memes”.

Alguns projetos que estão sendo desenvolvidos atualmente estão usando as partes implementadas deste pipeline, como se discorrerá a seguir.

- No CPDOC – Centro de Pesquisa e Documentação da FGV – os arquivos de documentos digitalizados estão sendo processados no sentido de se fornecerem verbetes para inclusão no DHBB, o Dicionário Histórico Biográfico Brasileiro. Desta forma, o processamento de linguagem natural auxilia na caracterização e representação do domínio.
- Outro projeto que tem usado NLP é o “Supremo em Números”, desenvolvido conjuntamente com a Escola de Direito da FGV. A base de mais de um milhão de processos do STF tem sido objeto de mineração textual, com resultados muito interessantes. No projeto Supremo em Números, houve a necessidade de criar pequenas taxonomias de assuntos, para ajudar na identificação destes temas nos processos, através de técnicas de PLN. Uma visualização gerada através da análise de textos compreendia a relação entre despachos de juízes e leis citadas. Desta forma, pudemos mapear o uso – não simétrico – do corpo de doutrina jurídica e os juízes individuais.
- Em outra parceria com a Escola de Direito da FGV, fazemos a extração de conceitos que acontecem comumente em despachos de juízes, para saber quais os principais temas dos litígios contra a Light – concessionária de energia do Estado do Rio de Janeiro - de modo a auxiliá-la a melhorar seus processos e serviços.

- Em um projeto com a Fiocruz, estamos, a partir de um corpus de textos científicos que versam sobre a Dengue, buscando extrair informações sobre quais os modelos matemáticos que são usados para tratar determinadas particularidades da doença, segundo a literatura. Esta associação permitirá facilitar a pesquisa de novos fenômenos, a partir do inventário de recursos matemáticos utilizados.

- Em um projeto mais recente, batizado de “Media Cloud BR” (inspirado no projeto Media Cloud), tem-se como objetivo o monitoramento da evolução dos conceitos e do gradiente terminológico da mídia brasileira. Desta forma, pode-se associar esta terminologia aos momentos e conjunturas políticos e sociais e aos moldes do projeto Culturomics, desenvolvido com o corpus do Google Books.

Estes projetos envolvem quatro das cinco principais aplicações contemporâneas de linguística de corpus, a saber: a) Estudos lexicográfico-terminológicos; b) Estilística, palavras-chave, textualização; c) Alimentação de ferramentas de sistemas de informação e bancos de dados; e d) Detecção de tendências emergentes. Acreditamos estar contribuindo para o avanço tanto epistemológico, quanto empírico da área, oferecendo à comunidade, através de projetos de cunho open source, recursos terminológicos e computacionais para o processamento de linguagem natural na língua portuguesa.

Referências

SARDINHA, Tony Berber. *Lingüística de Corpus: histórico e problemática.* DELTA, São Paulo, v. 16, n. 2, 2000 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso>. access on 27 Aug. 2012. http://dx.doi.org/10.1590/S0102-44502000000200005.

FIRTH, J. R. *Papers in Linguistics* – 1934-1951. Oxford: Oxford University Press. 1957.

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

SINCLAIR, J. Corpus, Concordance, Collocation. Oxford: Oxford University Press. 1991.