CORPORA, BANCOS DE DADOS E A WEB: ESTADO DA ARTE DA PESQUISA EM FRASEOLOGIA E LEXICOGRAFIA ASSISTIDA POR COMPUTADOR[1]

Noah Bubenhofer[2]
Stefaniya Ptashnyk[3]
Tradução: Cristiane Krause Kilian[4]
Revisão: Alessandra Gusatto[5]

Barack Obama ganhou as eleições presidenciais nos Estados Unidos em 2008 porque utilizou, em seus discursos para o eleitorado, as palavras-chave certas com as colocações adequadas. Referindo-se ao *Iraq,* ele mencionou com frequência lexemas e combinações como *critical issue, national issue, focus on al Qaeda, our troops, family, war, talk, Afghanistan, Pakistan* e muitas outras (ver Figura 1). Seu adversário, John McCain, ao contrário, empregou colocados como *we ... win, suceed, come home, they* e *I* (ver Figura 2). Isso mostra que McCain usou outras colocações com Iraque e que também falou muito sobre o tema, a diferença entre Obama e McCain, no entanto, está no uso, na maneira de falar (BUBENHOFER, 2009). O que chama a atenção não é só o tipo de colocado usado com Iraque, mas também o fato de que, na retórica de McCain, se observa uma variedade menor de colocados do que na de Obama. Concluindo, Obama parece ter preferido uma forma de expressão mais diferenciada (consequentemente também mais complicada) (ver BUBENHOFER et al, 2008a, b).

A afirmação de que as colocações corretas levam à presidência dos Estados Unidos pode ser um tanto exagerada, mas análises sobre padrões de uso linguístico em campanhas eleitorais mostram a relevância das colocações para caracterizar esse uso[6]. Para determinar o uso linguístico *típico* de determinados discursos, temas, pessoas, gêneros textuais, épocas etc.,

[1] Título original: „Korpora, Datenbanken und das Web: State of the Art computergestützter Forschung in der Phraseologie und Lexikographie“. Publicado em PATASHNYK, Stefaniya, HALLSTEINSDÓTTIR, Erla, BUBENHOFER, Noah (eds.) *Korpora, Web und Datenbanken: computergestützte Methoden in der mordernen Phraseologie und Lexikographie*. Baltmannsweiler: Schneider, 2010. p. 7-19. Traduzido com a permissão dos autores e da editora.

[2] Institut für Deutsche Sprache, R 5, 6-13, 68161 Mannheim, Deutschland (bubenhofer@ids-mannheim.de).
[3] Heidelberger Akademie der Wissenschaften, Deutsches Rechtswörterbuch, Karlstr. 4, 69117. Heidelberg, Deutschland (stefaniya.ptashnyk@adw.uni-heidelberg.de).
[4] Pós-doutoranda, bolsista CNPq-PDJ, Instituto de Letras – UFRGS.
[5] Tradutora e Professora de Alemão, Instituto Goethe – Porto Alegre.

[6] O que mostram também análises das eleições para o Parlamento Alemão em 2009 (BUBENHOFER et al., 2009).

interessam, em primeiro lugar, as colocações que, em comparação com *corpora* de referência, são estatisticamente significantes para cada área.

Não apenas combinação de palavras em sentido amplo, mas também fenômenos mais estritamente definidos como expressões idiomáticas e provérbios são centrais para definir os elementos-chave do uso linguístico. Na campanha eleitoral americana, por exemplo, a expressão *Joe the Plumber* ficou famosa, mas não significa literalmente *João, o encanador*; refere-se, entretanto, ao pequeno empresário exemplar pertencente à classe média baixa e que representa o programa econômico de McCain. Existe sim um Joe, que é encanador e é a inspiração para a expressão; durante a campanha eleitoral, no entanto, a locução se consolidou e passou, como uma expressão fixa, a ser um elemento frequente na retórica da campanha eleitoral dos Estados Unidos.

Noah Bubenhofer/Stefaniya Ptashnyk

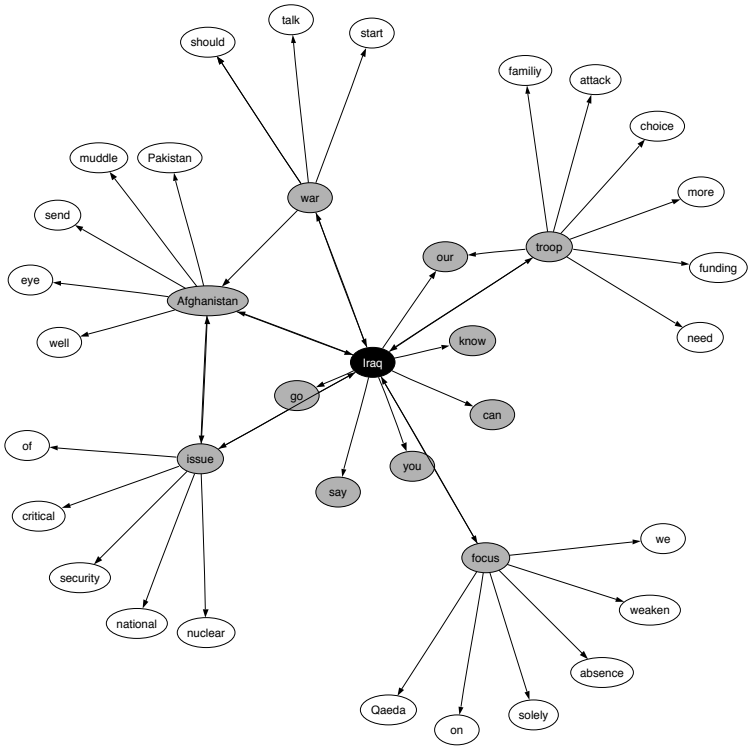

Abbildung 1: Kollokatoren zu *Iraq*, die Barack Obama typischerweise während der ersten TV-Debatte verwendete (vgl. Bubenhofer u. a. 2008a; Grafik: Forschergruppe semtracks).

Figura 1: Colocados de *Iraq* empregados usualmente por Barack Obama durante o primeiro debate televisivo (ver BUBENHOFER et al, 2008a; Gráfico: Grupo de Pesquisa semtracks).

Korpora, Datenbanken und das Web:

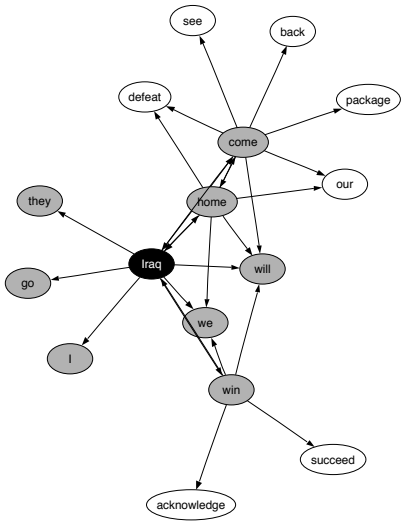

Abbildung 2: Kollokatoren zu *Iraq*, die John McCain typischerweise während der ersten TV-

Figura 2: Colocados de *Iraq* empregados usualmente por John MacCain durante o primeiro debate televisivo (ver BUBENHOFER et al, 2008a; Gráfico: Grupo de Pesquisa semtracks).

O breve exemplo analisado aponta diferentes aspectos do trabalho linguístico assistido por computador na área da Fraseologia e da Lexicografia:

- O trabalho com grandes quantidades de textos ficou mais fácil no que diz respeito aos métodos utilizáveis. A Linguística de *Corpus* coloca à disposição várias ferramentas básicas para calcular padrões linguísticos e colocacionais ou unidades multivocabulares típicas, bem como para anotar e buscar dados.

- O trabalho com grandes quantidades de textos ficou mais fácil também em relação às fontes de dados. A Web é uma mina de textos de diferentes gêneros disponíveis eletronicamente, sejam transcrições de discursos políticos, fóruns de discussões, blogs, jornais, revistas etc. Ainda há, em parte, problemas quanto aos direitos autorais que precisam ser solucionados, mas, pelo menos tecnicamente, a obtenção automática de dados textuais é relativamente fácil.

- Além da variada utilização de dados da Linguística de *Corpus* para estudos fraseológicos com foco no sistema linguístico, a perspectiva fraseológica ajuda também nos âmbitos da análise textual, discursiva, cultural e retórica. Os métodos linguísticos mencionados levam em consideração padrões de uso linguístico que são típicos para certa

área em comparação ao uso linguístico em geral. No entanto, em comum para todas as pesquisas é a convicção de que expressões multivocabulares (sejam definidas em sentido mais restrito ou mais amplo) são categorias de análise relevantes.

- Os resultados das análises podem ser reutilizados de diferentes maneiras, por exemplo, para fins meramente científicos (didáticos, lexicográficos, ou para descrição linguística). Também para um público interessado em política podem ser úteis, como aconteceu no caso das análises referentes às eleições americanas: os dados foram preparados de maneira a demonstrar as especificidades linguísticas da campanha eleitoral nos Estados Unidos.[7]

Os textos reunidos neste volume apresentam soluções e abordagens em todas as áreas. O ponto em comum está no fato de levarem em conta o desenvolvimento atual no cenário da Fraseologia e Lexicografia, no qual métodos de análise e pesquisa assistidos por computador e baseados em *corpus* vêm ganhando importância.

O emprego de métodos assistidos por computador na Fraseologia deve ser visto como uma evolução lógica da dinâmica apresentada até o momento nesta área da Linguística. Após pesquisas tradicionais, que tinham como objetivo a descrição dos fraseologismos em relação ao sistema linguístico, especialmente a sua semântica e sua tipologização, seguiram – como consequência da virada pragmática – questões baseadas em aspectos da Estilística textual e da Pragmática. O progresso da Linguística de *Corpus* e da Linguística Computacional das últimas duas décadas abriu novos caminhos para a pesquisa em Fraseologia: central em vários projetos de pesquisa atuais é a utilização de recursos eletrônicos em geral e de grandes *corpora* textuais em particular.

Neste contexto, já se constatou há muito tempo que, além dos *corpora* compilados com critérios científicos e bem estruturados, também os recursos da Web podem ser empregados como uma enorme base textual para pesquisas fraseológicas, por exemplo, portais de notícias, sites da Web, blogs etc.

Os artigos mostram claramente quais problemas podem ser contornados com a ajuda de recursos eletrônicos e de que maneira *corpora* textuais e a Web podem ser empregados de modo produtivo para fins científicos em geral e nas questões fraseológicas em especial.

As questões tratadas dizem respeito às vantagens e desvantagens da busca automática/formalizada por expressões fixas e da identificação de fraseologismos em *corpora* textuais de grandes proporções, ao emprego de análises guiadas pelo *corpus* na descrição de

7 As análises foram publicadas não apenas no meio científico, mas também em blog (www.semtracks.com), e foram objeto de várias reportagens em meios de comunicação.

padrões de uso linguístico (por exemplo, em certos gêneros textuais ou discursos), à utilidade de métodos estatísticos para pesquisas qualitativas, bem como aos *softwares* utilizados para análise de *corpora* mais extensos.

No contexto da crescente utilização de *corpora* textuais, o conceito de "colocação" foi novamente valorizado. A análise de colocações baseada em *corpus* como uma ferramenta importante na pesquisa em Fraseologia e na Lexicologia prática ganhou em importância. Assim, as várias definições para colocação são tematizadas (por exemplo, como um *continuum* entre expressões fixas e combinações livres, como construções organizadas hierarquicamente, ou como combinações de palavras definidas pelo critério de frequência), bem como a problemática terminológica da pesquisa sobre colocações. Além disso, a relevância das colocações nas áreas de Lexicografia, Didática de línguas, descrição de gêneros textuais e Análise do discurso é notória.

Por fim, são abordados, sob a perspectiva teórica, aspectos e questões que remetem ao entendimento geral dos fenômenos fraseológicos: pesquisas baseadas em *corpus* levam a um redimensionamento dos conceitos básicos em fraseologia, como o conceito de grau de fixação e variabilidade ou o de norma/forma fraseológica etc.? Quais são os pontos de contato que se pode constatar entre os fenômenos sintáticos, por exemplo, entre os padrões sintagmáticos e fraseologias e em que medida questões sintáticas podem ser abordadas nos estudos fraseológicos?

1 *Corpora* como fonte de dados e objeto de análise

A Linguística de *Corpus* moderna, que torna possíveis pesquisas em *corpora* de grandes proporções em formato digital, passou a apresentar-se como um método indispensável também para a Fraseologia: os *corpora* podem, por um lado, ser usados como enormes depósitos, nos quais são buscadas e analisadas comprovações para determinados fenômenos, bem como sua distribuição. Por outro lado, a análise estatística permite a descoberta, em grandes *corpora*, de usos linguísticos padronizados e sua posterior categorização.

Corpora são utilizados majoritariamente para a identificação de fraseologismos e para a análise da sua distribuição, variação e modos de uso. Os trabalhos aqui reunidos referem-se a esses objetivos, para cujas soluções é empregada uma diversidade de métodos.

1.1 Identificação de unidades multivocabulares

A extração automática de fraseologismos ainda é "um osso duro de roer", ou "uma pedra no sapato" para a Linguística de *Corpus* e Linguística Computacional (FILATKINA/KLEINE/MÜNCH[8])? Os trabalhos apresentados empregam métodos variados para abordar o problema, sendo que a expectativa em relação aos métodos não é uniforme. Por um lado, o interesse da pesquisa determina qual tipo de combinação será buscado: colocações lexicais, expressões idiomáticas, provérbios ou combinações multivocabulares não idiomáticas? Por outro lado, aceita-se o trabalho manual em diferentes graus para se obter como resultado a unidade desejada.

Como pano de fundo aparece a discordância sobre a real utilidade dos conceitos teóricos da Fraseologia para a análise quantitativa. Em princípio, há, por parte de vários pesquisadores, a tendência em alargar o conceito de fraseologia, bem como a necessidade de utilizar novos critérios, não tradicionais, para a caracterização de unidades multivocabulares. COLSON argumenta, por exemplo, a favor de uma definição de colocação operacionalizável estatisticamente, para obter resultados objetivos e passíveis de serem reproduzidos. Ele critica que critérios como idiomaticidade ou composicionabilidade são de natureza semântica e cognitiva e, consequentemente, de difícil objetivação. Por isso, colocações deveriam ser compreendidas como fenômenos que pudessem ser definidos com medidas de significância estatística.

Por outro lado, há objetivos concretos, como a elaboração de dicionários fraseológicos (ALEKSA; TOPOROWSKA GRONOSTAJ/SKÖLDBERG; ĎURČO etc.), ou como pesquisas para divulgação de expressões idiomáticas e provérbios que poderiam ser aplicados para fins didáticos (HRISZTOVA-GOTTHARDT; PETROVA etc.). Nesses casos, os *corpora* são verificados à procura dos fenômenos desejados com base em teorias definidas. Há aqui a expectativa de, com o auxílio de métodos automáticos, encontrar da forma mais exata possível os fenômenos que na teoria fraseológica tradicionalmente se definem como 'colocação', 'frasema', 'expressão idiomática' ou 'provérbio.'

Em uma posição central, encontram-se métodos que combinam medidas estatísticas com conhecimento sobre estruturas sintáticas e aspectos semânticos. Ferramentas como 'DeepDict Lexifier' (FJELD/NYGAARD/BICK) ou 'Sketch Engine' (ĎURČO) baseiam-se nesses procedimentos.

[8] Os nomes escritos em versalete remetem aos artigos do livro.

Nos artigos do livro, são empregadas tanto ferramentas simples quanto métodos mais elaborados que necessitam, em alguns casos, conhecimentos de programação. Mais especificamente, são ferramentas disponíveis para o grande público e métodos especialmente desenvolvidos, mencionados a seguir.

1 Softwares disponíveis

a) Programas que calculam colocações ou unidades multivocabulares de diferentes extensões com medidas de significância:

– Kwic Concordance for Windows: *http://www.chs.nihon-u.ac.jp/eng dpt/ tukamoto/kwic _e.html* (ver ALEKSA)

– NSP Ngram Statistics Package: http://ngram.sourceforge.net/ (ver ALEKSA)

– Collocation Extract: http://pioneer.chula.ac.th/~awirote/colloc/ (ver ALEKSA)

– Manatee/Bonito: http://nlp.fi.muni.cz/projekty/bonito/ (ver ĎURČO)

b) Sketch Engine: *http://www.sketchengine.co.uk/* (ver ĎURČO)

c) DeepDict Lexifier: *http://gramtrans.com/deepdict/* (ver FJELD/NYGAARD/BICK)

2 Métodos próprios

a) Busca de colocações do tipo Adjetivo + Substantivo na Web utilizando as interfaces de programação de aplicativos (APIs) das máquinas de busca da Web. O processo segue o princípio da busca por modificadores como *most, rather, quite, too* em contextos com um substantivo dado; em seguida, ocorre o processamento automático dos resultados, como extração de todos os adjetivos, verificação da frequência etc. (ver COLSON).

b) Verificação sistemática de todos os n-gramas em um texto dado através de interfaces de programação de aplicativos (APIs) das máquinas de busca da Web. Aqui, o grau de fixação, ou seja, a relação da forma exata com a forma variante é testada. Em trigramas, isso é verificado com a seguinte fórmula: frequência [palavra1 + palavra2 + palavra3] em relação a [palavra1 + qualquer palavra + palavra3] (ver COLSON).

c) Análises de frequência de uso de possíveis colocados em relação a uma palavra de busca através de buscas no Google. Esse método funciona segundo o princípio da análise de frequência para diferentes colocações; em seguida, as frequências das diversas colocações são comparadas entre si para se acharem os colocados mais frequentes da palavra de busca (ver KONECNY).

d) Combinação de diferentes critérios para extração automática de fraseologismos em *corpora* (ver QUASTHOFF/SCHMIDT/HALLSTEINSDÓTTIR). Aqui são exploradas várias características operacionalizáveis estatisticamente, como pouca variabilidade, combinação

típica de classes de palavras ou a existência de pelo menos duas "palavras importantes" (geralmente palavras lexicais) etc.

1.2 Testagem das formas de uso e dos limites da variação de fraseologismos

Enquanto a identificação de fraseologismos é extremamente complexa, as análises que partem de determinados fraseologismos (preestabelecidos) ou de seus constituintes são visivelmente mais fáceis. Sem dúvida, *corpora* extensos formam uma boa base para verificar a distribuição e as formas de uso de fraseologismos, bem como suas variantes usuais e modificações ocasionais (ver PTASHNYK, 2009). Vários trabalhos deste livro analisam a ocorrência de fraseologismos em determinados gêneros textuais, determinadas línguas ou áreas temáticas. WALLNER, por exemplo, tem seu foco de interesse nos diferentes usos de colocações em *corpora* científicos e da linguagem geral. Com base em testes de significância, verifica-se se as diferenças de frequência de colocados nos dois *corpora* ocorrem por acaso. HRISZTOVA-GOTTHARDT utiliza um conjunto de provérbios búlgaros como base para analisar a sua divulgação em textos jornalísticos. Também NIEMI et al. partem de expressões idiomáticas preestabelecidas: Os autores analisam as diferenças de uso de expressões idiomáticas relacionadas com o corpo em diferentes línguas.

Enquanto a busca de fraseologismos usuais acontece de forma relativamente simples através de seus constituintes e caracteres curinga, a extração de unidades modificadas é notavelmente mais difícil. Um exemplo para tal empreendimento é o Projeto HyperHamlet (ver QUASSDORF/HÄCKI BUHOFER), no qual uma classe específica de expressões fixas são objeto de pesquisa, ou seja, citações da obra de Shakespeare "Hamlet" e suas modificações.

No contexto de pesquisas baseadas em *corpus* que analisam os usos de fraseologismos, discute-se criticamente como é a correlação entre a frequência de ocorrência dessas unidades nos *corpora* e seu real uso por falantes de uma língua. DRÄGER/JUSKA-BACHER ressaltam que a frequência de uso não está necessariamente ligada ao fato de se conhecer ou não a expressão e sugerem, por isso, que estudos baseados em *corpus* sejam complementados com uma pesquisa *online*. Outros trabalhos, no entanto, indicam que, em caso de não haver correlação, pode-se tratar de um problema de *corpus*, principalmente de uma base de dados pequena demais (ver QUASTHOFF/SCHMIDT/HALLSTEINSDÓTTIR): fraseologismos são em parte um fenômeno da língua falada, fato pelo qual somente em *corpora* suficientemente extensos, que contenham vários gêneros textuais, é de se esperar que fraseologismos recorrentes sejam frequentes nos dados dos *corpora*. O trabalho de PETROVA mostra que o uso de newsgroups

como *corpus* é uma possibilidade interessante para aumentar a diversidade de gêneros textuais e também levar mais em consideração a língua falada.

Sob o prisma da Linguística de *Corpus*, o escopo da variabilidade dos fraseologismos se apresenta como um problema, sejam formas distintas de variantes usuais (p. ex. lexicais, estilísticas ou ortográficas) ou modificações ocasionais. Também a flexão dos fraseologismos, que geralmente é obrigatória na inserção das unidades multivocabulares em sintagmas da frase, dificulta em muitos casos as pesquisas. Do ponto de vista técnico, esse problema pode ser solucionado com relativa facilidade, por exemplo, com etiquetadores morfossintáticos ou lematizadores desenvolvidos pela Linguística Computacional (ver p. ex. o TreeTagger, SCHMID, 1994) e que já foram treinados para várias línguas. Se, além disso, estiver à disposição um *corpus* anotado manualmente e listas de lemas, esses etiquetadores podem ser treinados para outras línguas. Apesar disso, essas ferramentas parecem não pertencerem ao grupo de ferramentas básicas utilizadas por alguns pesquisadores na área da Fraseologia. Esse fato revela a necessidade de haver um *software* em Linguística de *Corpus* e Linguística Computacional de fácil uso, que possa também ser utilizado por pesquisadores sem experiência em programação.

Problemas com a variação ortográfica são relativamente fáceis de contornar: se o escopo é conhecido, esse pode ser levado em conta nas buscas, usando-se, por exemplo, as chamadas "expressões regulares", ou seja, uma linguagem complexa que permite, por meio de caracteres curinga e de formulações de condições, buscas que permitem extrair variantes. Na contribuição de FILATKINA/KLEINE/MÜNCH, são mencionados algoritmos que calculam a semelhança entre fraseologismos e, dessa forma, encontram expressões que apresentam variação mínima de ortografia.

No entanto, a busca por fraseologismos com modificações léxicas é uma tarefa mais complexa, como mostram HRISZTOVA-GOTTHARDT, PARIZOSKA E PETROVA. A solução poderia ser restringir a procura por palavras-chave do fraseologismo e considerar a estrutura sintática, como mostram QUASTHOFF/SCHMIDT/HALLSTEINSDÓTTIR. Também poderiam ser utilizados bancos de dados semânticos e ontologias, como WordNet, GermaNet etc., para integrar automaticamente sinônimos, hiperônimos e hipônimos na busca por fraseologismos modificados. Recursos como a compilação de fraseologismos com componentes arcaicos (ver RICHTER/SAILER/TRAWIŃSKI) podem ser igualmente úteis para diversos fins da análise automática.

Corpora prontos oferecem poucas possibilidades de buscas que permitam extrair variantes, embora essas buscas sejam muito úteis (ver, por exemplo, a pesquisa de PETROVA

baseada no Corpus Kielipankki). Assim, faz-se necessário um trabalho conjunto entre os profissionais da Linguística Computacional e Linguística de *Corpus*, que sabem programar instrumentos de pesquisa adequados e que, além disso, possuem larga experiência na análise automática de textos.

2 Compilação e uso de *corpora* de vários tipos

2.1 Web e corpus, Web como corpus

Já há muito tempo são usados na pesquisa linguística *corpora* compilados segundo critérios definidos que oferecem quantidade de dados relativamente balanceados. Há algum tempo, a Web é vista como um recurso que cada vez mais ganha importância, e, nesse contexto, surge a discussão do papel da Web como *corpus*. Há duas maneiras distintas para sua utilização: uma consiste em usar as máquinas de busca já existentes para procurar, em textos disponíveis eletronicamente, por certos fenômenos; aqui toda a Web é interpretada como um *corpus* gigantesco. A outra maneira consiste na possibilidade de, com base nos dados de acesso livre da Web, organizar *corpora* próprios, orientados rigorosamente pelos interesses de cada pesquisa.

A. Uso de máquinas de busca

Com máquinas de busca existentes, como Google e outras, pode-se buscar na Web por determinados fenômenos linguísticos. Dessa maneira, tem-se acesso, de forma rápida e sem maiores investimentos, a um *corpus* enorme. Caso se quisesse compilar um *corpus* do tamanho daqueles indexados pelos operadores das máquinas de busca e torná-lo útil para pesquisas, seria este um empreendimento de alto custo, tornando-se quase inviável para fins de pesquisa.

As desvantagens do uso de máquinas de busca existentes são, por um lado, o fato de elas não serem projetadas para pesquisas linguísticas e os algoritmos de busca não serem conhecidos em detalhes. Por outro lado, a base de dados é difícil de ser controlada. Os operadores de máquinas de busca não revelam a extensão do *corpus* indexado. Em consequência, não é possível estabelecer frequências em relação ao tamanho do *corpus*. Além disso, a composição do *corpus* muda, por sua natureza, constantemente.

Exemplos para tais processos são apresentados em KONECNY, que descreve a utilização do Google para estudos fraseológicos. COLSON contorna as limitações das

máquinas de busca utilizando suas interfaces de programação de aplicativos (APIs) para buscas automáticas, para, posteriormente, processar os dados com métodos próprios.

B. Compilação de *corpora* baseados na Web

Os textos publicados na Web e de acesso livre podem ser usados como base para a compilação de um *corpus* próprio. Nesses casos, os pesquisadores não dependem das possibilidades de uma máquina de busca, mas podem preparar os dados como quiserem e possuem, assim, controle total sobre a composição do *corpus*. QUASTHOFF/SCHMIDT/HALLSTEINSDÓTTIR mostram detalhadamente o seu método para uso da Web, com o qual compilam *corpora* próprios para o Projeto Lexicográfico da Universidade de Leipzig.

A obtenção automática dos documentos é relativamente fácil, como demonstrado em vários trabalhos (ver BARONI/BERNARDINI, 2006; FLETCHER, 2007; KILGARRIFF/GREFENSTETTE, 2003; SHAROFF, 2006). Há considerações, contudo, quanto aos direitos autorais, sendo que projetos de um *corpus* de código aberto podem apontar para uma solução na qual os dados para análise estão salvos, mas só são colocados à disposição na forma de listas de URL (SHAROFF, 2006).

2.2 Bancos de dados para pesquisa e gerenciamento de dados linguísticos

Com a crescente profissionalização das análises de *corpora*, sistemas cada vez mais complexos são empregados para o gerenciamento dos resultados obtidos. Na Fraseologia e Lexicografia, geralmente, a análise do *corpus* é apenas o primeiro passo para uma determinada pesquisa; em seguida, ocorre a avaliação e a categorização das ocorrências ou de resultados preliminares extraídos automaticamente, como os candidatos a fraseologismos. Assim, não é de surpreender que esses resultados sejam arquivados em bancos de dados e complementados com informações adicionais. Os bancos de dados oferecem a vantagem de que os dados neles reunidos e representados podem ser reutilizados sem problemas para diversos produtos finais. Com base em um banco de dados *online,* usado em toda sua complexidade durante o processo de pesquisa, pode-se produzir, por exemplo, um dicionário impresso, ou um portal dirigido a usuários leigos. DRÄGER/JUSKA-BACHER mostram esses múltiplos usos dos bancos de dados.

Enquanto na fase inicial dos bancos de dados era suficiente usar sistemas instalados localmente no computador de trabalho, a era da Web exige bancos de dados conectados em rede que possam ser utilizados por vários usuários ao mesmo tempo através da Internet. Esses

bancos de dados possibilitam uma nova forma de cooperação entre vários pesquisadores independentemente do seu local de trabalho. Até mesmo a integração de leigos (informados), seguindo o modelo de colaboração da Web, como, por exemplo, a Wikipédia, torna-se possível, o que é detalhado por DRÄGER/JUSKA-BACHER: trata-se de um tipo de 'crowdsourcing', a delegação de tarefas a uma grande massa de voluntários que colaboram, através da Web, em um projeto conjunto. É a soma dessas colaborações que faz com que os projetos se concretizem. A Wikipédia é um exemplo bem-sucedido para *crowdsourcing*. Esse modelo pode ser transferido para a Linguística: DRÄGER/JUSKA-BACHER descrevem, no seu artigo, o dicionário *online* de fraseologismos que ganha em qualidade com os comentários e as colaborações de usuários. Além disso, o comportamento dos usuários pode ser analisado: quais entradas são mais consultadas? Para quais entradas há uma necessidade maior de discussão? Dessa maneira, podem-se obter referências sobre o conhecimento, frequência de uso ou divergências acerca de fraseologismos. Outro exemplo para um banco de dados aberto é o "HyperHamlet" (ver QUASSDORF/HÄCKI BUHOFER).

Um grande desafio é o gerenciamento de dados históricos. FILATKINA/KLEINE/MÜNCH relatam sobre sistemas complexos de banco de dados para a fraseologia histórica: no Projeto "Linguagem Formulaica Histórica e Tradições de Expressão", exemplos para linguagem formulaica da Idade Média e início da Idade Moderna são compilados e categorizados em um banco de dados em rede (MySQL com interfaces correspondentes). Nesse contexto, novas possibilidades da Informática, que parecem apenas detalhes, mostram-se de enorme ajuda: somente com a definição do padrão unicode para codificação de caracteres (UTF-8) foi possível processar sistemas de escrita históricos (e também sistemas não ocidentais) com sistemas que não tenham sido desenvolvidos especialmente para esses fins. FILATKINA/KLEINE/MÜNCH evidenciam que bancos de dados podem apresentar recursos multimídia, como mostra o projeto "Conhecimento gnômico através de imagens", no qual dados textuais são combinados com imagens.

Um papel cada vez mais importante tem assumido a linguagem de marcação XML, com a qual metainformações e anotações podem ser acrescidas a diversos tipos de dados. RICHTER/SAILER/TRAWIŃSKI utilizam a linguagem XML para bancos de dados de arcaismos lexicais e itens de polaridade positiva e negativa. Através de um banco de dados como o 'eXist' (*exist.sourceforge.net*), pode-se gerenciar facilmente qualquer documento em XML. Com ele, as vantagens da linguagem XML são combinadas com as vantagens de um banco de dados: assimm é possível gerenciar os metadados de um texto ou porção textual no sistema de

banco de dados e, paralelamente, anotar o texto em XML que, por sua vez, pode ser analisado automaticamente.

Além do gerenciamento de ocorrências, os bancos de dados oferecem outra vantagem: a quantidade de conjunto de dados pode ser explorada facilmente de acordo com diversos critérios. Apertando somente uma tecla, obtêm-se várias informações estatísticas sobre os dados (sempre que eles estiverem divididos por categorias). Interessantes são os métodos que revelam as semelhanças entre diferentes conjuntos de dados através de algoritmos. Eles são muito úteis, por exemplo, para os bancos de dados da linguagem formulaica histórica, na qual as variantes ortográficas e sintáticas são de difícil lematização, já que, geralmente, não apresentam uma padronização. Como as variantes de uma unidade multivocabular são extraídas automaticamente em textos históricos com muita variação é abordado em FILATKINA/KLEINE/MÜNCH.

3 Resumo e Perspectivas

Os trabalhos reunidos no livro demonstram, por vezes detalhadamente e por outras exemplarmente, o leque de opções de métodos e procedimentos da Linguística de *Corpus* e do uso do computador os quais podem ser utilizados para pesquisas teóricas e práticas na Lexicologia e na Fraseologia, bem como na Lexicografia Aplicada e na Didática de Línguas. Novas e rápidas possibilidades de chegar-se aos dados empíricos consolidam-se como a maior vantagem tanto para pesquisadores quanto para os usuários dos resultados das pesquisas, sejam eles leigos ou especialistas.

No contexto dessa evolução, mostra-se, cada vez mais e com muita clareza, a questão sobre os limites do fraseológico, questão essa que já vem sendo discutida há muito na Fraseologia tradicional, mas que ainda não foi solucionada. Se o computador deve ser empregado para extrair automaticamente fraseologismos, o fenômeno 'fraseologismo' deve ser operacionalizável, de maneira que possa ser encontrado na superfície linguística através de regras claras. Os diversos algoritmos desenvolvidos abrangem, às vezes mais, às vezes menos, o que classicamente se define como fraseologismo. Há boas razões para entender-se o fenômeno de maneira mais ampla e, através de métodos estatísticos, incluir mais opções de usos linguísticos padronizados. Vários estudos já mostraram que esses fenômenos dos usos padronizados são de extrema importância para questões na Lexicografia, Didática e Linguística Textual, que, no entanto, nem sempre se enquadram na Fraseologia tradicional.

Qual resultado poderá trazer o alargamento dos interesses das pesquisas primeiramente focados nas questões de fraseologia em relação aos métodos mais modernos, que vem pondo em questão os limites da Fraseologia? Pesquisas futuras mostrarão se esse desenvolvimento fará com que a Fraseologia tome para si um objeto de estudo mais restrito e claramente definido, ou se na interface entre a Fraseologia tradicional e a Sintaxe afirmar-se-á uma nova subárea da Linguística.

Referências

BARONI, M., BERNARDINI, S. (eds.). *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT, 2006.

BUBENHOFER, N. *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge*. 2006. (*http://www.bubenhofer.com/korpuslinguistik/*).

BUBENHOFER, N. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskursund Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen; 4), 2006 .

BUBENHOFER, N., KLIMKE, M., SCHARLOTH, J. *political tracker – U.S. Presidential Campaign '08: A Semantic Matrix Analysis*. 2008a. (*http://semtracks.com/politicaltracker/*).

BUBENHOFER, N., KLIMKE, M., SCHARLOTH, J. The Word War: "Yes, He Did". How Obama won the (rhetorical) battle for the White House. In: *International Relations and Security Network, ISN ETH Zurich.* 2008b. (*http://www.isn.ethz.ch/Current-Affairs/Special-Reports/The-Word-War-Yes-He-Did/Analysis*).

BUBENHOFER, N., KLIMKE, M., SCHARLOTH, J. *political tracker – Bundestagswahl '09. Eine Semantische Matrixanalyse*. 2009.(*http://semtracks.com/politicaltracker/*).

FLETCHER, W. H. Implementing a BNC-Compare-able Web Corpus. In: FAIRON, C.; NAETS, H., KILGARRIFF, A., DE SCHRIJVER, G.-M. (eds.) *Building and Exploring Web Corpora – Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval (WAC3-2007, September 2007), UCL*, Louvain: Presses Universitaires de Louvain, 2007.

KILGARRIFF, A., GREFENSTETTE, G. Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics* 29, vl. 3, 2003, p. 333–347.

PTASHNYK, S. *Phraseologische Modifikationen und ihre Funktionen im Text. Eine Studie am Beispiel der deutschsprachigen Presse*. Baltmannsweiler: Schneider. 2009.

SCHMID, H. *Probabilistic Part-of-Speech Tagging Using Decision Trees.*1994. (*http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf*).

SHAROFF, S. Open-source Corpora: Using the Net to Fish for Linguistic Data. In: *International Journal of Corpus Linguistics,* v. 11, 2006. p. 435–462.