Kabatek, Johannes (ed.) (2016): *Lingüística de Corpus y Lingüística Histórica Iberorrománica.* Berlin / Boston: Walter de Gruyter, 448 pp.

A linguística de corpus, mais especificamente, de corpora diacrónicos, no contexto iberorromânico é o tema central do volume *Lingüística de Corpus y Lingüística Histórica Iberorrománica*, editado por Johannes Kabatek com a colaboração de Carlota Benito. Este volume nasce de dois encontros complementares: no verão de 2014, teve lugar na Universidade de Zurique o CODILI3 – *Terceiro Colóquio Internacional sobre Corpora Diacrónicos das Línguas Ibero-Românicas* e, em janeiro de 2016, decorreu em Kandersteg (Berna) o Curso de Inverno ALPES - *Abriendo Líneas en el Passado del Español*. Os textos publicados neste volume são representativos dos tópicos discutidos no âmbito destes encontros.

Para além da Introdução, o livro é composto por 18 trabalhos que o editor organizou de acordo com quatro grupos temáticos: I) Contribuições das línguas iberorromânicas para a linguística de corpus; II) Corpora iberorromânicos; III) Corpus e análises quantitativas e IV) Questões linguísticas diacrónicas iberorromânicas e linguística de corpus.

O volume apresenta-se como um manual de referência sobre os corpora diacrónicos em curso no contexto ibérico e hispânico. As várias reflexões sobre a linguística de corpus (grupo I) devem ser tidas em conta na constituição de corpus. É necessário atentar, entre outros aspetos, na arquitetura do corpus para que o utilizador tire o máximo proveito da informação disponível. A informação sobre os diferentes corpora em desenvolvimento (grupo II) e sobre possíveis formas de os explorar (grupos III e IV) reflete alguma investigação linguística baseada em corpora que é possível levar a cabo.

Na Introdução, Kabatek começa por chamar a atenção para a linguística histórica, para os corpora e para uma nova etapa com "más y mejores corpus, de más y mejores herramientas para el tratamiento de los datos y, finalmente, de una serie de nuevos estándares más o menos establecidos en la comunidad" (p. 1). Debaixo da designação de linguística de corpus estão abrangidas várias disciplinas que se entrelaçam e complementam: i) a disciplina que trata da criação de corpus, ii) a relação ente a linguística e a informática, ou seja, entre os dados, o seu tratamento e anotação, e a análise quantitativa e estatística que se faz desses mesmos dados e iii) a linguística que se faz com recurso a corpus. Não obstante, o editor refere a questão da representatividade e o perigo de se confundir corpus com língua: "Los corpus son colecciones de *textos* que nos permiten tener una visión indirecta de la lengua, ya que la producción de textos a partir de la competencia lingüística de los individuos está condicionada por una serie de factores que el corpus no permite ver" (p. 4). Mais adiante, Kabatek realça igualmente a atitude crítica sobre corpora existentes (por exemplo, o "CORDEmáforo") só possível graças ao desenvolvimento que a linguística de corpus teve nas últimas décadas.

A primeira secção abre com o texto de Andrés Enrique-Arias que fala das vantagens dos corpora paralelos sob o parâmetro de *perspectiva*, ou seja, a forma como o utilizador acede ao corpus. O corpus *Biblia Medieval* (BM) disponibiliza versões paralelas de textos em hebreu, latim e espanhol medieval e o autor enuncia as vantagens de pesquisa num corpus desta natureza contrapondo com o *CORDE* (*Corpus Diacrónico del español*) *e o CE (Corpus del Español)*. Não obstante, o autor chama a atenção para o perigo de usar-se o BM como única fonte de informação em estudos diacrónicos, uma vez que é necessário averiguar se a pesquisa encontrada só se aplica à tradução bíblica. Algumas das vantagens apontadas podem tornar-se desvantagens se o utilizador não dominar nenhuma outra língua das versões paralelas, neste caso, o hebreu e o latim.

No segundo texto ("Traducción y tradición en los corpus: nuevas perspectivas para la linguística histórica"), Santiago del Rey Quesada sublinha a importância da história da tradução para a história da língua. Na tradição dos *Corpus-based translation studies* (CTS), o autor faz o estado da arte no que se refere aos diferentes tipos de corpora: paralelos (como o BM), multilingues e comparáveis. Ainda que não haja corpora diacrónicos comparáveis, seria muito útil para os historiadores da língua a existência de corpora paralelos e comparáveis em línguas distintas.

O trabalho de Álvaro S. Octavio de Toledo y Huerta ("Aprovechamento del CORDE para el estúdio sintáctico del primer español moderno") explora as potencialidades de um corpus, o CORDE, para o estudo de vários fenómenos morfossintáticos num período específico do espanhol: de 1675 a 1825. Com o auxílio de análises quantitativas, são apresentados diferentes tipos de difusão e dinâmica variacional dos fenómenos considerados.

O trabalho seguinte ("Tres propuestas en el âmbito de la lingüística de corpus") é da autoria de Joan Torruella e pretende apresentar, a partir do CICA (*Corpus Informatitzat del Català Antic*, um corpus "textual, pequeño, histórico, diacrónico y concerniente a lengua em general", p. 91), uma proposta de reflexão sobre o desenho e construção de corpora. Destes fatores depende uma melhor exploração e consequente análise dos dados.

Com o trabalho intitulado "Iluminar los *Séculos Escuros*: Gondomar, un corpus para el estudio del gallego en la Edad Moderna", Rosario Álvarez e Ernesto González Seoane

inauguram a segunda secção do volume: *Corpus iberorrománicos*. É apresentado o corpus *Gondomar – Corpus dixital de textos galegos da Idade Moderna*, as suas especificidades e os desafios enfrentados na definição e construção de um corpus de um período temporal muito específico (séc. XVI-XVIII).

Em seguida, apresenta-se o texto "O CIPM – Corpus Informatizado do Português Medieval, fonte de um Dicionário exaustivo". Maria Francisca Xavier, uma das responsáveis deste projeto, dá a conhecer as várias etapas que, desde 1993, são constitutivas deste corpus e a sua relação com outros projetos, nomeadamente o *Dicionário de Verbos do Século Treze* (1999) e o futuro *Dicionário da Língua Portuguesa Medieval*. É também anunciada uma colaboração que juntará o TMILG (*Tesouro Medieval Informatizado da Língua Galega*) e o CIPM, originando o CIGPM - *Corpus Informatizado do Galego-Português Medieval*.

Dando continuidade à divulgação do período medieval, Vicente J. Marcet Rodríguez e Mª Nieves Sánchez González de Herrero, no texto "La documentación medieval de Miranda de Ebro: Presentación del corpus y rasgos linguísticos", abordam esta temática referente a uma região geográfica muito precisa: o norte de Burgos. Os documentos deste corpus são dos séculos XIII-XV, levando os autores a concluir que se está perante uma língua em formação, com traços arcaizantes e simultaneamente com fenómenos próximos do castelhano moderno.

Segue-se o texto, assinado por vários membros da equipa, que apresenta o projeto *P.S. Post Scriptum*, um corpus da Idade Moderna (séc. XVI-XIX) composto por cartas particulares em espanhol e em português. Dada a natureza específica da sua arquitetura, as principais potencialidades concentram-se na Sociolinguística Histórica, mas não se limita a essa área. Os autores apresentam vários casos de estudos baseados nos dados do corpus: 1) *pois*, como marcador discursivo; 2) o pronome relativo *cujo* e 3) o leísmo, laísmo e loísmo, em espanhol.

"*Citius, maius, melius*: del CREA al CORPES XXI" é o trabalho de Guillermo Rojo que descreve o caminho percorrido desde a construção do CREA (*Corpus de Referencia del español actual*), passando pelo CORDE, ambos de meados dos anos 90, até ao CORPES XXI (*Corpus del español del siglo XXI*), que se iniciou em 2007. A sua configuração mais moderna permite pesquisas mais avançadas e constitui-se como um corpus de referência do mundo hispânico.

A secção "Corpus y análisis cuantitativos" inicia-se com um texto de Dorien Nieuwenhuijsen, intitulado "Notas sobre la aportación del análisis estadístico a la lingüística de corpus". Usando um caso concreto, a variação do modo indicativo e conjuntivo nas orações interrogativas indiretas negativas introduzidas pelo verbo *saber* com diferentes sintagmas interrogativos (*no sé si/qué puedo/pueda*), pretende-se demonstrar de que forma é que a análise estatística pode conduzir a conhecimentos mais profundos sobre o fenómeno em estudo, usando textos do espanhol peninsular e americano extraídos do *Corpus del Español (CdE)*.

No estudo seguinte ("Entrenchment and frequency effects in the diffusion and replacement of modal perífrases in Spanish: a diachronic variationist analysis"), Kim Schulte e José Luis Blas Arroyo abordam as construções perifrásticas modais em espanhol, usando uma análise diacrónica variacionista. O objetivo é testar o peso estatístico de um leque de variáveis em diferentes estádios. Os autores partem de um corpus que contém documentos com informação pessoal e privada, i.e. correspondência pessoal.

O aparecimento de pronomes átonos em posição pós-verbal é o objeto de estudo do trabalho de Miriam Bouzouita. São analisados os contextos de futuros e condicionais sintéticos (*tornaré los*) e de futuros e condicionais analíticos (*tornar los é*). São apresentadas várias hipóteses para o surgimento destas formas sintéticas: 1) o contexto sintático; 2) a forma morfológica do verbo e 3) o modelo latino (e o efeito de *priming* sintático). O corpus *BM* serve de base para este trabalho.

María Jesus Torrens Álvarez e Hiroto Ueda exploram "El nacimiento de la letra jota como grafia consonántica". Partindo da versão paleográfica dos documentos do *Corpus Histórico del Español Norteño* (CORHEN), analisam as diferentes escritas (visigótica, carolina-gótica, entre outras) e tipos paleográficos que estão na origem da letra consonântica, tendo em conta a sua posição na palavra.

No texto seguinte, intitulado "El castellano en los orígenes del cambio gramatical: el pretérito imperfecto de la 2ª y 3ª conjugación (*-ié / -ía*)", Mª Carmen Moral del Hoyo também

recorre ao CORHEN para analisar formas do pretérito imperfeito na documentação asturo -leonesa, castelhana e navarro-aragonesa.

A terceira secção termina com o texto de Inés Carrasco Cantos y Livia Cristina García Aguiar intitulado "Análisis de la sufijación en el corpus DITECA". A partir dos textos de tradição jurídica que integram o suporte que deu origem ao *Dicionário de textos concejiles de Andalucía* (DITECA), as autoras propõem estabelecer as bases para poder estudar a vitalidade, produtividade ou o desgaste dos sufixos no período que vai do séc. XIII ao séc. XVIII.

Finalmente a última secção debruça-se sobre questões linguísticas iberorromânicas e linguística de corpus. Abre com o texto de Beatriz Arias Álvarez y Juan Antonio Hernández Mendoza sobre a caracterização do espanhol na Nova Espanha no séc. XVI, do ponto de vista dialetológico e sociolinguístico. Simultaneamente é apresentado o *Corpus Electrónico del Español Colonial Mexicano* (COREECOM) que serve de base para o estudo de algumas estruturas com artigo definido e o possessivo.

O texto de Marta Fernández Alcaide diz igualmente respeito ao espanhol da América e intitula-se "Manifestaciones de la variación del español colonial en un corpus epistolar multidimensional". São feitas considerações sobre a configuração de um corpus de espanhol americano, tendo em conta as tradições discursivas e as características linguísticas. A autora analisa a carta de um espanhol emigrante na América e aborda, entre outras questões linguísticas, a variação sintática aí encontrada.

O volume encerra com o texto de Olivier Iglesias "«Se le quedó mirando»: la atracción de clíticos en un corpus de idiolectos". O objetivo é analisar a posição do clítico num certo tipo de complexos verbais pouco frequentes e pouco estudados, sob o ponto de vista idioletal. Foram escolhidos registos escritos de quatro escritores espanhóis (sécs. XIX-XX) e produções de duas bloguistas espanholas (séc. XXI).

Há duas ideias centrais que percorrem todo o volume: por um lado, os vários textos evidenciam a utilidade de recorrer a corpus para estudar diversos tópicos em diferentes áreas da linguística (sintaxe, fonologia, morfologia, léxico) e sob diversas perspetivas teóricas. Há corpora que são a base de vários trabalhos, como o CE/CdE (cf. Enrique-Arias, Nieuwenhuijsen), o CORDE (cf. Enrique-Arias, Octavio de Toledo e Huerta, Rojo), que são corpora gerais, de referência, mas também corpora mais específicos como o CORHEN (cf. Torrens Álvarez, Moral del Hoyo). A especificidade de alguns corpora advém, por exemplo, da variedade regional que cobrem (CORHEN, CICÀ, COREECOM), do período histórico (CICA, CIPM, Gondomar), ou do tipo de texto (cartas pessoais: Post Scriptum; textos jurídicos: DITECA; textos bíblicos: BM) e o papel das tradições discursivas pode ser aqui muito bem explorado.

Paralelamente, é notório que não só a criação dos corpora está cada vez mais apurada e posta aos serviços dos utilizadores como a relação entre a informática e a linguística está mais próxima e são vários os trabalhos neste volume com incidência na análise estatística e quantitativa.

Em suma e como já foi dito anteriormente, este volume constitui-se como uma referência incontornável sobre os corpora diacrónicos no contexto iberorromânico. Adicionalmente, por se tratar de uma obra que contém trabalhos de reflexão e exploração desses corpora, assume uma dimensão ainda mais internacional do que os corpora que aqui estão representados.

Sandra Pereira