1. INTRODUCIÓN

Desde os primeiros corpus electrónicos creados na década dos sesenta, é sabido que a lingüística de corpus experimentou unha mellora e expansión progresivas, especialmente a partir dos avances tecnolóxicos que tiveron lugar nas últimas décadas. Actualmente, o emprego de computadoras e ferramentas informáticas cada vez máis avanzadas permite almacenar, procesar e analizar grandes cantidades de datos textuais cunha eficacia e rapidez inimaxinables hai non moitos anos, o que sitúa a lingüística de corpus, incluíndo á propia creación de corpus e demais recursos electrónicos, nunha área de seu dentro do estudo da linguaxe.

A lingüística histórica non foi allea a esta revolución tecnolóxica, como acredita o crecente interese no desenvolvemento de corpus históricos en diferentes linguas e xéneros (Claridge 2008, Xiao 2008: 401-408) e como confirma o estado actual da investigación diacrónica baseada en corpus electrónicos (Kytö 2011). Non obstante, na creación de recursos dixitais que sexan útiles para a investigación histórica das linguas cómpre afrontar alomenos dous tipos de problemas que non precisan atención cando se traballa con datos contemporáneos.

En primeiro lugar, o conxunto de datos dispoñible na elaboración de corpus históricos non admite comparación, nin cuantitativa nin cualitativamente, co conxunto de datos ao que temos acceso para construír corpus contemporáneos. Non podemos esquecer que a compilación e deseño dos primeiros está condicionada por certas limitacións ben coñecidas, que adoitan ser máis acusadas a medida que retrocedemos no tempo: carencia de datos falados, conservación fragmentaria de textos, dificultades de datación, problemas de contextualización, distribución errática de xéneros, etcétera (Kohnen 2007, Claridge 2008).

A estas limitacións, que son consubstanciais á investigación con documentación histórica, súmanse ademais certas cuestións de carácter técnico, específicas da lingüística de corpus e do procesamento de textos históricos, e que só nos úl-

timos anos comezan a ser abordadas con certa solvencia. Por un lado, preséntase o problema de delimitar a información do documento orixinal que vai formar parte do corpus e a que non. A metodoloxía convencional, como veremos, adoita centrar o foco preferiblemente no contido lingüístico, é dicir, palabras e signos de puntuación, poñendo unha menor atención nos aspectos relacionados coa dimensión paleográfica do documento fonte: disposición do texto, tipos de letra, emendas textuais, riscos, omisións, etcétera. Por outro lado, abrolla un problema relacionado coa ortografía orixinal dos textos recompilados. A variación ortográfica que presentan os textos non contemporáneos dificulta o procesamento automático dos anotadores lingüísticos, facendo necesaria unha normalización previa dos datos. En non poucas ocasións, este problema é presentado como unha disxuntiva (i.e. normalizar para aumentar as posibilidades de busca ou non normalizar para preservar a ortografía orixinal) claramente derivada das limitacións técnicas que impiden a consideración de ambos os niveis de edición nun mesmo corpus. Esta cuestión, ademais, está relacionada cun terceiro problema, o de como abordar a existencia de múltiples niveis de edición e anotación aplicables a un mesmo texto. A solución habitual é ofrecer versións separadas e independentes unhas das outras (o texto orixinal, o texto normalizado, o texto anotado), o que se traduce na práctica na creación e actualización de diferentes corpus, sen posibilidade de aplicar buscas cruzadas entre eles.

Este segundo conxunto de cuestións constitúe o punto de partida do presente traballo. O levantamento destes problemas de carácter técnico suxire a demanda dunha ferramenta que permita ao compilador de corpus históricos ofrecer os datos cumprindo alomenos cun triplo obxectivo: o de preservar as características textuais e peritextuais do documento fonte, o de potenciar a capacidade de busca de calquera aspecto incluído no corpus e o de garantir o mantemento do recurso electrónico mediante unha arquitectura que facilite posibles modificacións en calquera nivel de edición, permitindo igualmente a integración de novos niveis no futuro.

A plataforma *TEITOK* (Janssen 2016) foi pensada para dar resposta a problemas como os arriba citados e está deseñada especialmente para crear e manter corpus que combinan marcación textual e anotación lingüística. *TEITOK* permite integrar nun único sistema a edición diplomática dixital dun texto, segundo os estándares de codificación propostos polo consorcio TEI (*Text Encoding Ini-*

tiative), e o seu procesamento lingüístico, desde a normalización ortográfica ata a anotación sintáctica, incluíndo calquera outro nivel de edición que o compilador considere oportuno. Neste traballo expoñemos algunhas desvantaxes asociadas á metodoloxía convencional na construción de corpus históricos e explicamos as potencialidades de *TEITOK* como alternativa a dita metodoloxía.

2. A APROXIMACIÓN CONVENCIONAL NA CONSTRUCIÓN DUN CORPUS

A metodoloxía convencional na construción de corpus parte de extraer o contido lingüístico dos textos, basicamente palabras e signos de puntuación, separándoo ou "limpándoo" doutros aspectos que poidan entorpecer ou dificultar o almacenamento e o procesamento lingüístico. O proceso habitual na construción dun corpus, por conseguinte, pode ser resumido do xeito seguinte.

O punto de partida debe ser sempre o texto simple, é dicir, sen ningún tipo de formato. Se o compilador toma como fonte de datos edicións impresas, será preciso aplicar un proceso de limpeza para eliminar todo o que ten que ver con cuestións de estilo destinadas á representación gráfica do texto. Se o compilador parte de documentación manuscrita, o que é menos frecuente, obviará xeralmente todos os aspectos presentacionais relacionados coa propia campaña ou campañas de escrita do manuscrito. O obxectivo é ficar co contido puramente lingüístico e codificado preferiblemente nun conxunto de caracteres estandarizado que evite problemas de procesamento. Co conxunto de datos así almacenado, o seguinte paso é dividilo en *tokens*, o que seguramente obrigue a adoptar certas decisións para os casos de contraccións, enclíticos e formas semellantes, e transformalo nun formato máis manexable (habitualmente nunha disposición verticalizada cun *token* por liña) ben como resultado ou ben como paso previo á aplicación de programas de anotación automática, lematizadores, etcétera. Finalmente, todo este proceso adoita estar dirixido, en último termo, á implementación dunha interface con certas opcións de busca que permita recuperar resultados en forma de concordancias, listas de frecuencias ou calquera outro formato habitual na investigación lingüística baseada en corpus.

É claro que esta metodoloxía convencional está enfocada a analizar estatisticamente os datos e non a representar fielmente os textos. De feito, nos corpus

lingüísticos a miúdo a simple lectura dos textos, se é que están completos e non fragmentados, ou ben non é cómoda ou ben directamente non é posible, ofrecendo ao usuario, no mellor do casos, un contexto máis ou menos amplo da palabra buscada. Trátase, en definitiva, dunha estratexia válida para o tratamento e a construción de corpus a partir de datos contemporáneos, cuxo interese paleográfico é irrelevante ou mesmo inexistente; no entanto, a súa aplicación na elaboración de corpus históricos, e particularmente na elaboración de corpus a partir de documentación manuscrita, resulta nun produto imperfecto, pois omite toda unha serie de información paleográfica da fonte orixinal que non só resulta interesante a ollos do filólogo ou do historiador, senón que pode ser mesmo relevante na propia análise lingüística, como ilustramos máis adiante (véxase o apartado 3).

Esta perda de información paleográfica nos corpus históricos convencionais non só está ocasionada polas tarefas de procesamento encamiñadas a extraer o contido textual como valor único de análise, senón que deriva igualmente dunha práctica habitual na compilación de corpus históricos, a saber: o uso de edicións modernas como fonte de datos. En efecto, o emprego dunha edición moderna predomina claramente sobre a transcrición do texto orixinal como método de recompilación de datos históricos. A razón deste predominio non é difícil de imaxinar. A edición moderna dun texto antigo é, por regra xeral, de fácil acceso, evita a toma de decisións editoriais, pois o traballo filolóxico xa está feito de antemán e, sobre todo, axiliza considerablemente a tarefa de recompilación. Se a edición en cuestión xa existe en formato electrónico, o proceso limítase a unha exportación e acaso un procesado simple do texto; se a edición non está aínda dispoñible na rede, a tecnoloxía actual no eido do recoñecemento óptico de caracteres permite escanear un documento impreso de maneira rápida e eficaz. En definitiva, partir de edicións modernas supón unha maior accesibilidade, facilidade e rapidez de dixitalización, liberando ao lingüista do tempo e esforzo que implica transcribir e editar fontes primarias non contemporáneas. Non é de estrañar, polo tanto, que esta situación de *philological outsourcing*, como é denominada por Dollinger (2004), constitúa a solución habitual na compilación de corpus históricos:

> The compiler is confronted with the task of computerization and would like to use, and in many cases due to time and labour constraints is bound to use, the work of philolo-

> gists as a base. If an edition of a given text can be found, why should any time be dedicated to the transcription of texts from manuscript sources? (Dollinger 2004: 5).

Desde un punto de vista práctico, o uso de edicións modernas está fóra de toda dúbida, sobre todo no caso de macrocorpus e corpus de referencia, para os que a reedición *ad hoc* do material utilizado suporía unha empresa colosal. Desde un punto de vista lingüístico, non obstante, cómpre subliñar que non é a opción máis axeitada; ao contrario, o emprego de edicións como alternativa á transcrición do texto orixinal comporta unha serie de desvantaxes que xa foron debidamente sinaladas en diferentes estudos: mestura de diferentes criterios editoriais, representación superficial do manuscrito, adulteración da ortografía orixinal, etcétera (Lass 2004, Dollinger 2004, Grund 2006, Claridge 2008: 250-251, Honkapohja *et al.* 2009: 456-460).

A modo de recapitulación, digamos que na elaboración de corpus históricos existen dúas decisións relevantes que caracterizan á metodoloxía convencional. No momento de obter os datos, os corpus tradicionais tenden a usar edicións modernas de textos históricos en lugar de partir da transcrición do orixinal. No momento de procesar os datos, os corpus tradicionais tenden a centrar o interese nas expresións puramente lingüísticas en lugar de integrar os diferentes aspectos que conforman un documento, tanto textuais como paratextuais. En ambos os casos, por tanto, a fidelidade ao texto orixinal e a información paleográfica vense claramente prexudicadas.

Un corre o risco de pensar que esta dimensión paleográfica (a disposición do texto no documento, o tipo de grafía, a presenza de texto riscado ou ilexible, as adicións fóra de liña e outro tipo de cuestións de semellante natureza) constitúen cuestións periféricas e pouco menos que irrelevantes na análise lingüística e que, polo tanto, supón unha perda asumible no proceso de edición dun corpus. Contra esta visión monodisciplinar, xorden voces que demandan unha aproximación multidisciplinar da análise textual e, por conseguinte, a integración da información non puramente lingüística nas edicións que vaian nutrir un corpus (Meurman-Solin / Tyrkkö 2013, Marttila 2014). Esta liña de traballo sostén que un corpus cinguido a recoller unicamente as expresións lingüísticas de documentos históricos ten de feito limitada a súa utilidade para o estudo da lingua, podendo conducir mesmo a conclusións erróneas:

> Now that diachronic corpora provide us with large quantities of data, errors resulting from the misinterpretation of insufficiently contextualised linguistic items may be multiplied in the analysis. In other words, if we edit digitally and annotate the *language* of texts exclusively and reproduce quite imprecisely, or not at all, non-linguistic features such as layout, script type, and contracted and abbreviated word-forms, the consequences may be serious in both quantitative and qualitative analysis (Meurman-Solin / Tyrkkö 2013).

Ofrecemos a continuación un par de exemplos para ilustrar a importancia que pode adquirir o nivel paleográfico na análise lingüística. Ambos están tomados do corpus epistolar *Post Scriptum* (CLUL 2014) e refírense a cuestións de carácter gramatical da lingua española.

O primeiro caso (Figura 1) está tomado dun estudo sobre a variación pronominal dos clíticos de terceira persoa e, concretamente, sobre a ocorrencia das formas *la* e *las* en función de obxecto indirecto no citado corpus (Vaamonde 2015). Un dos obxectivos deste estudo foi a exportación a un mapa da localización de autores laístas rexistrados no corpus e a súa comparación coa situación dialectal deste fenómeno na época actual. O exemplo en cuestión, que ofrecemos a continuación, constitúe un dos poucos casos que se afastan da zona laísta actual (trátase dun autor procedente do sur de Estremadura):

> Vm me haga m(e)r(ce)d darme noticia de mi s(eño)ra D(oñ)a m(arí)a Lopes de castro a quien escrevi de Amberes q(ue) estimare **la vaia bien**

O que importa destacar aquí é o feito de que o manuscrito revela unha pequena corrosión da tinta na vogal do clítico (liña 8 na Figura 1), o que esperta certas reservas sobre a forma realmente empregada por este autor: *la* ou *le*? Aqueles casos nos que, debido ao estado de conservación do manuscrito, a ocorrencia *la* é dubidosa poden ser transcritos como *lo*, *la* ou *le*, dependendo da interpretación do transcritor no momento de codificar o texto. Estes casos non deberían ser tidos en conta nunha análise sobre laísmo ou, alomenos, deberían poden ser claramente identificados para que o investigador decida o que facer con eles. Non obstante, tales ocorrencias de laísmo conxecturado son indetectables nos corpus históricos convencionais.

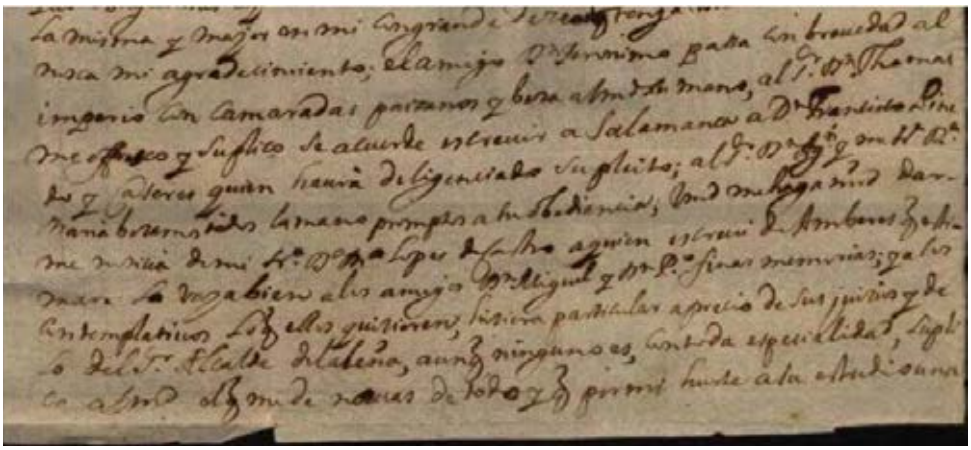

Figura 1. Fragmento de carta de 1687 (referencia no corpus: PS6044)

O segundo exemplo (Figura 2) está relacionado coa variación producida nas oracións subordinadas substantivas encabezadas por *que* e *de que*, referida con frecuencia na bibliografía cos termos de queísmo (emprego da conxunción *que* en contextos onde a construción normativa é *de que*) e dequeísmo (emprego de *de que* en contextos onde a construción normativa é *que*). O fragmento que interesa rescatar aquí é o que recollemos a continuación:

> prudente y sabio; cuyo dictamen firmado quisiera ver para **convencer mi mucha ignorancia de que** con tan notables defectos no puede persuadirse a que la Ley permita empecer y abochornar a una familia tan ilustre con semejante mezcla

O manuscrito revela que nunha primeira campaña de escrita o autor escolle a variante *que* para a subordinada substantiva (o que denotaría un caso de queísmo, dado que o verbo *convencer* esixe a presenza da preposición), pero despois adiciona a preposición *de* por riba da liña (liña 2 na Figura 2).

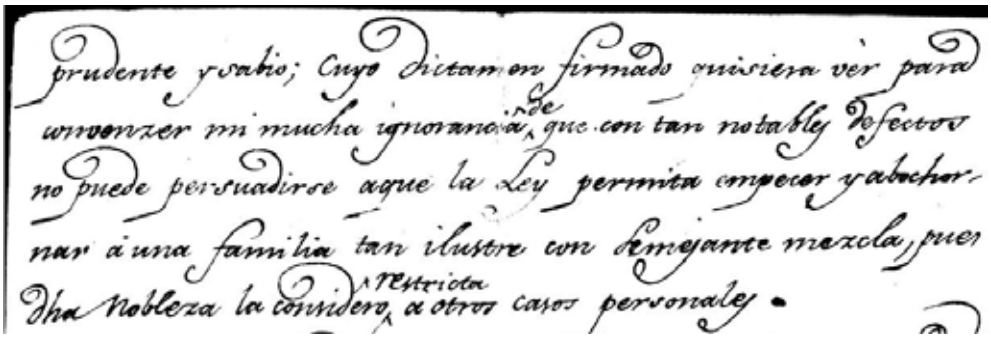

Figura 2. Fragmento de carta de 1796 (referencia no corpus: PS5131)

Desde o punto de vista cualitativo, casos coma estes poden resultar especialmente interesantes para o estudo do (de)queísmo, pois reflicten na escrita o conflito entre norma e uso que está a funcionar nesta variación gramatical. De novo, non obstante, é preciso contar con edicións dixitais que integren sistematicamente estes e outros aspectos semellantes para poder ser identificados no momento de obter os datos do corpus.

3. A EDICIÓN DIXITAL EN LINGUAXE TEI

Na procura de corpus históricos que permitan integrar e explorar tanto a dimensión lingüística como a dimensión paleográfica dos textos, constitúe unha peza decisiva o traballo que desde hai varios anos se ben facendo no eido das humanidades dixitais e, particularmente, na creación dun estándar de marcación para a representación de textos en formato electrónico. Este estándar, coñecido como TEI e baseado na linguaxe XML, consta na súa versión máis recente (P5) de máis de 500 elementos e decenas de atributos que permiten codificar dixitalmente aspectos textuais de moi diversa índole: cuestións estruturais (capítulos, apartados, títulos, páxinas, liñas, versos), cuestións presentacionais (adicións, riscos, lagoas, roturas, subliñados), cuestións conceptuais (nomes de lugares ou de persoas), etcétera.

Publicado inicialmente no ano 1987, TEI representa na actualidade un estándar plenamente consolidado na comunidade científica das humanidades dixitais.

Non obstante, o seu emprego na compilación de corpus históricos e, en xeral, no ámbito da lingüística de corpus, ten sido infrecuente ou practicamente inexistente:

> The searchability of a corpus is crucially dependent on how the corpus has been annotated. Again, there is a lack of consensus on this point, and compilers of historical corpora have been slow or even reluctant to apply standards such as the Text Encoding Initiative (TEI) Guidelines (P5). Many of the better known corpora are annotated for the main textual features but not all, and not as exhaustively as could have been the case (Kytö 2011: 437).

Con todo, si existen corpus históricos que fan uso das directrices TEI, como é o caso do *British National Corpus* ou *The Lampeter Corpus of Early Modern English Tracts* e, de feito, semella que a lista vai en aumento nos últimos anos. Podemos citar aquí *The Corpus of Middle English Prose and Verse*, *The Coruña Corpus of English Scientific Writing*, *The Corpus of Modern Greek Poetry* ou *The Seville Corpus of Northern English*.

O obxectivo da linguaxe TEI é ofrecer á comunidade científica un estándar rigoroso e completo para representar os textos en formato electrónico. Polo tanto, a súa razón de ser é a de servir como guía na tarefa de marcación. Unha vez o texto está debidamente marcado, non obstante, é necesario ou ben crear ferramentas propias ou ben recorrer a ferramentas externas para publicar e analizar os datos. Este salto que vai desde a marcación á publicación pode resultar complexa e problemática, como xa demostraron Burghart / Rehbein (2012) mediante unha enquisa destinada aos membros da comunidade TEI que traballaban precisamente con documentación manuscrita:

> The results demonstrate the existence of a steep learning curve for the TEI, where many practitioners are self-taught and where learning-by-doing dominates; there exists a long gap between the first encounter with the TEI and its actual use in projects. Survey results highlight the need for user-friendly, bespoke tools facilitating the processing, analysis, and publishing of TEI-encoded texts (Burghart / Rehbein 2012).

Os resultados da enquisa elaborada por Burghart / Rehbein suxiren a necesidade de contar con ferramentas que faciliten ao usuario a dispoñibilización dos seus ficheiros TEI na rede. Actualmente, existen cada vez máis plataformas que

cumpren este cometido. Tal é o caso de *Versioning Machine* (Clement / Schreibman 2010), *TEICHI* (Pape *et al.* 2012) ou *TEIViewer* (Schlitz / Bodine 2009), por citar só algunhas. Non obstante, estas e outras plataformas similares non foron deseñadas para cumprir as demandas do lingüista de corpus (senón máis ben as do filólogo) e, en consecuencia, as súas opcións de consulta e de visualización de resultados son neste sentido de pouca utilidade. Existe, polo tanto, a necesidade de contar cunha plataforma pensada para visualizar edicións diplomáticas dixitais en TEI e que, simultaneamente, permita construír, xestionar e explotar corpus históricos compilados a partir desas edicións diplomáticas. É aquí onde entra en xogo a plataforma *TEITOK*.

4. *TEIKOK*: DA EDICIÓN DIXITAL AO CORPUS LINGÜÍSTICO

Na plataforma web *TEITOK*, un corpus consiste nun conxunto de ficheiros en formato XML e, preferentemente, en TEI/XML, é dicir, codificados segundo as directrices de marcación de textos propostas por dito consorcio. Cada un destes ficheiros, que é editado de xeito independente, contén a descrición dos aspectos máis relevantes do documento fonte ao que representa. O estándar TEI asegura a marcación dun amplo tipo de fenómenos de carácter paleográfico: adicións, cancelacións, cambios de man, letras capitais, cores, etcétera.

TEITOK proporciona varias formas de realizar transcricións a partir das imaxes facsimilares, coa idea de que os filólogos poidan crear edicións dixitais dos textos. Por suposto, estas edicións dixitais poden adoptar como punto de partida outra edición xa existente; no entanto, neste traballo entendemos que a opción máis adecuada é a de crear edicións propias e que, idealmente, todos os documentos utilizados na construción do corpus histórico compartan os mesmos criterios de transcrición.

No propio ficheiro XML e como complemento á marcación TEI, a plataforma *TEITOK* permite realizar a tokenización do texto e, sucesivamente, engadir anotación lingüística a cada token. A creación do corpus é feita de xeito automático a partir do conxunto de ficheiros TEI/XML e a súa explotación permite dar resposta a cuestións de natureza lingüística.

4.1. Tokenización e anotación

Por cada documento TEI almacenado na plataforma *TEITOK*, o sistema aplica unha tokenización interna (*inline*), isto é, os *tokens* represéntanse directamente como nodos XML dentro do propio documento. A división en *tokens* é algo habitual no procesamento dun texto e, de feito, está contemplada polo estándar TEI, que suxire marcar as palabras co elemento <*w*> e a puntuación co elemento <*c*>. Cómpre apuntar, non obstante, que moitos dos recursos que aplican o estándar TEI para tokenizar os textos non están a operar na verdade cun documento TEI ao que se lle aplica unha tokenización interna, senón cun corpus construído segundo a estratexia convencional (i.e. unha palabra por liña) que é logo convertido en linguaxe TEI/XML (tal é o caso do BNC, por exemplo). A estratexia usada en *TEITOK* parte de considerar que o documento TEI contén información relevante que debe ser conservada, e por iso a tokenización é engadida sobre a propia marcación xa existente e non como parte dun proceso externo.

Unha vez que o documento está tokenizado é posible engadir a anotación lingüística pertinente para cada *token*. A anotación pode ser de dous tipos. Por unha banda, considérase a anotación propiamente dita, é dicir, etiquetas que informan sobre a clase de palabra (nome, adxectivo, verbo, etc.) e as súas correspondentes formas lematizadas. Por outra banda, considéranse as realizacións ortográficas alternativas da forma orixinal, como son as formas con grafía normalizada ou as abreviaturas desenvolvidas. A razón fundamental de normalizar as palabras dun corpus é a de mellorar os resultados de etiquetadores e *parsers* na anotación automática (Baron / Rayson 2008). Nos corpus históricos, no entanto, a normalización ortográfica serve para outros propósitos non menos importantes: facilita a lectura dos textos, constitúe unha interpretación da intención autorial, fai posible a busca de palabras con independencia da súa ortografía orixinal, etcétera. Polo tanto, no deseño de *TEITOK* a normalización ortográfica é un aspecto que ten relevancia en si mesmo e por si mesmo, sendo o incremento na precisión de ferramentas de tratamento automático tan só unha vantaxe adicional.

Con todo, convén lembrar que non existe un concepto único de normalización: as palabras poden ser normalizadas na ortografía máis habitual utilizada polo propio autor, ou segundo a variedade estándar correspondente ao período no que foi escrito o texto, ou en función da ortografía actual da lingua. Todas

estas posibilidades responden a intereses diferentes e atopan acomodo en diferentes edicións posibles para un mesmo texto: edición diplomática, edición moderna, edición semipaleográfica, etcétera. Así as cousas, o sistema permite traballar non só cun nivel de regularización ortográfica, senón con tantos como sexan necesarios para satisfacer as necesidades do editor ou do compilador. O nivel básico é ou debe ser sempre unha transcrición o máis fiel posible ao orixinal, e é neste nivel inicial onde intervén o estándar TEI co seu rico conxunto de elementos destinados a marcar calquera complexidade presente no documento fonte. O resto de niveis constitúen anotacións adicionais da forma inicial e nun formato puramente textual, o que quere dicir que as alternativas ortográficas non poden conter código XML. Esta estratexia de normalización difire da utilizada no estándar TEI/XML, onde todas as variantes son tratadas como entidades XML.

4.2. Visualización en HTML

Os documentos en *TEITOK* poden ser consultados individualmente nun navegador. A interface aplica diferentes cores para marcar as diferentes anotacións do texto, como son as adicións, os riscos, as palabras ilexibles, etcétera. Posto que *TEITOK* é un sistema pensado para traballar con corpus, a visualización por defecto dos documentos está orientada ao texto e, en consecuencia, presenta a transcrición en forma de texto corrido. Non obstante, o sistema obedece a un deseño modular no que un mesmo documento TEI/XML pode ser visualizado de diversas maneiras, dependendo do tipo de documento. No caso de documentación facsimilar, *TEITOK* ofrece unha visualización similar á de plataformas como *EVT* (Rosselli Del Turco *et al.* 2014), coa imaxe do facsímile á esquerda e a transcrición á dereita.

Unha vez tokenizado o documento TEI, cada palabra pode estar asociada a múltiples realizacións ortográficas; por iso na interfaz de *TEITOK* o texto pode ser presentado de diferentes formas. Mediante o uso de simples botóns, o usuario pode seleccionar a versión orixinal de cada palabra, pero tamén a súa forma normalizada, a súa forma desenvolvida (no caso de abreviaturas) e así sucesivamente. Isto significa que a partir dun mesmo ficheiro XML é posible visualizar e xerar a versión semipaleográfica do texto ou a versión diplomática ou a versión normalizada, xa que todas elas están contidas no propio ficheiro. Desta forma, a

normalización do texto non se utiliza só para buscar e procesar, senón que pode ser usada para xerar as diferentes edicións do documento a partir dunha única fonte de datos.

Finalmente, o feito de que as diferentes edicións poidan ser creadas a partir dun único ficheiro XML ten outra vantaxe non menos importante. Só existe un única fonte de datos que precisa ser actualizada con posibles correccións ou modificacións, fronte á problemática que suscita ter múltiples versións independentes dun mesmo documento. Este aspecto resulta ser especialmente vantaxoso no traballo filolóxico.

4.3. Edición en HTML

A forma convencional de construír un corpus anotado é mediante unha arquitectura canalizada dos datos: por un lado entran os documentos orixinais, por outro lado sae o corpus. Este proceso divídese en fases, de forma que cada fase trata o documento sobre o resultado da fase anterior: primeiro límpase o texto, despois é normalizado, despois é etiquetado e lematizado, e despois é convertido nun corpus. Esta forma de traballar é unidireccional, é dicir, complétase unha fase e o proceso continúa cara á seguinte. Asúmese, ademais, que os datos de entrada non serán alterados unha vez que comeza o proceso, unha condición por outro lado irrelevante sempre que o punto de partida sexa un texto nacido xa en formato electrónico. Non obstante, no caso de documentación histórica o punto de partida é por regra xeral un manuscrito, cuxa transcrición sempre pode conter algún erro. Ademais, a normalización ortográfica de textos históricos e manuscritos tamén é máis propensa a conter erros, pois a variación ortográfica característica deste tipo de documentación implica en moitos casos un traballo de interpretación que merece ser revisado. E, por último, o tratamento automático de textos históricos tende a ser menos preciso, o que esixe unha corrección manual das etiquetas morfosintácticas e dos lemas.

Debido a estas particularidades, *TEITOK* asume que calquera fase no proceso de construción do corpus está suxeita a revisión manual e, precisamente por iso, facilita diferentes opcións de edición. A forma máis básica de edición, aínda que tamén a máis flexible, é permitir diferentes vías de acceso ao documento XML a través do propio navegador. O usuario, unha vez rexistrado no sistema como

administrador, pode premer en calquera palabra do texto e editar posibles erros a partir dun formulario HTML básico, isto é, sen necesidade de consultar o código XML subxacente.

4.4. Explotación de corpus en *TEITOK*

Para facilitar e potenciar as buscas sobre os documentos XML, o sistema parte de ditos documentos e crea un corpus na plataforma *Corpus Workbench* (Evert / Hardy 2011), que vén integrado coa potente linguaxe de consulta *CWB Query Language (CQL)*. No proceso de creación de corpus, que se fai automaticamente mediante un simple clic ou ben pode ser programado para ser executado periodicamente, o sistema abre cada ficheiro XML e crea o corpus a partir de cada palabra contida nos documentos, rexistrando igualmente a ortografía normalizada, a anotación lingüística, as adicións, os riscos, etcétera. A interface de busca permite obter resultados segundo diferentes opcións: concordancias, *keywords*, colocacións, etcétera.

Toda a anotación lingüística do corpus é recuperable en linguaxe CQL. Polo tanto, o usuario non só pode buscar palabras na súa grafía orixinal, senón tamén na grafía normalizada. A seguinte consulta buscará todas as palabras do corpus que presentan a forma regularizada *ação*, con independencia da súa grafía orixinal no corpus:

[reg="ação"]

Tamén é posible combinar diferentes realizacións ortográfica para buscar, por exemplo, todas as palabras do corpus que terminan en *-çon*, pero que non presentan unha forma regularizada terminada en *-ção*:

*[form=".*çon" & reg != ".*ção"]*

No momento de construír o corpus en CWB, *TEITOK* rexistra a localización exacta de cada palabra dentro do documento XML. Isto permite ao sistema utilizar o resultado dunha busca e identificar o fragmento correspondente no XML orixinal, podendo así presentar este último en lugar do resultado en texto plano. Noutras palabras, os resultados dunha consulta ofrecen toda a marcación

TEI presente no documento orixinal, incluíndo as palabras riscadas que foron excluídas do corpus.

O feito de que toda a marcación TEI da edición dixital poida ser recuperada nos resultados da consulta axuda a identificar rapidamente exemplos que poidan ser particularmente interesantes así como a descartar outros que poidan ser problemáticos. Por exemplo, e retomando o caso de (de)queísmo apuntado no apartado 2, podemos estar interesados en estudar o comportamento do verbo *convencer* en relación con esta variación sintáctica, para o cal executaríamos a seguinte consulta en linguaxe CQL:

[lemma="convencer"] []{0,4} [form="que"]

Esta consulta nos devolverá todas as ocorrencias do lema *convencer* seguido da forma *que* e, opcionalmente, ata un máximo de catro palabras entre o verbo e a forma *que*. O resultado desta consulta no corpus *Post Scriptum* (incluíndo datos do español e do portugués) é ofrecido na Figura 3 a continuación:

context	uno producir desde aqui.	**Combencette que**	no podemos miramos sin
context	Sn Martin | me a	**Conbencido diciendome que**	las Cobrancas de | las Mulas
context	ninguno. Debe U igualmente	**combencerse de que**	en esta Ciudad escasean los
context	. | Por el mismo se	**convencera U que**	conozco antes | de ahora la
context	q V E | ja estará	**convencido, que**	na publicação do
context	minhas chaves; Estou | seguramte	**convencido que foi elle | que**	me roubou, e pello
context	e estou | tão	**convensido disto que**	não duvidaria | publicálo,
context	dictamen firmado quisiera ver para	**convenzer mi mucha ignorancia de que**	con tan notables defectos | no

Figura 3. Exemplo dun resultado de consulta en *TEITOK* (corpus Post Scriptum)

O último resultado ofrecido por *TEITOK* mostra a preposición *de* en azul e por riba da liña, o que indica que esa palabra foi engadida no manuscrito nunha segunda campaña. Como xa foi comentado, esta adición *a posteriori* da preposición converte o exemplo nun caso cualitativamene interesante para o estudo do (de)queísmo; non obstante, en termos cuantitativos pode ser preferible non contabilizalo nin como caso de queísmo nin como caso de emprego normativo con preposición *de*. Sexa como for, a visualización dos resultados en *TEITOK* permite identificar este e outros exemplos semellantes cun simple golpe de vista, quedando á vontade do investigador a súa inclusión ou non nas análises que rea-

lice. De xeito similar, un usuario interesado na variación pronominal do español pode recuperar, por exemplo, todas as ocorrencias do clítico *la*. Posto que na visualización de resultados en *TEITOK* as palabras de lectura dubidosa aparecen destacadas en fondo verde, o usuario pode facilmente desestimar eses resultados para traballar só con ocorrencias inequívocas do clítico. Se o usuario está interesado en obter máis información, o botón *context* situado á esquerda de cada exemplo facilita o acceso ao texto completo xunto coa imaxe do facsímile. Desta forma, é posible estudar fenómenos particulares dos documentos históricos sen perder de vista a evidencia subxacente no manuscrito.

Cómpre sinalar que aínda que as adicións (ou as lecturas dubidosas) sexan visibles directamente nos resultados da consulta, como se amosa na Figura 3, iso non significa que esa información sexa tida en conta ao consultar datos estatísticos. Sería teoricamente posible ter en conta esa información, pero requiriría unha consulta moito máis complexa e facilmente esquecible. Ademais, só funcionaría para palabras enteiras, xa que unha arquitectura como a do CWB está baseada no *token* e non pode representar información relativa a caracteres específicos dentro dunha palabra. En suma, este método permite visualizar rapidamente aspectos paleográficos que poden ser relevantes nos resultados lingüísticos, e facilita así mesmo o acceso á imaxe do manuscrito, pero aínda así segue sendo necesario tomar con certa cautela os resultados obtidos a partir de corpus en lingüística histórica.

5. CONCLUSIÓNS

Como explicamos neste traballo, os corpus electrónicos constitúen unha ferramenta particularmente importante para a lingüística histórica. Certo é que as limitacións que deben afrontar os corpus históricos en termos de representatividade e conservación de fontes dificultan ou ás veces mesmo invalidan a aplicación de métodos estatísticos propios da lingüística de corpus. Pero, mesmo cando un corpus é usado simplemente como fonte de datos para atopar exemplos dun determinado fenómeno, resulta crucial que os documentos históricos teñan sido minuciosamente tratados, pois os segmentos riscados, engadidos, ilexibles ou conxecturados deben de ser utilizados con especial cautela na investigación de fenómenos lingüísticos. Non obstante, as técnicas tradicionais na lingüística

de corpus non sempre conservan a información relativa a tales segmentos, o que dificulta a correcta explotación de corpus históricos.

A aproximación que se defende neste traballo, e que toma corpo na plataforma *TEITOK*, permite superar este problema mediante o uso de documentos TEI/XML como punto de partida para a creación de corpus lingüísticos, que deste xeito inclúen toda a marcación relevante do documento en cuestión. En *TEITOK*, o corpus é creado a partir dos documentos TEI/XML e os resultados da busca recuperan o fragmento TEI/XML correspondente, o que permite visibilizar directamente os segmentos engadidos, cancelados ou conxecturados. En caso de dúbida, o sistema permite redirixir ao usuario desde a páxina de resultados á imaxe facsimilar para consultar e verificar calquera aspecto do documento orixinal. Ademais, *TEITOK* ofrece a posibilidade de traballar con diferentes niveis de regularización ortográfica e anotación lingüística, cuxa información tamén é rexistrada no proceso de creación co corpus. Esta información pode ser recuperada e cruzada mediante buscas simples ou complexas en linguaxe CQL.

En suma, o ambiente de traballo *TEITOK* dá resposta a todos os problemas apuntados na introdución do presente traballo. En primeiro lugar, permite aos filólogos crear edicións dixitais en formato TEI/XML, deixando así rexistrada toda a información do documento fonte que se considere relevante. En segundo lugar, proporciona unha interface sinxela para crear ou editar os ficheiros XML, que poden chegar a conter moita información, sen necesidade de traballar directamente no propio código. En terceiro lugar, unha vez que o documento XML foi creado na plataforma (ou importado a ela), a súa dispoñibilidade en liña é automática, facendo prescindible a necesidade dun especialista informático que solucione o paso da marcación á publicación. En cuarto lugar, posto que as edicións dixitais son creadas especificamente para a creación do corpus, tamén neste último está asegurado un mesmo e único conxunto de criterios paleográficos. E, finalmente, o sistema recupera e amosa a marcación paleográfica dos ficheiros XML ao longo de todas as buscas e visualizacións, o que significa que toda a información que é importada ao corpus permanece dispoñible en calquera momento.

TEITOK é unha ferramenta útil e potente que está a ser usada na actualidade por un número cada vez maior de corpus históricos, entre os que cabe citar *Post Scriptum* (CLUL 2014), *Gondomar* (Álvarez / González Seoane 2017), *Oralia Diacrónica del Español* (Calderón Campos / García-Godoy 2018) ou *Corpus de*

Textos Antigos (Sobral *et al.* 2015). A aproximación adoptada neste traballo xira arredor da realización de buscas en corpus históricos formados por manuscritos individuais ou monografías. A aplicación de buscas en corpus formados por múltiples versións dun único documento, conservado ou perdido, require unha implementación máis complexa cuxa discusión queda fóra dos límites do presente traballo.

Referencias bibliográficas

Álvarez, Rosario / Ernesto X. González Seoane (2017): *Gondomar. Corpus dixital de textos galegos da Idade Moderna.* Santiago de Compostela: Instituto da Lingua Galega. <http://ilg.usc.gal/gondomar/>.

Baron, Alistair / Paul Rayson (2008): «VARD 2: A tool for dealing with spelling variation in historical corpora», en *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University.

Burghart, Marjorie / Malte Rehbein (2012): «The Present and Future of the TEI Community for Manuscript Encoding», *Journal of the Text Encoding Initiative* 2. <https://doi.org/10.4000/jtei.372>.

Calderón Campos, Miguel / M.ª Teresa García-Godoy (2018): *Oralia diacrónica del español (ODE). Documentación manuscrita de los siglos XVI a XIX*. Granada: Universidad de Granada. <http://corpora.ugr.es.ode>.

Claridge, Claudia (2008): «Historical corpora», en A. Lüdeling / M. Kytö (eds.), *Corpus Linguistics*: *An International Handbook*. Berlin / New York: Walter de Gruyter, 242-259.

Clement, Tanya / Susan Schreibman (2010): «The Newest Version of the Versioning Machine (V4)», *2010 Conference and Members' Meeting of the Text Encoding Initiative*. Zadar, Croatia, 8-14.

CLUL (ed.) (2014): *P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna*. <http://ps.clul.ul.pt>.

Sobral, Cristina *et al.* (2015): *Corpus de Textos Antigos em Português até 1525*. <http://alfclul.clul.ul.pt/teitok/cta/>.

Dollinger, Stefan (2004): «'Philological computing' vs. 'philological outsourcing' and the compilation of historical corpora: a Late Modern English test case», en Christiane Dalton-Puffer *et al.* (eds.), *Vienna English Working Papers (VIEWS)* 13, 3-23.

Evert, Stefan / Andrew Hardie (2011): «Twenty- first century corpus workbench: Updating a query architecture for the new millennium», en *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham: University of Birmingham, <https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf>

Grund, Peter (2006): «Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights», *Journal of English Linguistics* 34, 105-125. <https://doi.org/10.1177/0075424206290255>.

Honkapohja, Alpo / Samuli Kaislaniemi / Ville Marttila (2009): «Digital Editions for Corpus Linguistics: Representing Manuscript Reality in Electronic Corpora», en Andreas H. Jucker / Daniel Schreier / Marianne Hundt (eds.), *Corpora: Pragmatics and Discourse*, Amsterdam / New York: Rodopi, 451-475. <https://doi.org/10.1163/9789042029101_023>.

Janssen, Maarten (2016): «TEITOK: Text-Faithful Annotated Corpora», en *Proceedings of the Language Resources and Evaluation Conference (LREC 2016) ELRA*. Portorož, Slovenia, 4037-4043.

Kohnen, Thomas (2007): «From Helsinki through the centuries: the design and development of English diachronic corpora», en Päivi Paht / Irma Taavitsainen / Terttu Navelainen / Jukka Tyrkko (eds.), *Studies in Variation, Contacts and Change in English. Volume 2: Towards Multimedia in Corpus Studies*. <http://www.helsinki.fi/varieng/series/volumes/02/kohnen/>.

Kytö, Merja (2011): «Corpora and historical linguistics», *Revista Brasileira de Linguística Aplicada* 11 (2), 417-457. <https://doi.org/10.1590/S1984-63982011000200007>.

Lass, Roger (2004): «Ut custodiant litteras: Editions, Corpora and Witnesshood», en Marina Dossena / Roger Lass (eds.), *Methods and Data in English Historical Dialectology* (Linguistic Insights 16). Bern: Peter Lang, 21-48.

Marttila, Ville (2014): *Creating Digital Editions for Corpus Linguistics. The case of Potage Dyvers, a family of six Middle English recipe collections*. Tese de doutoramento. Universidade de Helsinki.

Meurman-Solin, Anneli / Jukka Tyrkkö (2013): «Introduction», en Anneli Meurman-Solin / Jukka Tyrkkö (eds.), *Studies in Variation, Contacts and Change in English. Volume 14: Principles and Practices for the Digital Editing and Annotation of Diachronic Data* <http://www.helsinki.fi/varieng/series/volumes/14/introduction.html>.

Pape, Sebastian / Christof Schöch / Lutz Wegner (2012): «TEICHI and the Tools Paradox. Developing and Publishing Framework for Digital Editions», *Journal of the Text Encoding Initiative* 2. <https://doi.org/10.4000/jtei.432>.

Rosselli Del Turco, Roberto / Giancarlo Buomprisco / Chiara Di Pietro / Julia Kenny / Raffaele Masotti / Jacopo Pugliese (2014): «Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions», *Journal of the Text Encoding Initiative* 8. <https://journals.openedition.org/jtei/1077>.

Schlitz, Stephanie A. / Garrick S. Bodine (2009): «The TEIViewer: Facilitating the transition from XML to web display», *Literary and Linguistic Computing* 24 (3), 339-346. <https://doi.org/10.1093/llc/fqp022>.

Xiao, Richard (2008): «Well-known and influential corpora», en A. Lüdeling / M. Kytö (eds.), *Corpus Linguistics*: *An International Handbook*. Berlin / New York: Walter de Gruyter, 383-456.

Vaamonde, Gael (2015): «Distribución de leísmo, laísmo y loísmo en un corpus diacrónico epistolar», *Res Diachronicae* 13, 58-79. <https://resdi.net/volumen-xiii/>.