INFORMÁTICA E ANÁLISE TEXTUAL

Alberto Álvarez Lugrís
Universidade de Vigo
aalvarez@airas.cirp.es

1. Os obxectos da análise

1.1. Os textos electrónicos

1.1.1. Que son e onde podemos atopalos

Podemos defini-lo texto electrónico como unha versión doutro texto[1] que é manipulable mediante un ordenador e certos programas. Esto implica importantes diferencias coa idea tradicional de texto:

---

[1] Entendemos *texto* como calquera acto comunicativo, o cal quere dicir que o orixinal a partir do cal construímo-lo texto electrónico non ten por que ser necesariamente un texto impreso, un libro. Podemos crear textos electrónicos, por exemplo, a partir de textos orais non transcritos nin impresos, senón gravados en soporte de audio.

-o texto electrónico, en realidade, non é máis que unha sucesión de caracteres ASCII, é dicir, unha sucesión de bytes almacenada nun soporte determinado (en CD-ROM, en disquete ou, normalmente nos discos duros de potentes estacións de traballo situadas nos centros de textos electrónicos).

- o texto electrónico, polo tanto, non ten entidade física, non é un obxecto material: non o podemos ver nin tocar.

- por este mesmo motivo, o texto electrónico non posúe formatos de ningunha clase[2]: é un simple ficheiro dos chamados de texto puro ou ASCII puro, libre de efectos tipográficos como tamaños e tipos de letra, negriñas, cursivas, versaletas, debuxos, etc.

- o único xeito de acceder ó contido do texto electrónico (por exemplo, para lérmolo) é a través da pantalla do ordenador ou ben mediante impresión, co cal, compre ter en conta, non estamos accedendo ó texto electrónico, senón a unha representación do mesmo.

- un texto electrónico soe incorporar, ademais do texto propiamente dito, outros dous tipos de información: primeiro a chamada *cabeceira*[3], unha especie de introducción e presentación do texto na que se inclúe, polo menos, unha pormenorizada descrición bibliográfica da fonte orixinal (impresa, polo xeral) que serviu de base para elabora-lo texto electrónico, o(s) encargado(s) de transcribir e

---

[2] Aínda que, como veremos, dependendo dos obxectivos de cada estudio ou corpus concreto, os textos electrónicos poden incorporar certo tipo de marcación para representar todos eses efectos tipográficos. Pensemos, por exemplo, en córpora que conteñan textos nos que a disposición física dos textos orixinais sexa importante: caligramas en poesía, novelas con efectos especiais (*The Sound and the Fury*), etc. A maior parte dos estudios, sen embargo, non soen ter en conta estas características dos textos.

[3] O tipo de cabeceira máis completo é a *cabeceira TEI* (TEI header). Cf. o apéndice nº1, onde inclúo un exemplo. Cf. tamén as seccións 2.4 e 2.5.3.

codifica-lo texto e a súa posible vinculación a grupos de investigación ou centros de textos electrónicos, criterios de transcrición, criterios de codificación, marcación utilizada e a súa xerarquía, data de codificación, data(s) da(s) posible(s) modificación(s), e estatus legal do texto. E, segundo, certo tipo de marcación especializada coa que se especifica a información «metatextual» que cada usuario precise[4].

A historia dos textos electrónicos remóntase ó ano 1949, cando Busa elaborou o *Index Thomisticus*, e ás primeiras concordancias de textos, sobre todo bíblicos. Dende aquela data, o número de textos aumentou sen cesar. Ata hai relativamente moi pouco tempo foi difícil facer inventarios dos fondos existentes, pero hoxe sabemos que estes textos se contan por decenas de milleiros. Só os textos de OTA (Oxford Text Archive), ARTFL (American Research on the Treasury of the French Language), *Thesaurus Linguae Graecae* e do Instituto de Lingüística Computacional de Pisa sobrepasan con moito o número de 11.000[5].

Pero hai só trinta anos que os textos electrónicos se utilizan con asiduidade nas investigacións humanísticas. Cunha frecuencia cada vez maior, os estudiosos de tódalas ramas das Humanidades constrúen os seus propios textos (ou, máis recentemente, utilizan as cada vez máis numerosas coleccións) para construír concordancias, establecer frecuencias de palabras e/ou estructuras, elaborar índices de obras individuais ou de autores completos, facer análises estilísticas que sirvan de base a estudios sobre autoría, redactar diccionarios de

---

[4] Cf. a sección 2.5.

[5] Segundo Hockey (1994:3).

autoridades, realizar estudios de sintaxe, léxico ou morfoloxía, preparar edicións críticas, etc.

Nun principio, cada investigador ou grupo de investigadores construía, cos rudimentarios medios dos que dispuña, o seu corpus particular ou colección de textos para uso propio. Como consecuencia desta disgregación, era difícil acceder ós fondos de cada investigador, os criterios de elaboración variaban grandemente segundo as fins perseguidas, e mesmo era imposible saber cantos textos existían, cales eran ou onde estaban; a situación legal de cada texto ou colección soía ser máis que dubidosa, xa que rara vez se lle pedía permiso ó propietario legal do texto, e cando se pedía, moi poucas veces se concedía (o que aínda hoxe non é nada infrecuente). A calidade e a fiabilidade dos textos canto á súa exactitude na transcrición deixaba ás veces moito que desexar, xa que tamén era frecuente que os investigadores carecesen do equipamento axeitado.

Na actualidade, sen embargo, a situación é ben distinta: a calidade está moito máis coidada e son cada vez máis numerosos os catálogos de textos electrónicos e as coleccións de uso público, en aluguer ou á venda.

Podemos atopar catálogos e listados máis ou menos exhaustivos en publicacións como *Humanities Computing Yearbook 1*[6], *Humanities Computing Yearbook 2*[7], *CETH Newsletter, Computer and Texts, Computing and the Classics, Craft, Reach: Research and*

[6] Lancashire e McCarty (eds.) (1989).
[7] Lancashire e McCarty (eds.) (1991).

Educational Applications of Computers in the Humanities, ReCALL, Arts & Humanities Search, Wilsonline, etc.

1.1.2. Inventarios, arquivos e outras fontes

No apartado anterior falamos xa dalgúns dos repertorios de textos electrónicos máis importantes. Vexámolos máis en concreto[8]:

- *ARTFL*: colección de máis de 2000 textos en lingua francesa; abrangue literatura, historia, filosofía e obras científicas publicadas entre os séculos XVII e XX. Universidade de Chicago.
- *British Library General Catalogue of Printed Books on CD-ROM*.
- *Canadian Periodical Index*: textos xornalísticos canadianos dende o ano 1988.
- *CETEDOC Library of Christian Latin Texts on CDROM*: texto completo do *Corpus Patrum Latinorum*. Universidade de Lovaina, Bélxica.
- *Cumullative Content Index*: millóns de artigos de xornais dos séculos XIX e XX.
- *Dartmouth Dante Database*: base de datos da *Divina Comedia*.
- *DIALOG*: organización que ten á venda diferentes productos como *Historical Abstracts, Artbibliographies*, etc.
- *Dissertation Abstracts International*: teses doutorais canadianas, estadounidenses e doutros países dende 1871.

[8] Para máis información, cf. o apartado 1.1.3, especialmente a nota 12.

- *English Poetry Full-Text Database*: recolle as obras de 1350 poetas ingleses dende o período anglo-saxón ata o século XIX. Chadwyck-Healey, Ltd.
- *OTA*: colección de textos en diferentes linguas, maioritariamente en inglés. Varios dos textos son de uso público (*Volpone or the Fox, The Beggar's Opera, The Bell Rope, Defense of Poesie, A Christmal Carol*) . Universidade de Oxford.
- *FRANCIS (Ficher de recherches bibliographiques automatisées sur les noveautés, la communication et l'information en sciences socials et humaines)*: inclúe textos de literatura, arte, historia, dereito, relixión, filosofía, etc. Telesystemes-Questel.
- *Le Robert Electronique*. Chadwyck-Healey, Ltd.
- *Religion Index*: textos publicados entre 1949 e 1975.
- *Thesaurus Linguae Graecae*: repertorio de textos gregos clásicos dende Homero ata o século VI d.X.C., 61 millóns de palabras. Universidade de California Irvine.
- *Thesaurus Linguae Latinae*: textos en latín clásico ata o ano 200 d.X.C. Packard Humanities Institute.
- *ARTHUS* (Archivo de Textos Hispánicos de la Universidad de Santiago de Compostela).
-O Centro de Investigacións Lingüísticas e Literarias Ramón Piñeiro está elaborando actualmente o primeiro corpus de lingua galega moderna, con textos literarios (novela e teatro) e ensaios publicados dende 1975 e xornais dende 1994.
- Colección do *University of Virginia Library Electronic Text Center*. Contén certos textos de uso público, como *The Diary of John Woolman*.

- *The Research Arquive* do Instituto Oriental da Universidade de Chicago. Textos das civilizacións de Oriente Medio: Val do Nilo, Mesopotamia, Persia, Imperio Otomano, países bíblicos.
- *The English Server* da Universidade Carnegie Mellon. Uso restrinxido.
- *Project Gutemberg* de Illinois. Uso público.
- *ETexts Archive* da Universidade de Michigan.

1.1.3. Centros de textos electrónicos

Cada vez son máis os centros de textos electrónicos (*Electronic Text Centre*), centros de informática aplicada ás humanidades (*Humanities Computing Centre*) e arquivos que se crean nas distintas universidades ou vencellados a programas de investigación concretos (como o Dartmouth Dante Project, ARTFL, etc). Estes centros son os encargados de elabora-los textos electrónicos e as coleccións e córpora correspondentes; ademais, soen ofrecer outros servicios como centros de computación e laboratorios multimedia. Algúns centros, así mesmo, ofrecen cursos e seminarios de introducción e especialización durante todo o ano, especialmente o CETH (Centre for Electronic Texts in the Humanities), que ofrece tódolos veráns dende hai catro anos un valiosísimo curso intensivo de dúas semanas titulado «Electronic Texts in the Humanities: Methods and Tools»[9].

[9] Para máis información, véxase http://cethmac.princeton.edu/CETH.

1.1.4. En Internet

Hoxe en día, e gracias ás grandes posibilidades de intercomunicación que nos ofrece Internet, podemos atopar na rede un número considerable de textos electrónicos espallados por todo o mundo. Todos eles son de uso público, aínda que o usuario se compromete moralmente a cumprir certas condicións cando os utilice, como que non vai utiliza-los textos electrónicos para facer edicións piratas dos textos orixinais, que non vai altera-lo texto, ou que se o fai seguirá unha determinadas directrices (por exemplo, explicar con claridade na cabeceira que cambios fixo quen, cando e porque) ou que sempre citará a fonte da que tomou o texto en cuestión (é dicir, incluirá sempre a cabeceira). Podemos observar algunhas destas condicións no seguinte fragmentos de cabeceira:

The Beggar's Opera, John Gay.
Transcribed, with an Introduction, Notes, and Bibliography,
by Richard Bear, University of Oregon, August 1992.
--Note on the e-text version--

Many search programs balk at the introduction of nonstandard or formatting characters into the text. For the convenience of those who wish to conduct searches, there is no markup provided with this text, nor even any indication of italics within the play itself. Italics (or underlines) in the introduction, notes, and bibliography are indicated with an _underline_ character before and after the affected material. If you wish, you may substitute your own formatting characters for these as desired, or remove them. Endnotes are indicated within the text by numbers within braces: { }. As these may interfere with searches, you may wish to remove them, or, if you wish to print out a formatted copy, replace them with superscripts. No pagination or white space has been provided with this edition, and you may wish to go through it with your word processor before printing out. You may, if you wish, remove the front and back matter or substitute your own for that provided. The transcriber knows of no copyright laws that have been violated in producing this edition; _The Beggar's Opera_ was printed originally in 1728, and this electronic text follows that of the printed text of 1765. It is placed in the public domain to aid in scholarship and teaching as well as reading enjoyment, provided always that the name of the transcriber remains associated with any copies distributed.

É posible tamén atopar fragmentos de grandes córpora que os autores poñen en Internet temporalmente ó alcance de tódolos usuarios para facer probas de funcionamento. Así, é ou foi posible nalgún momento, atopar fragmentos de tamaño considerable da base de datos COBUILD, do Lund Corpus, do OTA, e outros córpora como o Corpus de Xornais Holandeses.

1.2. Creación do texto electrónico

1.2.1. Teclado vs. escáner óptico

Se o investigador precisa construí-lo seu propio texto electrónico ou o seu propio corpus ou colección, ten dúas alternativas: o teclado ou o escáner. O uso do teclado non é unha opción tan primitiva e inusual como puidese parecer. Implica ter que transcribir (mecanografar) o texto ou textos completos, dende logo, pero hai ocasións nas que non queda outra solución. Pensemos, por exemplo, na construcción de córpora electrónicos de textos orais. Aínda que existen sintetizadores que transcriben a fala, soen estar preparados exclusivamente para o inglés, e normalmente só para a fala estándar; moi pouco hai feito para o castelán e nada para o galego, aínda que se está a traballar neste campo. Outro caso no que a única vía posible é o teclado é cando non contamos cun texto base axeitado para o escáner: textos antigos que só se conservan en manuscritos ou edicións impresas moi antigas (que ou ben non son axeitadas para explorar ou ben poderían sufrir no proceso) ou textos modernos pero escritos a man. Existen, como veremos máis adiante, programas de recoñecemento da escrita «adestrables»[10] como o *Newton*, pero non son completamente fiables.

Unha vez mecanografado o texto[11], o seguinte paso é a corrección, ben mediante correctores ortográficos ou ben visualmente. Debemos ter en conta que, se ben os correctores ortográficos soen ser «adestrables» (é dicir, admiten a posibilidade de almacenar na memoria palabras novas que non contiñan), teñen o inconveniente de que non distinguen homónimos, polo que se no texto aparece «Nacía, ca mañán, a terra de Ávila. lus extraña aos sacos infindos, xiados», no canto de «Nacía, ca mañán, a terra de Ávila. Lus extraña aos sucos infindos, xiados»[12], o corrector ortográfico non correxirá «sacos». A maioría dos correctores si correxirán «lus» con minúscula despois de punto, pero non aceptarán esta palabra como válida porque non recoñecen os dialectalismos. Así mesmo teremos problemas en galego, por exemplo, coas segundas formas do artigo, xa que «escribi-lo libro» conta como tres palabras, das que dúas («escribi» e «lo» non existen en galego). A corrección visual, xa que logo, será sempre precisa.

É moi aconsellable que nesta fase de corrección traballen alomenos dúas persoas para asegurármonos de que o texto transcrito queda completamente libre de erros ortográficos. Ademais,

[10] É dicir, que poden almacenar na memoria as características de determinadas escrituras.
[11] No apartado 2 ocuparémonos de cales son as partes do texto que nos interesa conservar (e polo tanto mecanografar) para os nosos estudios.
[12] Otero Pedrayo (1980).

probablemente será preciso que cada texto se corrixa, alomenos, dúas veces.

1.2.1.1. O escáner. Imaxes en formato electrónico

Un escáner (do inglés *scanner*, derivado de *scan* facer un varrido, escrutar, escudriñar) é un aparato electrónico que transforma unha imaxe impresa nunha imaxe electrónica mediante un varrido con feixes de luz paralelos. Debemos ter moi presente que o resultado que obtemos logo de utilizar un escáner non é un texto electrónico, senón unha imaxe dese texto; é dicir, non obtemos unha cadea de caracteres cos que podemos traballar, senón unha fotografía almacenada en soporte magnético. Como veremos no apartado 1.2.1.2, este é un paso previo necesario antes de utiliza-lo OCR.

Pero ¿ten algún sentido conserva-las imaxes dos textos? Deixando á parte a utilidade que ten para ilustrar calquera tipo de investigación, o almacenamento de imaxes en formato electrónico utilízase, en relación co tema que nos ocupa para dúas fins fundamentais:

- a conservación de documentos en perigo de desintegración, que consiste, por unha parte, na transcrición do texto e, pola outra, no almacenamento da imaxe do texto, para conservar tamén unha representación do seu aspecto físico. Pensemos, por exemplo, en antigos manuscritos, papiros, etc. Aínda que non estean en perigo de desingegración, a creación destas imaxes multiplicaría as posibilidades de calquera persoa de traballar cos «orixinais» directamente: para propor novas lecturas, facer estudios caligráficos, etc.

- a inclusión, como apéndice do texto electrónico, dunha serie de arquivos que conteñan as imaxes de partes importantes do texto, como ilustracións, emblemas, etc. e incluso o texto enteiro, co cal podemos ter acceso, ó mesmo tempo, ó texto propiamente dito e ó seu aspecto físico.

O almacenamento destas imaxes presenta, sen embargo unha serie de problemas:

- manipulación: as imaxes ocupan moitísima memoria, co cal a velocidade do procesador diminúe e diminúe tamén a velocidade do vídeo. Traballar con moitas imaxes supón precisarmos unha grande cantidade de memoria RAM.
- como consecuencia do seu tamaño, precísase tamén moito espacio no disco duro para almacenalas.
- a reproducción das imaxes na pantalla implica sempre unha certa perda de nitidez da imaxe, mesmo nas grandes estacións de traballo. A degradación da imaxe ó imprimila soe ser aínda maior.
- hai tamén un inconveniente para o intercambio de imaxes, xa que existen numerosos formatos distintos dependendo dos progrmas e das plataformas nas que foron creadas.

1.2.1.1.2. Texto impreso: paso previo ó OCR

A *exploración*[13] (ou *scanning*) dun documento é, como xa vimos, o paso previo necesario antes da aplicación do OCR. Consiste en sacar unha «foto» do texto, para o cal debémolo preparar, sempre en

[13] Termo adoptado de Gómez Guinovart e Lorenzo Suárez (1994).

función dos nosos obxectivos, tal e como explicaremos na sección 1.2.2.

1.2.1.2. OCR: interpretación de imaxes e conversión en texto

1.2.1.2.1. Tipos básicos: automático e semiautomático

Os programas de OCR (*Optical Character Recognition*) ou Recoñecemento Óptico de Caracteres son sistemas capaces de enfrontarse a unha imaxe electrónica dun texto, descifrala (é dicir, «lela») e transformala en texto ASCII. Con isto a imaxe orixinal non se perde, senón que, ó final do proceso, obtemos dous arquivos: un coa imaxe e outro co texto co que xa podemos traballar.

O programa parte duns datos almacenados na memoria, confróntaos con distintos fragmentos da imaxe facendo un varrido de esquerda a dereita e de arriba a abaixo, e identifica os distintos caracteres.

Hai dous tipos de OCRs: automático e semiautomático. Os programas semiautomáticos admiten «adestramento» é dicir, cando atopan un caracter que non recoñecen, o usuario pode especificar de que caracter se trata e o programa gárdao na memoria; cando o volva a atopar non terá problemas para recoñecelo. Os programas automáticos non admiten este tipo de intervencións por parte do usuario, senón que exploran a imaxe ininterrumpidamente; cando atopan algo que non recoñecen escollen o caracter ou caracteres máis semellantes. Tanto uns como outros soen admitir unha especificación de lingua: podemos elixir traballar co alfabeto galego, inglés, francés, etc, co cal recoñecerá ou non os distintos acentos, letras, etc. Ademais, certos OCRs recoñecen ou non combinacións de letras segundo a lingua que se lles especifique. Por exemplo, se estamos a traballar co alfabeto galego (basicamente o mesmo que o castelán), non recoñecerá grupos como «-ck-», «-chr-», «-ng», etc.

1.2.1.2.2. A fiabilidade do OCR

A fiabilidade do OCR é inversamente proporcional ó número de erros por páxina. Ningún programa de recoñecemento de caracteres é efectivo ó 100%. Mesmo os mellores dos «adestrables» presentan fallos, aínda que a porcentaxe de erros diminúe logo das primeiras páxinas.

Xunto coa fiabilidade do programa, unha segunda cracterística importante que debe ter o OCR que elixamos é a velocidade. O proceso conxunto de exploración co escáner e de interpretación co OCR non debería levar máis de un minuto por páxina. A diferencia entre un minuto e un minuto e medio pode parecer agora insignificante, pero calculémo-la diferencia se queremos introducir no ordenador 10.000 páxinas de texto: máis de 83 horas de traballo, é dicir, máis de dez xornadas de oito horas (pero é imposible, e pouco recomendable, estar explorando textos oito horas seguidas tódolos días, polo que as 83 horas se poden convertir nun mes de traballo).

1.2.1.2.3. Os problemas máis comúns do OCR

A preparación dun texto impreso previa á exploración mediante escáner é un proceso tan importante como a exploración mesma. Debemos procurar, en primeiro lugar, que o texto de partida teña a máxima calidade posible: papel branco e sen manchas, caracteres nítidos e perfectamente debuxados, cor da tinta uniforme, follas enteiras e sen enrugas, etc. Se temos que recorrer a edicións antigas por non dispormos doutras, o mellor será sacarlle unha fotocopia ó texto cunha máquina de alta resolución que limpe as manchas do texto orixinal. Por outra parte, coa fotocopia evitamos posibles estragos ó desencadernármo-lo libro.

Hai certos problemas, non embargantes, que son difíciles de solucionar:

- caracteres que se tocan: nas maiúsculas das edicións antigas é doado atopar dous ou máis caracteres que están en contacto, o que interfire na aplicación do OCR. É un problema imposible de evitar e que require unha corrección posterior a man.
- caracteres incompletos: provocados por defectos na impresión do texto orixinal. Moi frecuentes en edicións antigas. Tampouco teñen solución.
- variación excesiva da liña base (a liña imaxinaria sobre a que se imprime unha liña de texto). É tamén unha chata das edicións antigas.
- inconsistencia excesiva na forma dos caracteres: como consecuencia de defectos de impresión, da calidade do papel ou do desgaste dos tipos (no caso de imprentas de chumbo).

- variedade excesiva de tipos de letra e existencia de letras iluminadas, que son irrecoñecibles para o OCR.
- letras capitais que ocupan máis de unha liña: o OCR non sabe como interpretalas.
- transparencias dunha páxina a outra, o que provoca que a imaxe electrónica non sexa nítida.
- decoloracións das follas: pódese evitar tamén coas fotocopias.
- anotacións a man ou debuxos nas páxinas do texto.
- en xeral calquera outro ruído visual que impida unha perfecta visión do texto.

1.2.1.4. A escrita: sistemas «adestrables»

Nos casos nos que faga falla explorar un texto escrito a man, podemos contar coa axuda de programas especializados en recoñecer escrita manual que ademais son «adestrables», como por exemplo o *Newton*. Así e todo, o texto orixinal debe estar moi limpo, o trazado das letras debe ser moi consistente, etc., o que fai que estes OCRs sexan bastante lentos e pouco fiables. Nestes casos, normalmente temos que recorrer ó mecanografado dos textos.

1.2.2. Proceso de creación do texto electrónico

1.2.2.1. Finalidade do arquivo: privado ou de uso público

Se queremos construír un texto electrónico ou unha colección para uso privado é probable que non teñamos traba ningunha dende ningún punto de vista. Poderemos utiliza-las convencións que prefiramos e incluso inventarmos as nosas propias; non teremos que pedir permisos legais, nin sequera no caso de que publiquémo-los resultados das nosas investigacións; poderemos prescindir de certos detalles imprescindibles se fixesemos un texto de uso público ou para a venda: a cabeceira, portadas, índices, prólogos etc (no caso de que nos interese só o texto propiamente dito).

1.2.2.2. Elección da edición de partida: cuestións legais e prácticas

Cando construímos unha colección de textos electrónicos da que vai poder dispor o público en xeral (xa sexa que a poñamos á venda, que a aluguemos ou que simplemente permitamos acceso ós fondos) estaremos obrigados a elixi-los textos que a formarán en función das nosas posibilidades legais e económicas. É cada vez menos frecuente que os propietarios dos dereitos de autor se neguen a colaborar coas institucións que elaboran textos electrónicos[14], mais debemos sempre pedir permiso para utilizalos, e non será infrecuente que teñamos que pagar a cambio. Se ben hai certos textos que semellan non ter propietarios, ben porque sexan de autor anónimo, ben porque sexan moi antigos, debemos ter en conta que o seu beneficiario será sempre a Sociedade Xeral de Autores, polo que tamén neses casos cumprirá pedir permiso. Nin que dicir ten que ese permiso terá que ser, en tódolos casos, por escrito, con data e sinatura.

[14] A situación cambia cando se trata de investigadores particulares.

Hai tamén certas cuestións prácticas que haberá que considerar na nosa elección da edición de partida:

- custo da edición impresa, xa que é posible que o libro se deteriore.

- facilidade para utiliza-lo libro no escáner: se o volume é demasiado groso e non se pode abrir ben, a imaxe será defectuosa.

- calidade da impresión, como xa vimos no apartado 1.2.1.2.3.

- posibilidade de desencaderna-lo volume para exploralo e de volver a encadernalo despois. A desencadernación é, no caso de utilizar edicións de pouco valor, unha alternativa fronte á fotocopia: é máis barata, máis rápida e evitaranos posibles problemas cos dereitos de copia.

Deberemos sopesar todos estes factores e decidirnos pola relación calidade-precio que máis nos conveña.

1.2.2.3. Creación do texto electrónico

Unha vez escollida a edición de partida, o proceso de creación dun texto electrónico comeza pola preparación dos orixinais. Se eliximos utilizar fotocopias, deberemos utilizar unha fotocopiadora de alta resolución para evitar todo tipo de ruídos visuais. Cada folla do texto orixinal debe ocupar unha fotocopia, sexa cal sexa o tamaño, para facilita-lo proceso do OCR (que fai varridos horizontais).

Se decidimos traballar co texto impreso directamente, temos dúas opcións: ou ben se trata dun libro que se abre perfectamente (co cal se soen estropear) e se pode adaptar á superficie do escáner ou ben o desencadernamos. A desencadernación consiste no descosido dos cadernos que compoñen o libro ou no guillotinado do libro polo lombo; o corte deberá ser o mínimo posible para evitar que as marxes das follas se fagan excesivamente pequenas, o que dificultaría a posterior encadernación.

Unha vez preparadas as follas, explóranse no escáner que, se ten alimentador, alixeirará o traballo. É aconsellable facer primeiro unhas probas cunhas cantas follas para asegurarnos de que temos correctamente especificada a resolución, tamaño, etc. Para comeza-lo traballo con certa seguridade, é igualmente conveniente que realicemos unhas probas anres de empeza-la fase de interpretación co OCR.

Ó igual que cos textos que introducimos no ordenador mediante o teclado, a corrección manual, a ser posible con varios correctores, é imprescindible se queremos te-la seguridade de que o texto resultante está libre de erros nunha porcentaxe razoable. Na meirande parte dos casos, a limpeza dos textos inclúe a supresión de números de páxina, notas ó pé, liñas en branco e, en xeral, calquera outro atributo do texto impreso orixinal que non nos interese.

Cada imaxe que nos proporciona o escáner está almacenada nun arquivo; cada vez que o OCR se enfronta a un destes arquivos crea outro co texto que interpreta. Deste xeito, un libro de duascentas páxinas proporcionaríanos catrocentos arquivos[15] que finalmente teremos que «unir» na orde correcta para remata-lo texto electrónico;

[15] Por iso o máis sensato é numera-los arquivos no canto de lles pór nomes.

esta é unha operación que se pode facer con calquera procesador de texto.

1.3. Textos limpos vs. textos etiquetados

Rematado o proceso descrito anteriormente temos listo xa o noso texto electrónico, que consiste, como xa vimos, nunha sucesión de caracteres ASCII, tamén chamado *texto puro* ou *só texto*.

Se o noso interese final consiste en facer un simple índice de frecuencias, o traballo de preparación do texto rematou aquí. Podémolo utilizar como arquivo de entrada do programa que consideremos máis axeitado (vid. sección 3). A maior parte dos repertorios de textos electrónicos, non embargantes, utilízanse para máis fins: facer concordancias, establecer densidades léxicas, estudios estilísticos, etc. Por este motivo soen incorporar, á parte do texto propiamente dito, unha serie de caracteres extra ou marcas (vid. sección 2.5) que identifican as divisións do texto en libros, volumes, capítulos, parágrafos e/ou oracións; as diferentes partes da oración ou clases de palabras; as personaxes que falan nunha obra de teatro; acotacións; estrofas e versos; e en xeral calquera información que os autores do texto electrónico coiden de interese para calquera investigación. Estes caracteres, tamén denominados *marcas* ou *etiquetas*, constitúen o que se chama unha *linguaxe de marcación*. Existen varias destas linguaxes e algunhas son moi elaboradas, o que nos permite marcar, como xa veremos máis adiante, dende os menores rasgos fónicos ou prosódicos ata córpora enteiros.

A maioría dos textos electrónicos que podemos atopar dispoñibles na rede ou nos centros de textos electrónicos están marcados con algunha desas linguaxes, principalmente SGML.

Se ben a inclusión destas marcas representa un grande gaño na cantidade de información codificada no texto electrónico, agacha tamén un inconveniente: dependendo da linguaxe utilizada verémonos na obriga de utilizar certos programas de análise e non outros, co cal a eficacia e o custo económico e en tempo do noso proxecto poden variar grandemente.

2. Preparación do texto electrónico. Marcación do texto.

Neste apartado revisaremos tódolos pasos previos á utilización dos diferentes programas de análise lingüística. Partiremos do ficheiro de texto ASCII e incluiremos toda a información precisa para chegarmos ó texto electrónico que mellor se adapte ás nosas posibilidades e ás nosas fins.

2.1. Problemas que pode presentar o texto

Segundo cal sexa a fin para a que utilicémo-lo texto electrónico, deberemos ter en conta certas características do mesmo que, segundo os casos, poderemos modificar ou eliminar segundo a nosa conveniencia. Se o texto electrónico é para uso exclusivamente persoal, non haberá problema ningún en prescindir de determinada información, mais se estamos a deseñar un corpus de uso máis ou menos público non poderemos omitir nada, xa que debemos procurar que os textos lles sexan útiles á maior cantidade de usuarios posible.

Ó longo dos apartados 2.1.1 a 2.1.18 referireime a «programas (de análise)» sen especificar máis. Na sección 3 veremos con detalle e en funcionamento as solucións propostas para algúns dos problemas.

2.1.1. Puntuación

En xeral, é conveniente conserva-la puntuación do texto orixinal, xa que nos serve para delimitar oracións, pero, se cadra, hai casos nos que debemos eliminar ou modificar certos usos. Á parte da separación de oracións, o punto serve tamén para indicar unidades de millar e, por veces, millóns, nos números; para marcar certas abreviaturas ou siglas; para separa-los datos en certos formatos de datas (01.06.95), etc. Estes usos poderían provocar interferencias nas nosas buscas de oracións completas con ordes do tipo «localiza-lo segmento comprendido entre dous puntos».

Se tomamos como punto de referencia para as nosas búsquedas as comas, compre termos en conta que tamén se utiliza ese signo de puntuación nos números decimais[16], polo que é probable que tamén poidan presentar problemas.

---

[16] Se estamos a analizar un texto en inglés, debemos lembrar que o uso da coma e do punto para as unidades de millar e para os decimais é inverso ó noso.

Unha das solucións pode ser prescindir de tódolos números, se non teñen importancia para a nosa investigación; ou elimina-los signos de puntuación nos números e nas abreviaturas.

2.1.2. A segunda forma do artigo

Dependendo de como teñamos configurado o noso programa de análise, as segundas formas do artigo pódennos crear problemas, xa que unha cadea como «recibi-la carta» podería ser interpretada como tres palabras no canto de dúas, co que as cadeas «la, las, lo, los» se convertirían en palabras galegas que se utilizan cunha frecuencia relativamente alta. O problema está no trazo que separa o artigo do verbo, xa que normalmente funciona como separador. Unha posible solución é enganar ó programa de análise e especificar nas ordes que o trazo é un diacrítico, como se fose un acento. Isto ten como desvantaxe que, se o texto posúe palabras separadas entre dúas liñas, tamén nese caso o trazo será analizado como un diacrítco. Vexamos unha posible solución no apartado 2.1.3.

2.1.3. Trazos.

O problema que quedaba aberto no apartado anterior soluciónase, en parte, se modificamos parcialmente a apariencia do texto de xeito que non haxa palabras divididas entre dúas liñas; esas palabras pasarían, por exemplo, á primeira liña.

Hai, non embargantes, outro uso do trazo que nos presentará un problema de máis difícil solución: o trazo como separador de compostos. ¿Que facemos con palabras como «luso-galaica»? A barra ou a contrabarra serían posibles substitutos do trazo, sempre que non se teñan xa utilizado no texto con outra fin. Unha segunda posibilidade sería prescindir de marca-los compostos; pero ¿e se queremos estudialos? Non queda máis remedio que buscar sempre unha solución intermedia.

2.1.4. Citas noutras linguas

Se estamos a analizar un texto con abundantes citas noutras linguas, sería útil identificar cada cita con cadansúa lingua ou, se cadra, eliminalas na análise para que non inflúan nas frecuencias de uso, estatísticas de lonxitude de palabras, etc. Podemos dicir que a supresión sería a mellor opción no caso de estarmos estudiando o estilo dun determinado autor, xa que as citas non responderían en absoluto a ese estilo.

2.1.5. Homónimos

Unha das desvantaxes da maioría dos programas de análise é que non contan cunha utilidade de lematización, o que lles impide diferenciar homónimos se non están marcados como tal polo usuario. Polo tanto, se queremos un listado das formas do verbo comer e pedímolo cunha orde do tipo «buscar tódalas cadeas do tipo com*», obteremos, á parte das formas verbais, «como» conxunción, «coma» substantivo (estado clínico ou signo de puntuación). Sería útil, polo tanto, un certo grao de diferenciación por medio dalgún tipo de marcación.

2.1.6. Variantes ortográficas ou de realización

Traballando con textos dialectais ou orais, decatarémonos das grandes dificultades que nos presenta determinado léxico coas súas respectivas variantes. Se queremos conservar un texto o máis próximo posible á fala real, non poderemos corrixir estas variantes, pero teremos moitos problemas para, por exemplo, establecer frecuencias de uso de determinadas palabras se no texto ou no corpus hai variantes. Busca-las frecuencias da palabra «ganancia» nun texto onde apareza realizada como «ganancia, ghanancia, ghanansia, ghañansia, ganansia, etc» multiplicaría o traballo se non contamos con algún tipo de marcación: por exemplo, que conserve a palabra tal e como foi pronunciada e que indique ó lado a forma normativa, na cal se basearían as buscas.

Un problema semellante preséntao a segunda forma do artigo: se queremos buscar con que frecuencia se usa, teremos que buscar tanto «a, as, o, os» como «-la, -las, -lo, -los».

2.1.7. Notas: ó pé, ó final, do autor, do editor, etc.

As notas ó pé, ó final, do autor, etc., que nunha edición impresa aparecen fóra do corpo do texto, no texto electrónico non poden aparecer no mesmo lugar, xa que, como vimos, carece de formato. Deberemos decidir se nos interesa conservalas (p. ex. as notas do autor) ou non (probablemente as notas do traductor, editor, referencias bibliográficas, etc, xa que non pertencen ó texto en si). Se as conservamos podemos situalas no medio do texto, introducindo unha marca que signifique «isto é unha nota ó pé/ó final», ou ó final, para o que teremos que utilizar unha chamada:

> ... utilizaba versos leoninos <nota13: esta clase de versos era moi común na Idade Media, e aparece nas inscricións das cruces da Catedral de Santiago. Fin nota> en tódalas súas composicións.

2.1.8. *Marginalia*

Se a edición que utilizamos como base do noso texto electrónico é unha edición crítica comentada, é probable que presente anotacións marxinais. Unhas veces conteñen resumes dos parágrafos ós que están engadidas, outras veces ofrecen lecturas distintas, comentarios, etc. Como non son nunca parte integrante do texto, é probable que non nos interesen e que poidamos prescindir delas. Se as queremos conservar, teremos que facer como coas notas ó pé, etc das que xa falamos no apartado 2.1.7.

2.1.9. Diferentes lecturas

Outras edicións críticas presentan as lecturas diferentes na mesma liña do texto, o cal dificulta a limpeza se non as queremos conservar:

> Nam eum alii Lunae [Lini] filium, alii Orphei volunt.

2.1.10. Listados

Normalmente tódolos elementos dun listado poderán ir a liña seguida, co cal aforraremos espacio e interferencias entre a división en liñas e en parágrafos. Doutro xeito, por exemplo, a media de palabras por parágrafo diminuiría notablemente.

2.1.11. Números: arábigos e romanos

Os números arábigos non soen dar problemas (á parte dos vistos no apartado 2.1.1.) á hora de analizar un texto, pero os romanos si, xa que o programa os pode confundir con letras: VI = vi, CLI = cli, MI = mi, MIL = mil, DIX = dix, etc., o que daría lugar a combinacións de letras inesperadas segundo que linguas. Novamente, podemos optar por eliminalos ou marcalos.

2.1.12. Datas

Convén conservalas se estamos a realizar unha investigación cronolóxica para data-las aparicicións de determinadas palabras e a súa evolución no tempo. Para iso habería que transforma-las datas en marcas ou ben xerar marcas a partir das datas. Podémonos atopar co problema de que haxa máis de un sistema de calendario no mesmo texto: en textos en latín podemos atopa-la data da era hispánica e mais da cristiana.

2.1.13. Nomes: persoais, de lugar, de organizacións, etc.

Cando analizamos un corpus de textos de xornais, descubrimos que unha altísima porcentaxe das palabras utilizadas son nomes. Isto débese a que no estilo de redactar noticias se lle soe dar importancia ós protagonistas (persoas, pero tamén institucións, organizacións, etc) e ós lugares onde se desenvolven os feitos. Se cadra, para determinadas análises sería aconsellable prescindir deste tipo de textos, ou, cando menos, dos nomes que conteñan, xa que senón poderiamos chegar a conclusión de que palabras como Santiago, A Coruña, etc. ou os nomes de determinados políticos son as palabras máis frecuentes do galego.

2.1.14. Portada, prólogo, dedicatorias, citas, etc.

O texto dun libro non sempre comeza onde di, por exemplo, «Capítulo I». Moitas veces comeza mesmo na portada, onde o autor pode incluír (sobre todo na literatura dos séculos pasados) citas, dedicatorias ou aclaracións sobre o texto. Por este motivo, é preciso, moitas veces, incluír esa información dentro do texto electrónico codificada axeitadamente: normalmente se prescindirá de debuxos e adobíos, pero poderíanse conservar en imaxes asociadas. Debemos decidir tamén se conservamos un prólogo que non é do autor do libro, as citas doutros autores, etc.

2.1.15. Epílogo, apéndices, índices, contraportada, etc.

Do mesmo xeito que un libro se pode estender cara adiante, pode facelo cara atrás: epílogo e apéndices soen ser partes integrantes dun texto. Non así os índices (que ás veces non son máis que listados) e a contraportada. Haberá que decidir que facer con eles.

2.1.16. Tipografía e formato: ¿son importantes?

Xa vimos no apartado 1.1.1 que os textos electrónicos non adoitan conservar formato ningún, pero ¿que sucede cando son características relevantes do texto? Pensemos, por exemplo, nos caligramas e noutras disposicións estratéxicas dun texto nunha páxina. O máis axeitado sería facer por un lado un texto electrónico sen formato ningún, sen debuxos e sen efectos e, por outro lado, asociarlle unha imaxes do texto impreso con tódolos seus efectos. Deste xeito poderiamos analiza-lo texto sen problemas e tamén recupera-lo seu aspecto orixinal cando o precisemos.

2.1.17. Textos falados

Se traballamos con textos falados é moi probable que nos interese codificar dalgún xeito quen está a falar en cada momento, en que situación e, en xeral, calquera circunstancia que poida afectar á comunicación: ruídos, interrupcións, etc. Podemos tamén codificar información sobre a prosodia, do tipo «entoación ascendente» etc.

2.1.18. Acotacións

Un tipo especial de textos falados son as obras de teatro. Á parte dos diálogos, as obras de teatro conteñen outra información en forma de acotacións do autor. Se o que nos interesa é simplemente a fala e non o estilo do autor, teremos que prescindir desa información ou marcala dalgún xeito para que o programa de análise non a teña en conta.

2.2. Estructura do texto

Neste apartado trataremos de conseguir representar no texto electrónico a estructura do texto orixinal, de xeito que a poidamos explotar cos programas de análise.

2.2.1. Estructura xerárquica simple

Esta é, se cadra, a estructura máis doada de representar. Básicamente responde a unha ordenación semellante á seguinte:

```
Texto.
        Capítulo.
                Sección.
                        Apartado.
                                Parágrafo.
                                        Oración.
```

Para utilizar un texto con esta estructura no OCP, por exemplo, non teremos máis que utilizar, ó comezo de cada unidade, marcas do estilo de:

```
<T Historia de Galicia>
        <C Historia Antiga>
                <S Galicia Primitiva>
                        <A Desde o 2.416 ó 2332 a.X.C.>
                                <P 1>
                                        <O 1>
```

onde as letras maiúsculas representan cada un dos apartados.

2.2.2. Estructura aniñada

A estructura pode complicarse cando se presentan unidades recurrentes, é dicir, unidades que poden aparecer, directamente, dentro de máis de unha unidade maior. Isto crea o que se chama unha estructura aniñada. Un exemplo podería ser:

```
<T Historia de Galicia>
    <C Historia Antiga>
        <P Introducción á Historia Antiga>
        <S Galicia Primitiva>
            <P Introducción á Galicia Primitiva>
            <A Dende o 2.416 ó 2332 a.X.C.>
                <P 1>
                    <O 1>
```

onde existen unidades parágrafo en varios niveis. O capítulo de Historia antiga, por exemplo, compónse de varias seccións que van precedidas dun ou máis parágrafos introductorios pero non constitúen unha sección en si mesmos. Para lograr unha descrición exacta desta estructura teremos que recorrer a outro sistema de marcación distinto ó utilizado anteriormente, por exemplo a SGML[17] (vid. o apartado 2.4.3). Precisariamos definir unha estructura do tipo:

[17] Sobre a SGML e as directrices da TEI, vid. o capítulo de Fernando Magán neste mesmo volume.

```
<!ELEMENT texto          - o (capítulo+)-->
<!ELEMENT capítulo       - o (parágafo?, sección+)-->
<!ELEMENT sección        - o (parágrafo?, apartado+)-->
etc.
```

que básicamente quere dicir que os capítulos poden conter, antes das seccións, parágrafos de texto. Este tipo de marcación leva incluído un sistema de numeración das diferentes unidades, polo que non temos que escribi-los números correspondentes.

2.2.3. Numeración de liñas

É moi probable que se analizamos textos poéticos utilicémo-los versos como referencia, polo que convén que as liñas estean numeradas. Con isto chegamos a unha terceira representación das estructuras xerárquicas. Programas como o OCP numeran automaticamente as liñas, pero non os versos; por iso, cando os versos ocupan máis dunha liña (por exemplo, nas edicións críticas con varias lecturas que nos interesa conservar) é conveniente numera-los versos e prescindir da numeración das liñas. Podemos utilizar un sistema semellante ó visto no apartado 2.2.1, como nos seguintes versos[18]:

<V 1>Anque ora é a hora dos vagueiros corazóns que non ousan
<V 2>e estás na banda das ribas altas e os mármores de olvido,
<V 3>vir veu o tempo de que xa falen as herbas mudas e as pedras
<V 4>e nos cons bate a ola salgada dunhas palabras verdadeiras.

[18] Iglesia Alvariño (1961).

Existen outros sistemas semellantes que veremos no apartado 2.4.2.

2.3. Documentación do texto electrónico, do orixinal e da codificación

Como xa vimos na sección 1, un texto electrónico debe incorporar, á parte do texto propiamente dito, unha serie de informacións relativas a certos aspectos externos. Toda esta información, que desglosaremos nos apartados 2.3.1 a 2.3.4 pode ser codificada de varios xeitos, aínda que o máis completo e o que se está a facer cada vez máis común, é o que marcan as directrices da TEI, como veremos no apartado 2.3.5.

2.3.1. Información bibliográfica sobre o texto orixinal

Compre especifica-la maior cantidade posible de datos sobre a edición impresa orixinal para facilitarlles ós investigadores ó acceso á mesma no caso de que queiran, por exemplo, comproba-la fidelidade do texto electrónico. Estes datos serán os mesmos, basicamente, que se inclúen nas fichas bibliográficas dos libros: autor(es), editor(es), título, subtítulo(s), editorial, ano de edición, reimpresión(s) se as houbese, etc. Incluirase tamén o ISBN e calquera outra identificación que posúa o libro. Ás veces é tamén conveniente especificar se o texto electrónico se fixo a partir do texto orixinal ou de fotocopias, xa que se pretende dar unha información o máis exhaustiva posible.

2.3.2. Información sobre o texto electrónico

Incluiremos a data de comezo e remate de creación, os autores dos diferentes procesos (exploración, corrección e limpeza, marcación, etc), institución ou organismo do que dependen, motivos da creación, se pertence a un corpus, descrición do proxecto, o director do proxecto ou outra persoa responsable, o tamaño do ficheiro resultante, etc.

2.3.3. Principios de codificación

Esta é a información máis importante dende o punto de vista do investigador que vaia utiliza-lo texto unha vez creado, xa que recolle os principios e directrices seguidos na exploración do texto, na súa corrección e limpeza e na marcación. Incluiranse datos sobre que partes do texto foron omitidas e por que (números de páxina, ilustracións, notas, segundas lecturas, etc) e como afecta iso ó resultado xeral. As directrices de marcación deberán especificar o número de marcas ou etiquetas utilizadas, a súa ubicación e significado, así como unha descrición da súa xerarquía. Estes datos son necesarios para que o investigador se poida mover polo texto con soltura e sabendo en cada momento que pode buscar e que vai atopar.

Aínda que non tódolos textos electrónicos recollen a totalidade destes datos, a tendencia é incluílos todos. Vexamos un exemplo de *Volpone*:

```
<author Ben Jonson> <work Volpone> <edition Brooke and Paradise>
<<scanned 12/92; preliminary edit by Willard McCarty 12/92>>
<<proofread & corrected by Karen Arthur 2/93>>
```

<<line-numbers added by W.M. 2/93>>
<<additional speaker label added, minor corrections, WM, 5/93>>
<<speaker `label' removed, speaker tag replaced by TACT label, WM 7/93>> <act -> <scene -> <actscene -.-> <L ->-|poet|-

2.3.5. A cabeceira TEI

A información máis completa é a que se codifica na chamada *cabeceira TEI* (*Text Encoding Initiative*), que nas súas directrices especifica o *maximum* desexable de datos. No Apéndice 1 temos un exemplo de tódolos datos que se poden incluír, aínda que non sempre é preciso ou posible cubrilos todos. No seguinte exemplo podemos ve-la cabeceira TEI dunha colección de contos de Bioy Casares (*El lado de la sombra*), creada pola Asociación Arxentina Corpus:

<TEI. O> <!_ _ Encabezador TEI de texto> <encabezador TEI tipo texto 11> <!_ _ Descripción.archivo> <descripción. archivo> <declaración. título> <título>El lado de la sombra. Edición electrónica</título> </declaración. título> <declaración. responsabilidad> <autor. texto> Bioy Casares, Adolfo </autor. texto> <creador. archivo> Asociación Argentina Corpus </creador. archivo> <patrocinante> Sociedad Estatal Quinto Centenario </patrocinante> </declaración. responsabilidad> <declaración. edición> <edición> primera informatizada </edición> <responsabilidad. edición>Segret y asoc. </responsabilidad.edición> </declaración. edición> <declaración. extensión> Cuarenta y siete mil </declaración.extensión> <declaración. publicación> <editor> Asociación Argentina Corpus </editor> <fecha. creación> Abril 1992 </fecha. creación> </declaración. publicación> <descripción. fuente> <autor>Bioy Casares, Adolfo</autor> <editor>Tusquets Editores. Colección Andanzas</editor> <título>El lado de la sombra</título> <impresión> <ciudad> Buenos Aires </ciudad> <fecha>Noviembre 1991</fecha> <nota especial>ISBN 950-9779-21-0</nota especial> </impresión> </descripción. fuente> </descripción. archivo> <!_ _ Declaraciones.codificación> <declaraciones. codificación> <objetivo> almacenar un texto literario.</objetivo> <principios. edición> <corrección>No se transcribe el Indice inicial ni la lista último títulos del final</corrección>

</principios. edición> <sistema. referencia>se han marcado con membretes libro, solapa, cuento, acápite, dedicatoria, contratapa, párrafo y corte de página.</sistema referencia> <niveles.codificación> se han marcado con membretes título, nombre propio de persona y lugar, cápite, dedicatoria, firma, fuente cita, destacado (comillas, negrita, bastardilla), abreviaturas, palabras extranjeras. </niveles.codificación> </declaraciones. codificación> </encabezador TEI tipo texto 11>

2.4 Marcación do texto segundo os nosos obxectivos

Ó igual que outros dos aspectos dos textos electrónicos vistos nos apartados anteriores, a marcación dun texto dependerá fundamentalmente de se o seu uso será persoal ou público. No caso de construírmos un texto para uso doutros investigadores, deberemos utilizar unha marcación estándar que sexa doada de entender ou, que posúa, cando menos, documentación abonda para estudiala. Se só o imos utilizar nós, as marcas poden ser completamente persoais.

2.4.1. Marcas persoais para traballo privado

Despois de termos decidido claramente o que queremos marcar, deberemos facer unha relación das marcas que utilizaremos. En principio, as marcas deberían ser caracteres que non aparecesen no texto, para evitar erros nas buscas; por exemplo, non deberiamos marca-los substantivos como «s» por que se facemos unha busca ou un reconto de tódalas «s» estaremos contanto tanto a etiqueta como a letra. Sóense utiliza-los seguintes caracteres: | @ # $ % & / \ < > [ ] { } *. A estes pódenselles engadir, se estamos seguros de que non aparecen no texto, ( ) = _ ç Ç. Dado o limitado número de símbolos, compre pensarmos detidamente que cousas imos marcar e como. Unha solución para obter máis símbolos é facer combinacións como:

substantivo masculino singular: $$$
substantivo masculino plural: $$%
substantivo feminino singular: $#$
substantivo feminino plural: $#%
verbo transitivo: @
verbo intransitivo: &
adxectivo masculino singular: *$$

Unha segunda solución é utilizar algún destes símbolos para «enmarca»-las diferentes etiquetas:

<substantivo masculino singular> ou
[substantivo masculino singular] ou
{substantivo masculino singular}

2.4.2. Estándares

Tanto cando traballamos con textos para uso persoal como cando o facemos cos de uso público, o máis seguro é utilizar estándares de marcación. As vantaxes fundamentais destes sistemas son:

- xa están feitos e non hai que pensar en combinacións de signos.
- as súas etiquetas soen ser mnemotécnicas e, xa que logo, doadas de lembrar e entender.
- están documentados, polo que é doado estudia-lo seu uso e resolver dúbidas.
- existen programas de análise que «entenden» estes sistemas.

Véxamos brevemente algúns deles.

2.4.2.1. WordCruncher

Este sistema propón etiquetas ca estructura «|etiqueta_valor». Por exemplo[19]:

|aI (= acto I)
|e1 (= escena 1)
|l1 (= liña 1)
|tt (= tipo de texto)
|p (= personaxe)

|t Teatro para a xente.
|a Eduardo Blanco Amor.
|o Fas e Nefas ou o castelo enmeigado número 5000 e pico.
|tt dramatis personae.
...
|p curuxa primeira: ¿Que che parez, Violante, se pousamos aquí un bocadiño?

2.4.2.2. Para OCP

2.4.2.2.1. Sen marcas

O programa OCP pode levar a cabo certas análises sen necesidade de marcación: índices, concordancias, estatísticas e listados. O resultado será, por exemplo, un simple listado sen máis especificación que o número de liña no que se atopa a palabra; se precisamos máis información teremos que marca-lo texto.

---

[19] Texto tomado de Blanco Amor (1974).

2.4.2.2.2. COCOA

O sistema COCOA, que é o máis axeitado para OCP, ten a seguinte estructura: <nome_etiqueta_valor>. Por exemplo[20]:

<título> ou <t>
<autor> ou <a>
<capítulo> ou <c>
<sección> ou <s>

<título Devalar> <autor Otero Pedrayo> ... <capítulo X> <sección 1> <título sección "As palmas de Herbón"> <parágrafo 1> O libro de Pauliños Fontenla foi imprentado en "Nós" i-apareceu o mesmo día nas librarías da Rúa do Vilar, da Rúa da Oliva, da Praza de Ourense, da Rúa Policarpo, no Cantón de Porlier. Foi mirado por moitos, mercado por poucos, lido por algúns ... <sección 2> <título sección A néboa da ría> <parágrafo 1> Fradiño do San Francisco, ou neno ceibado nunha serán das palmetas escolásticas...

2.4.2.2.3. Formato fixo

Chámase así o sistema que reserva un número determinado de caracteres de cada liña para as marcas que utiliza. A súa estructura é «nome_valor_localización no texto»[21]:

VPASdV2CMSR001Abenzoada rosa
VPASdV2CMSR002con pétalas de pau enfrorescida,
VPASdV2CMSR003tecedeira de sonos e de rumos,
VPASdV2CMSR004a trelucir como unha lúa solta ...
VPASdV3MHCR001Si arrente do cantil das beiramares,
VPASdV3MHCR002pola escuma dos sonos salferido,
VPASdV3MHCR003a lira dos piñeiros ...

---

[20] Texto de Otero Pedrayo (1935).
[21] Texto de Paz Andrade (1968).

que se le: Valentín Paz Andrade, *Sementeira do Vento*, sección II (Cantarol do mar), composición «Salmo da Rosa»; Valentín Paz Andrade, *Sementeira do Vento*, sección III (Memorial do Home), composición «Canto da Remigración».

2.4.2.3. Etiqueta (*label*)

Consta só dun número entre chaves. Por exemplo[22]:

{1}R. Cabanillas
{2}Obra: Samos
{3}Composición: Oración
{4}Da nube baixa á decaída rosa
{5}o orballo que a sostén fresca e vizosa,
{6}e a rosa, agradecida
{7}á graciosa mercede recibida,
{8}fai do orballo fragancia deleitosa...

2.4.3. SGML e as directrices da TEI

A SGML (*Standard Generalized Markup Language*) ou Linguaxe de Marcación Xeralizada Estándar merece un capítulo de seu e diferenciado dos outros sistemas de marcación, xa que hoxe por hoxe é a máis completa e se está a estender cada vez máis.

---

[22] Texto tomado de Cabanillas (1958).

2.4.3.1. SGML: un pouco de historia[23]

A SGML tal e como hoxe a coñecemos é o resultado de 30 anos de investigacións levadas a cabo, fundamentalmente nos Estado Unidos. Nun principio, calquera documento en formato electrónico contiña certos códigos de control que definían o formato final do mesmo cando era impreso. Nos anos 60, varios investigadores como Tunnicliffe e Rice comezaron a pensar nun sistema de codificación xenérico que usase etiquetas ou marcas descritivas. O seu obxectivo era separa-lo contido do documento do seu formato.

No 1969, Charles Goldfarb, Edward Mosher e Raymond Lorie desenvolveron a chamada GML (que a parte de se-las iniciais dos seus apelidos significa *Generalized Markup Language*, Linguaxe de marcación xeralizada), que co tempo remataría por se converter na SGML. O seu obxectivo fundamental era facilita-la edición, formatado e transmisión de información independentemente das plataformas.

No 1978 o American National Standard Institute (ANSI) creou un comité especial para linguaxes de ordenador e para o procesamento de textos. Goldfarb colaborou con este comité para convertir SGML nun estándar aceptable e completo. O primeiro borrador de SGML foi publicado no 1980 e pronto se estendeu a Europa (ISO/EC), publicándose unha versión internacional no 1985. Nese ano creouse no Reino Unido un grupo de usuarios de SGML que dende entón ven

[23] Tomada de SGML User's Group (1985) "Standard Generalized Markup Language (International Standard ISO 8879). A Brief History".

fomentando o uso desta linguaxe e serve de axuda ós particulares e organizacións que a utilizan.

Entre as primeiras aplicacións da SGML está o Proxecto de Manuscritos Electrónicos da Asociación de Editores dos Estados Unidos (AAP), que tiña como obxectivo desenvolver unha aplicación para a creación de libros, xornais e artigos que facilitase o intercambio de manuscritos entre autores e editores.

SGML é, antes que nada, un xeito de concebir e describir documentos mediante a descrición da súa estructura e independentemente de calquera plataforma e, sobre todo, independentemente de calquera formato. Por exemplo, para describi-lo título dun capítulo, certos procesadores de texto usan códigos como os seguintes:

.nf	
.ps 42	(= tamaño 42 puntos)
.ffam BKL	(= tipo de letra Bookman light)
.ft I	(= cursiva)
.ce	(= centrado)
Este é o título	

No sistema de SGML, só habería que dicir:

<TI> Este é o título </TI>

E noutro lugar se especificarían as características de **`<TI>`**. Na codificación ordinaria, as etiquetas soen representar instruccións de formato; as etiquetas de SGML só especifican a estructura do texto e non din nada sobre o seu formato.

2.4.3.2. A filosofía da TEI: información exhaustiva

A TEI (*Text Encoding Initiative*, Iniciativa para a Codificación de Textos) é un proxecto internacional da ACH (*Association for Computers and the Humanities*), ACL (*Association for Computational Linguistics*) e a ALLC (*Association for Literary and Linguistic Computing*) e está subvencionada polo goberno dos Estados Unidos, pola Comunidade Europea, pola Fundación Mellow e polo Consello de Investigación das Ciencias Sociais e das Humanidades de Canadá. O seu obxectivo é desenvolver e difundir unhas directrices para a creación e intercambio de textos electrónicos. A necesidade dun único sistema de codificación que eliminase os problemas que supoñía a existencia de varios estándares levou á creación da TEI durante unha conferencia da ACH celebrada en Nova Iorque no 1987.

Os obxectivos que se marcou a TEI son:

- especificar un sistema de codificación de textos electrónicos único.
- redactar unha serie de recomendacións para a creación de novos textos: que é o que hai que codificar e como.
- documenta-los principais sistemas de codificación que xa existían e crear unha metalinguaxe que os describise.

Por razóns obvias (a independecia con respecto a sistemas e plataformas) a TEI elixiu a SGML para crear esa metalinguaxe de descrición. A SGML non é un sistema de codificación propiamente dito, senón un entorno dentro do cal se poden crear sistemas de codificación; é dicir, trátase dunha linguaxe coa cal se poden crear sistemas de marcación. Isto supón unha vantaxe moi grande, xa que fai que o número de etiquetas posibles sexa aberto, é dicir, cada usuario pode crea-las súas propias etiquetas para os seus usos específicos; a única condición é que as etiquetas estean correctamente descritas.

A TEI ven redactando unhas directrices dende 1987 e dende o mes de xuño de 1994 está publicada a terceira versión, corrixida e aumentada, das mesmas[24]. Están dirixidas ós investigadores de tódalas linguas e considéranse recomendacións de uso, non normas estrictas que haxa que seguir obrigatoriamente. Preténdese codifica-la maior cantidade de información posible, pero das máis de catrocentas etiquetas que ten definidas, só unhas poucas son obrigatorias; a filosofía básica da TEI pódese resumir como «se queres codificar isto, faino deste xeito».

As directrices céntranse exclusivamente na marcación descritiva, non de procesos (formatado) e parten do presuposto de que tódolos textos comparten unha serie de características básicas comúns, a partir das cales e cunhas limitadas variacións se pode xerar todo tipo de textos. O proceso de marcación, por outra parte, é progresivo, polo que se poden introducir máis etiquetas en calquera momento e segundo as necesidades dos usuarios.

[24] Cf. Sperberg-McQueen & Burnard (eds.) (1994).

A TEI posúe un boletín electrónico chamado TEI-L no que se publican novas relacionadas coa TEI e que serve tamén de foro de discusión sobre tódolos problemas relacionados coas directrices. Igualmente posúe un rexistro de documentos que leva conta de tódalas publicacións, investigacións e traballos relacionados coa TEI[25].

A principal desvantaxe que a TEI presenta, e podemos dicir que tamén a única, é que existen moi poucos programas que recoñezan as súas etiquetas e ademais son caros. Non embargantes, este problema vaise solventando ós poucos, xa que o número de usuarios crece e con el a cantidade de fabricantes interesados en crea-los programas axeitados. Entre os distintos programas que podemos atopar están:
- editores: *Author/Editor, Grif, WriteStation, WriteIt, SGMLTagger, Intellitag*.
- conversores: *MarkIt Application Language, SGML Translator, Omnimark, Tolk*.
- xeradores de etiquetas: *dataTAG, fastTAG, Magic Markup*.
- formatadores: *DynaText, IADS*.
- analizadores: *MarkIt, ParseStation, sgmls* (uso público)*, ARCSGML*.
- sistemas de xestión: *Basis+, DynaText, PAT*.

[25] Para subscribirse ó boletín, enviar unha mensaxe ó enderezo de Internet LISTSERV@UICVM.UIC.EDU coa instrucción "SUBSCRIBE TEI-L". Para recibir un listado dos documentos dispoñibles no rexistro, mandar ó mesmo enderezo a instrucción "GET TEI-L FILELIST".

3. Manipulación do texto electrónico

Nesta derradeira sección analizaremos brevemente un dos programas de análise textual máis estendidos. A limitación de espacio e tempo, e non criterios de calidade, imponnos esta selección.

3.1. Oxford Concordance Programme (OCP)

O OCP é un programa de análise textual que pode xerar listados de palabras (*wordlist*), índices (*index*) e concordancias (*concordances*) e as súas respectivas estatísticas a partir de textos que poden estar escritos en varias linguas e alfabetos. Un listado de palabras é unha lista de tódalas palabras do texto co número de aparicións de cada unha. Un índice é unha lista de palabras co seu número de aparicións e a súa localización. Unha concordancia é unha lista de palabras co seu número de aparicións, a súa localización e o seu contexto.

Os requisitos para instalar OCP son mínimos e dependen sobre todo do tamaño dos textos que se vaian analizar: dous megas de RAM ou máis, procesador o máis veloz posible (286 ou superior) e a maior cantidade de espacio libre no disco duro que se poida.

OCP precisa dous arquivos para comezar a traballar: un arquivo de texto ASCII (*text file*) e un arquivo de ordes (*commands file*) que ten a extensión .COM. Cando remate, producirá un terceiro arquivo de saída (*output file*) que conterá os resultados da análise.

3.1.1. O arquivo de texto

Como xa dixemos, os datos de entrada do OCP teñen que formar un arquivo de texto ASCII. A meirande parte dos porcesadores de texto posúen unha utilidade para crear este tipo de arquivos. Teremos que ter en conta un pequeno pero importante detalle: naqueles procesadores que nos ofrezan a posibilidade de crear arquivos dos tipos «só texto, só texto con saltos de liña, texto MS-DOS e texto MS-DOS con saltos de liña», deberemos elixi-la opción *texto MS-DOS con saltos de liña*, xa que a opción só texto non recoñece os caracteres acentuados nin o ñ, e as opcións *sen saltos de liña* convertirían cada parágrafo nunha soa liña, o que produciría erros no funcionamento do OCP.

Outro consello importante é que non debemos ter dentro do mesmo directorio do OCP demasiados arquivos cos mesmo nomes, xa que o OCP se traba. Por exemplo, imaxinemos que queremos facer varias análises dun mesmo texto e almacenalas en arquivos separados. Poderiamos te-los arquivos de texto texto1.txt, texto2.txt, texto3.txt, etc; os arquivos de ordes correspondentes ordes1.com, ordes2.com, etc. e os arquivos de saída saída1.fin, saída2.fin, etc. Isto causaríalle problemas Ó OCP, xa que confundiría uns nomes tan semellantes. Unha solución é porlle o número diante a cada nome: 1texto.txt, 2texto.txt, etc.

3.1.2. Arquivos de ordes. Sintaxe

Os arquivos de ordes do OCP poden ser creados dentro do mesmo programa ou ben en calquera outro editor de textos ASCII. O editor de ordes do OCP está limitado a cincuenta liñas por arquivo, polo que se precisamos un arquivo máis longo teremos que recorrer a outro editor. En todo caso, o número de caracteres por liña non debe sobrepasa-los oitenta.

As ordes divídense en catro apartados: *INPUT, *WORDS, *ACTION, *FORMAT, máis a orde final *GO que pon en marcha o programa. En *INPUT especificámo-las características do texto que queremos analizar: se está marcado, se ten referencias, se queremos analizalo todo, etc. En *WORDS definímo-lo alfabeto e outro símbolos que imos utilizar, a puntuación, etc. Na sección *ACTION eliximos que queremos facer e como: listados alfabéticos inversos de tódalas palabras, concordancias das palabras que posúan o ditongo -ia-, etc. Por último,. en *FORMAT definímo-lo formato co que imprimirémo-los resultados.

Cada unha destas seccións pode conter unha serie de ordes, pero cada unha das catro ten tamén un valor por defecto que veremos no seu momento. Dentro de cada sección as ordes non teñen unha xerarquía fixa, polo que podemos comezar ou rematar por calquera delas. Se queremos utilizar tódolos valores por defecto, non fará falla especificar ningunha orde, e mesmo poderemos omiti-los nomes das seccións, por exemplo «*GO » sería unha orde completa que nos devolvería un índice de tódalas palabras por orde alfabética, xa que esta é a orde por defecto.

Unha das grandes vantaxes do OCP é que a sintaxe das súas ordes é moi sinxela e moi semellante á linguaxe natural, pero pola contra, ten a desvantaxe de que esa linguaxe é o inglés, polo que haberá que aprender certas expresións, ordes e palabras clave. Outras características da linguaxe das ordes do programa son:

- Cada unha das ordes componse dunha palabra clave (PICK, DO, etc.), seguida de un ou máis dos seguintes elementos: a) unha palabra reservada como FREQUENCY, WORDS, CONCORDANCE, etc., b) números enteiros e c) cadeas ou secuencias de caracteres, que sempre deben ir entre comiñas («ter», «polo», «na», etc.).
- Calquera das ordes se pode especificar completa ou só coas primeiras catro letras: CONCORDANCE = CONC.
- As ordes poden partirse (para cambiar de liña) por calquera parte excepto polo medio dunha das chamadas palabras reservadas.
- Os elementos das ordes poden ir separados por un ou máis espacios.
- OCP é capaz de corrixir unha orde mal escrita sempre que non se diferencie moito do que debería dicir: entende que RIHGT debería ser RIGHT.
- Tódalas ordes teñen que rematar cun punto.

3.1.3. Construcción de ordes

A utilidade de construcción de arquivos de ordes de OCP ofrécenos dúas posibilidades. Podemos segui-la guía do programa e escoller entre unha serie de ordes predefinidas ou ben modificar un arquivo xa existente para introducirlle nós mesmos outras ordes diferentes das predefinidas.

3.1.3.1. Sección *input

Nesta sección definirémo-las características do texto que imos analizar. Dúas das súas opcións son:

(a) REFERENCES. Define o formato das referencias que se utilicen. OCP utilizará as referencias para localiza-las palabras ou concordancias no texto. As máis aconsellables para o OCP son COCOA, que se definen como «REFERENCES COCOA» ou ben «REFERENCES COCOA "<" TO ">"».

(b) COMMENTS BETWEEN «*cadea1*» TO «*cadea2*». Define as marcas que identifican os comentarios. Os comentarios son aquelas palabras que nunca serán seleccionadas como palabra clave pero que si poderán aparecer como contexto. Por exemplo, para especificar que os comentarios van entre dobres parénteses a orde sería «COMMENTS BETWEEN "((" TO "))"».

3.1.3.2. Sección *words

Nesta sección definirémo-lo alfabeto que imos utilizar. Compre ter en conta que amais das letras do alfabeto galego, deberiamos incluír todas aquelas que poidan aparecer no texto que analicemos, xa que isto nos evitaría problemas á hora de facer listados no caso de que houbese no texto palabras de idiomas estranxeiros. Por defecto, OCP usa o alfabeto inglés, polo que non ten acentos nin *ñ* e a *ch* conta como dúas letras. O OCP posúe ademais, xa predefinidos, os alfabetos español, italiano, alemán, francés, a transcrición do grego estándar e a transcrición do Thesaurus Linguae Graecae. Na descrición do alfabeto hai varias partes:

(a) ALPHABET. A relación de letras do alfabeto. Para o galego poderiamos utiliza-lo seguinte:

> "A=Á=a=á B=b C=c D=d E=É=e=é F=f G=g H=h I=i L=l LL=Ll=ll M=m N=n Ñ=ñ O=Ó=o=ó P=p Q=q R=r S=s T=t U=Ú=u=ú V=v X=x Z=z"

no que poderiamos engadir "J=j K=k W=w Y=y" e as vocais con acentos graves, circunflexos e diéreses se fose preciso.

(b) DIACRITICS. Define a lista de caracteres que deberán ser tratados como diacríticos ou acentos. Os caracteres declarados nesta sección serán a clave secundaria para a ordenación alfabética. É aconsellable que as vocais acentuadas se declaren neste apartado e no anterior ó mesmo tempo, xa que doutro xeito o programa pode dar problemas. Por exemplo:

> DIACRITICS "´ Á=á=É=é=Í=í=Ó=ó=Ú=ú" ZERO LENGTH.

onde ZERO LENGTH indica que o acento non inflúe na lonxitude da palabra.

Por defecto, o único diacrítico é o trazo. En galego non debemos especifica-lo trazo como diacrítico, xa que entón interpretaría «colle-lo» como «collelo» dividido a final de liña. O máis seguro é reserva-lo trazo para a segunda forma do artigo e non dividi-las palabras a final de liña.

(c) PADDING (*recheo, palla*). Son caracteres que queremos que se ignoren ó ordena-las palabras alfabeticamente. Por defecto non hai ningún caracter especificado como recheo, pero deberiamos incluír todos aqueles caracteres que utilicemos como marcas e que non sexan letras. Por exemplo: «PADDING "[ ] { } ( )"».

(d) PUNCTUATION. Define os signos de puntuación que aparecen no texto. Por exemplo: «PUNCTUATION ". , ; :: ? ¿ ¡ !"». Nótese que os dous puntos aparecen dúas veces. Isto é debido a que os dous puntos teñen un uso concreto na sintaxe do OCP: indicar que o que segue se usa co seu valor habitual. Por exemplo: o asterisco utilízase para representar calquera cadea de calquer número de caracteres incluído ningún. Pero se fose tamén unha letra do alfabeto (por exemplo nalgún tipo de transcrición fonética), para construír unha orde para que o OCP buscase tódalas palabras que comezasen por asterisco, precisariamos escribir «:**».

3.1.3.3. Sección *action

Nesta sección, a orde «DO» serve para especifica-lo tipo de análise que queremos facer. Os seus posibles argumentos son: «DO WORDLIST» (para facer un listado de palabras), «DO INDEX», (para facer un índice), «DO CONCORDANCE» (para facer unha concordancia), e «DO STATS» (para face-las estatísticas).

3.1.3.4. Sección *format

Nesta sección definirémo-lo aspecto da saída pola pantalla ou pola impresora das nosas análises. A opción por defecto ten moi poucos efectos visuais para non interferir en ningún caso nin cos procesadores de texto que despois se poidan utilizar nin cos intereses do usuario.

3.1.4. Demostracións: tipos de análises

Neste apartado amosaremos diversos tipos de análise coas súas ordes e saídas respectivas. Utilizaremos como arquivo de entrada catro capítulos dun libro traducido polo autor destas notas: *Galicia, a Suíza de España*, da autora inglesa de principios de século Annette M.B. Meakin. Velaquí unha mostra do texto[26]:

> <[Título: Galicia, a Suíza de España]> <[Autora: Annette M.B. Meakin]> <[Capítulos II a V]> <[Traductor: Alberto Álvarez Lugrís]> <[Data traducción: Decembro 1994]> <[Texto electrónico creado por Alberto Álvarez Lugrís]> <[Creado o 17 de xuño do 1995]> <[Utilizado como exemplo para o programa OCP no Curso de Verán Informática e análise textual (Informática e lingua)]> <[Marcación de capítulos, títulos e parágrafos tipo COCOA]> <[Por motivos prácticos, elimináronse as ilustracións, notas e todo tipo de formatos e non se mantén a paxinación orixinal]> <[Estas notas entre ángulos e corchetes serán consideradas como comentarios en OCP]> <C -> <T -> <P par-> <C cap2>CAPÍTULO II <T xeoG>XEOGRAFÍA DE GALICIA <P par2.1> Os límites de Galicia - Somontes dos Pirineos - As rías - Vexetación exuberante - O gando - Ríos - O "Río do esquecemento" - O Miño e o Sil - Industria de espadas - Ptolomeo - O primeiro mapa - Francia e a literatura sobre xeografía - Os máis fermosos portos de Europa - Colón e Galicia - Abundancia de reliquias do pasado <P

[26] Meakin (1908).

> par2.2> Galicia limita ó norte co Golfo de Biscaia, ó sur con Portugal e ó leste coas provincias de Asturias e León. Esta provincia é a máis occidental e ó mesmo tempo a máis setentrional de España, e o seu Cabo Fisterra foi unha vez o límite do Imperio Romano. Fisterra recibiu o seu nome dos romanos: a Fin da Terra. Os Pirineos, que se...

3.1.4.1. Listado con frecuencias

Comezaremos por un listado das palabras do texto co número de aparicións de cada unha delas. O arquivo de ordes que utilizaremos será o seguinte:

```
*input
references COCOA "<" to ">".
comments between "<[" to "]>".
*words
alph "A=Ä=Á=À=A-=a=á=ä=à=a- Æ=æ B=b C=c=Ç=ç CH=Ch=ch D=d
E=É=È=Ë=E-=e=é=ë=è=e- F=f G=g H=h I=Í=Ì=Ï=I-=i=í=ì=ï=i- J=j K=k
L=l LL=Ll=ll M=m N=n Ñ=ñ O=Ó=Ò=Ö=o=ó=ö=ò P=p Q=q R=r
RR=Rr=rr S=s T=t U=Ú=Ù=Ü=u=ú=ù=ü V=v W=w X=x Y=y Z=z -
0 1 2 3 4 5 6 7 8 9".
diac "'`¨ Á=À=Ä=A-=É=È=Ë=E-=Í=Ì=Ï=I-=Ó=Ò=Ö=Ú=Ù=Ü
        á=à=ä=a-=é=è=ë=e-=í=ì=ï=i-=ó=ò=ö=ú=ù=ü" zero length.
punc "_ . , :: ; ? ¿ ¡ ! :" ".
padd "( ) [ ]". {Galego con tódolos diacríticos}
*action
do wordlist.
*go
```

Obsérvese que incluímos no alfabeto tódolos diacríticos para non ter problemas cos nomes estranxeiros, citas noutras linguas, etc. Ademais, para evitar confusións cos trazos da segunda forma do artigo, identificamos como posibles letras do alfabeto as vocais seguidas por un guión e sen espacio: *a-, e-*, e *i-*. Especificamos tamén tódolos signos de puntuación e diacríticos posibles. Finalmente, como caracteres de recheo (*padding*) temos os parénteses e os corchertes. Non especificaremos ningún formato para comprobar cal é o formato por defecto do programa. O resultado é o seguinte:

a	535	abstrusos	1	acontecía	1
á	53	absurdo	1	Acquitaine	1
ab	1	Abundancia	1	Acquitanae	1
abade	5	abundante	1	activo	1
abadesa	5	abundantes	1	activos	1
abaixo	2	acada-la	1	actuais	1
abas	2	Acadou	2	actual	4
Abbatissae	1	accederon	1	actualidade	1
abertas	1	aceite	1	actualmente	1
abofé	1	acendidos	1	actuou	1
abonda	2	acepataron	1	acusación	1
abondo	3	aceptado	1	acusado	1
abondos	1	aceptar	1	acusar	1
about	1	acertadamente	1	achaba	1
Abraham	1	aciaga	1	achacábanlles	1
Abrahe	1	acoirazados	1	achádego	3
abraiante	2	acolle-las	1	áchanse	1
abraiantes	1	acolleuna	1	achegase	1
abraiar	1	acompañada	2	achou	2
abrande	1	acompañado	1	ad	9
abrangue	1	acompañaron	1	Ademais	2
abranguendo	1	acompañasen	1	Ademar	3
abrangueu	1	aconselloulle	1	adestrou	1
absoluto	2	acontecementos	3	adhuc	1
...					
Xerome	9	xustas	2	1013	1
Xerusalén	8	xustifique	1	1054	1
xesuita	1	Xustino	2	1066	1
xesuítas	1	xustiza	1	1087	1
Xesús	1	xusto	2	1098	1
XI	6	XVI	4	1105	1
XII	8	XVII	1	1110	1
XIII	7	XVIII	4	1140	1
Xinzo	4	y	1	1151	1
XIV	2	Yepes	1	116	1
XIX	2	zona	1	1274	1

```
xograres          1       zonas         1   1285                    1
1350-69           1       368           1
1453                      1      378              1
1455                      1      379              1
1471                      1      385              2
1499                      1      386              1
1550                      1      390              1
1615                      1      395              1
TOTAL WORDS READ                 =       15716
TOTAL WORDS SELECTED             =       15716
TOTAL WORDS PICKED               =       15716
TOTAL WORDS SAMPLED              =       15716
TOTAL WORDS KEPT                 =       15716
TOTAL VOCABULARY                 =        4148
```

3.1.4.2. Índice

Faremos agora un listado de palabras coa súa localización no texto, é dicir, un índice. A orde que utilizaremos será a seguinte:

```
*input
references COCOA "<" to ">".
comments between "<[" to "]>".
*words
alph "A=Ä=Á=À=A-=a=á=ä=à=a- Æ=æ B=b C=c=Ç=ç CH=Ch=ch D=d
E=É=È=Ë=E-=e=é=ë=è=e- F=f G=g H=h I=Í=Ì=Ï=I-=i=í=ì=ï=i- J=j K=k
L=l LL=Ll=ll M=m N=n Ñ=ñ O=Ó=Ò=Ö=o=ó=ö=ò P=p Q=q R=r
RR=Rr=rr S=s T=t U=Ú=Ù=Ü=u=ú=ù=ü V=v W=w X=x Y=y Z=z -
0 1 2 3 4 5 6 7 8 9".
diac "'`¨ Á=À=Ä=A-=É=È=Ë=E-=Í=Ì=Ï=I-=Ó=Ò=Ö=Ú=Ù=Ü=
        á=à=ä=a-=é=è=ë=e-=í=ì=ï=i-=ó=ò=ö=ú=ù=ü" zero length.
punc "_ . , :: ; ? ¿ ¡ ! :" ".
padd "( ) [ ]". {Galego con tódolos diacríticos}
*action
do index.
references P = 8.
*format
layout columns 2 and width 35 and gap 4 and left margin 4
and depth 40.
```

title "Demostración curso verán '95" centre and page 1 right
and headwords left cycle.
*go

Como resultado obteremos un índice coas referencias dos parágrafos nos que aparece cada palabra, disposto en dúas columnas de 35 caracteres de ancho cunha separación de catro caracteres e cunha marxe esquerda de 4 caracteres. Terá un título centrado, a numeración de páxina comezará no número 1 e a primeira e a derradeira entrada de cada páxina irán consignadas nunha cabeceira que alternará entre a dereita e a esquerda. Podemos comprobar todos estes aspectos no seguinte fragmento:

valde - velo Demostración curso verán '95

valde 1 par3.43

valente 1 par4.28

valentes 1 par3.2

Valerio 13 par3.34, par3.35,
par3.35, par3.36, par3.37, par3.38,
par3.38, par3.39, par3.45, par3.45,
par3.45, par3.49, par3.52

Valerius 1 par3.50

vales 5 par2.3, par2.4,
par5.17, par2.5, par2.5

valía 2 par4.12, par4.26

Valmar 12 par5.1, par5.19,
par5.21, par5.21, par5.21, par5.23,
par5.26, par5.28, par5.29, par5.29,
par5.29, par5.7

par3.45, par4.21, par5.20,

variedade 2 par2.3,
par5.19

varios 4 par3.19,
par3.25, par4.20, par4.23,

Varnefried 1 par4.23

waróns 1 par3.3

vasto 1 par5.20

Vaticana 2 par5.17,
par5.26

Vaticano 1 par3.52

vatum 1 par5.12

Vázquez 1 par3.7

veces 6 par2.3, par2.5,

valor 3 par3.6
par5.12, par5.20,
valle 1 par4.19

van 3 par2.2, par4.6, par5.28

vándalos 2 par3.18, par3.25

vao 1 par3.30

varía 1 par2.4

varias 2 par2.2, par3.34,
par3.35, par3.4, par3.45,
par5.20

veciño 2 par3.6, par3.7

Vega 1 par5.40

Velaiquí 1 par3.29

velo 1 par3.43

-147-

TOTAL WORDS READ	=	15716
TOTAL WORDS SELECTED	=	15716
TOTAL WORDS PICKED	=	15716
TOTAL WORDS SAMPLED	=	15716
TOTAL WORDS KEPT	=	15716
TOTAL VOCABULARY	=	4148

3.1.4.3. Concordancias

Para rematar faremos unhas concordancias de determinadas palabras e colocacións do texto. En primeiro lugar, buscarémo-las aparicións, cos seus contexto respectivos, das palabras Francia, Inglaterra, Portugal, Italia, Alemaña. As palabras clave ds concordancias irán situadas á esquerda e cada contexto ocupa´ra un máximo de catro liñas. Os contextos terán unha sanguría de dous caracteres e estarán separados por unha liña en branco. A orde será:

```
*input
references COCOA "<" to ">".
comments between "<[" to "]>".
*words
alph "A=Ä=Á=À=A-=‘A=a=á=ä=à=a-=‘a Æ=æ B=b C=c=Ç=ç
CH=Ch=ch D=d E=É=È=Ë=E-=‘E=e=é=ë=è=e-=‘e=‘è=‘é F=f G=g H=h
```

```
I=Í=Ì=Ï=I-=‘I=i=í=ì=ï=i-=‘i J=j K=k L=l LL=Ll=ll M=m N=n Ñ=ñ
O=Ó=Ò=Ö=‘O=o=ó=ö=ò=‘o P=p Q=q R=r RR=Rr=rr S=s T=t
U=Ú=Ù=Ü=‘U=u=ú=ù=ü=‘u V=v W=w X=x Y=y Z=z -
0 1 2 3 4 5 6 7 8 9".
diac "´`¨ Á=À=Ä=A-=É=È=Ë=E-=Í=Ì=Ï=I-=Ó=Ò=Ö=Ú=Ù=Ü=
        á=à=ä=a-=é=è=ë=e-=í=ì=ï=i-=ó=ò=ö=ú=ù=ü" zero length.
punc "_ . , :: ; ? ¿ ¡ ! :" ".
padd "( ) [ ]". {Galego con tódolos diacríticos}
*action
do concordance.
references P = 8.
maximum context left upto ". ; ¿ ¡" and span L and right upto ". ; ? !"
and span L.
pick words "Portugal, Inglaterra, Francia, Italia, Alemaña".
*format
context size 4 and indent 2.
layout length 75 and depth 45 and left margin 0 and lines 1 below
contexts.
headwords left.
titles "Concordancias" centre and page 1 right.
*go
```

Como resultado obteremos:

```
                              Concordancias

     Alemaña    2
  par3.24   A primeira edición apareceu en Alemaña xa no 1471

  par4.30                                    Non só
   en Inglaterra, senón tamén en Francia, Bélxica, Alemaña e mesmo
   en España se cría ata agora que Eadmer fora o primeiro en
   defender esta teoría

     Francia    9
  par2.1    Somontes dos Pirineos - As rías - Vexetación exuberante
   - O gando - Ríos - O "Río do esquecemento" - O Miño e o Sil -
   Industria de espadas - Ptolomeo - O primeiro mapa - Francia e a
   literatura sobre xeografía - Os máis fermosos portos de Europa -
   Colón e Galicia - Abundancia de reliquias
```

par4.1 Quen escribiu a Salve Regina? - Alfonso o Sabio - As súas Cantigas - A oración do mariñeiro - San Gregorio - Autoridades estranxeiras - Como a Salve chegou a Francia e Italia - A teoría do doutor Oviedo - Un artigo abraiante - O dogma da Inmaculada Concepción - De Consolatione
par4.22 No século XVI era xa popular en Francia, Portugal e Italia

-1-

TOTAL WORDS READ	=	15716
TOTAL WORDS SELECTED	=	15716
TOTAL WORDS PICKED	=	38
TOTAL WORDS SAMPLED	=	38
TOTAL WORDS KEPT	=	38
TOTAL VOCABULARY	=	5

Por último buscarémo-las concordancias de varias colocacións que poden resultar interesantes. Primeiro trataremos de ver que relación existe entre as preposicións «de» e «en»; despois buscaremos os usos do verbo viaxar neste texto de literatura de viaxes.

Para o primeiro caso, a orde será básicamente a mesma que no caso anterior, facendo só uns cambios no apartado *action:

```
*action
do concordance.
references P = 8.
maximum context left upto ". ; ¿ ¡" and span L and right upto ". ; ? !" and span L.
include only collocates "n@" upto 3 "d@", "n@@" upto 3 "d@@", "n@" upto 3 "d@@", "n@@" upto 3 "d@".
```

O resultado da operación será algo semellante a:

Concordancias preposicións EN - DE

nai de 1

par4.31

Decídeme esa que
mereceu a honra de se-la nai de Cristo, ¿foi
concebida sen pecado orixinal ou con el

Nai do 2

par4.31 foi concebida sen pecado
orixinal ou con el? A primeira posibilidade
semella, verdadeiramente a máis probable porque penso que á Virxe
Nai do noso Señor se lle concederon tódalas virtudes que podería
ter

par5.26 Esta lingua sería finalmente a nai do portugués

no das 1

par3.7 e de Nocela, en canto se decataron das intencións dos
visitantes, se reuniron no adro da pequena igrexa e protestaron
contra a retirada das pedras

-12-

TOTAL WORDS READ	=	15716
TOTAL WORDS SELECTED	=	15716
TOTAL WORDS PICKED	=	0
TOTAL WORDS SAMPLED	=	0
TOTAL WORDS KEPT	=	113
TOTAL VOCABULARY	=	30

Obsérvese que, dada a evidente limitación do programa para especificarlle as palabras que ten que buscar, amais das preposicións, devólvenos outras palabras que encaixan no patrón da busca, como por exemplo «nai de» e «nai do».

A busca dos usos do verbo *viaxar* sería semellante a esta, xa que na sección *action teriamos que especifica-lo seguinte.

```
*action
do concordance.
references P = 8.
maximum context left upto ". ; ¿ ¡" and span L and right upto ". ; ? !"
and span L.
pick words "viax*".
contexts sorted by right of keys.
```

que nos daría como resultado o seguinte:

```
Viaxar e derivados en Meakin
                                        viaxa   1
par3.1    to - Civitas Limicorum - Unha historia divertida - Os
primeiros anos de Idacio - Arrianismo - San Xerome - Paulo Orosio
- A traducción de King Alfred - San Agostiño e Orosio - Orosio
viaxa a Xerusalén - Os peregrinos romanos - Exeria - Unha abadesa
afouta - As súas visitas ós Santos Lugares

                                        viaxar 1
par5.41                 e cando un día o noso fiel amante estaba a
viaxar só, atopou a causa das súas coitas: dirixíase cara a el a
súa dama, e como recompensa polo enorme servicio que lle prestara

                                        viaxarmos      1
par2.7                  Ó viaxarmos por Galicia", escribía un monxe de
Oseira no século XVII, "experiméntase a cada milla ou, mellor
dito, a cada paso

                                        viaxase        1
par3.25                 Sobre o ano 414 San Agostiño aconselloulle a
Orosio que viaxase a Palestina para estudia-la herexía da Orixe in
situ e, ó mesmo tempo, consultar con San Xerome sobre algunha das
dúas
  -1-
TOTAL WORDS READ            =    15716
TOTAL WORDS SELECTED        =    15716
TOTAL WORDS PICKED          =       30
TOTAL WORDS SAMPLED         =       30
TOTAL WORDS KEPT            =       30
TOTAL VOCABULARY            =        9
```

APÉNDICE 1: Cabeceira TEI

```
<TEIHEADER>

<FILEDESC>
  <TITLESTMT>
    <TITLE> </TITLE>
    <AUTHOR> </AUTHOR>
    <SPONSOR> </SPONSOR>
    <FUNDER> </FUNDER>
    <PRINCIPAL> </PRINCIPAL>
    <RESPSTMT>
      <RESP> </RESP>
      <NAME> </NAME>
    </RESPSTMT>
  </TITLESTMT>
  <EDITIONSTMT>
    <EDITION>
      <DATE> </DATE>
      <RESPSTMT>
        <RESP> </RESP>
        <NAME> </NAME>
      </RESPSTMT>
    </EDITION>
  </EDITIONSTMT>
  <EXTENT> </EXTENT>
  <PUBLICATIONSTMT>
    <PUBLISHER> </PUBLISHER>
      <DATE> </DATE>
      <PUBPLACE> </PUBPLACE>
      <ADDRESS> </ADDRESS>
      <IDNO> </IDNO>
      <AVAILABILITY> </AVAILABILITY>
    <DISTRIBUTOR> </DISTRIBUTOR>
      <ADDRESS> </ADDRESS>
      <IDNO> </IDNO>
    <AUTHORITY> </AUTHORITY>
      <ADDRESS> </ADDRESS>
      <AVAILABILITY> </AVAILABILITY>
  </PUBLICATIONSTMT>
  <SERIESSTMT>
    <TITLE> </TITLE>
    <IDNO> </IDNO>
    <RESPSTMT>
      <RESP> </RESP>
```

```
      <NAME> </NAME>
    </RESPSTMT>
  </SERIESSTMT>
  <NOTESSTMT>
    <NOTE> </NOTE>
  </NOTESSTMT>
  <SOURCEDESC>
    <BIBLFULL>
      <AUTHOR> </AUTHOR>
      <TITLE> </TITLE>
      <TITLE TYPE =UB> </TITLE>
      <IMPRINT>
        <PUBPLACE> </PUBPLACE>
        <PUBLISHER> </PUBLISHER>
        <DATE> </DATE>
        <NUM> </NUM>
        <IDNO>ISBN: </IDNO>
        <IDNO TYPE="depósito legal"> </IDNO>
      </IMPRINT>
    </BIBLFULL>
    <SCRIPTSTMT> </SCRIPTSTMT>
    <RECORDINGSTMT>
      <RECORDING> </RECORDING>
      <EQUIPMENT> </EQUIPMENT>
      <BROADCAST> </BROADCAST>
    <RECORDINGSTMT> </RECORDINGSTMT>
  </SOURCEDESC>
</FILEDESC>

<ENCODINGDESC>
  <PROJECTDESC> </PROJECTDESC>
  <SAMPLINGDECL> </SAMPLINGDECL>
  <EDITORIALDECL>
    <CORRECTION> </CORRECTION>
    <NORMALIZATION> </NORMALIZATION>
    <QUOTATION> </QUOTATION>
    <HYPHENATION> </HYPHENATION>
    <SEGMENTATION> </SEGMENTATION>
    <STDVALS> </STDVALS>
    <INTERPRETATION> </INTERPRETATION>
  </EDITORIALDECL>
  <TAGSDECL>
    <RENDITION> </RENDITION>
    <TAGUSAGE> </TAGUSAGE>
  <REFSDECL>
    <STEP> </STEP>
    <STATE> </STATE>
```

```
  </REFSDECL>
  <CLASSDECL>
    <TAXONOMY>
      <CATEGORY>
        <CATDESC> </CATDESC>
      </CATEGORY>
    </TAXONOMY>
  </CLASSDECL>
  <FSDDECL> </FSDDECL>
  <METDECL>
    <SYMBOL>
  </METDECL>
  <VARIANTENCODING> </VARIANTENCODING>
</ENCODINGDESC>

<PROFILEDESC>
  <CREATION>
    <DATE VALUE=""> </DATE>
  </CREATION>
  <LANGUSAGE>
    <LANGUAGE ID = "galego"> Galego (100%)
      </LANGUAGE>
  </LANGUSAGE>
  <TEXTCLASS>
    <KEYWORDS>
      <LIST>
        <ITEM> </ITEM>
      </LIST>
    </KEYWORDS>
    <CLASSCODE> </CLASSCODE>
    <CATREF> </CATREF>
</PROFILEDESC>

<REVISIONDESC>
  <CHANGE>
    <DATE> </DATE>
    <RESPSTMT>
      <NAME> </NAME>
      <RESP> </RESP>
    </RESPSTMT>
      <ITEM> </ITEM>
  </CHANGE>
</REVISIONDESC>

</TEIHEADER>
```

REFERENCIAS BIBLIOGRÁFICAS

BLANCO-AMOR, E. (1974), *Teatro pra a xente*. Vigo: Galaxia.

CABANILLAS, R. (1958), *Samos*. Vigo: Galaxia.

GÓMEZ GUINOVART, J. e LORENZO SUÁREZ, A. (1994), *Vocabulario de Informática Galego-Inglés-Castelán*. Vigo: Servicio de Normalización Lingüística da Universidade de Vigo.

HOCKEY, S. (1994), «The Center for Electronic Texts in the Humanities» en Hockey, S. et. al. (eds) *Electronic Texts in the Humanities: Methods and Tools. Course Handbook*, pp. 11-23.

IGLESIA ALVARIÑO, A. (1961), *Nenias*. Vigo: Galaxia.

LANCASHIRE e MCCARTHY (eds.) (1989), *Humanities Computing Yearbook 1*. Oxford: Oxford University Press.

— (eds.) (1991), *Humanities Computing Yearbook 2*. Oxford: Oxford University Press.

MEAKIN, A. M. B. (1908), *Galicia, a Suíza de España*. Traducción de Alberto Álvarez Lugrís. Vigo: Galaxia, 1995. En prensa.

OTERO PEDRAYO, R. (1930), *Arredor de si*. Vigo: Galaxia.

— (1935), *Devalar*. Santiago: Nos.

PAZ ANDRADE, V. (1968), *Sementeira do Vento*. Vigo: Galaxia.

SGML User's Group (1985), «Standard Generalized Markup Language (International Standard ISO 8879). A Brief History».

SPERBERG-MCQUEEN, C.M. e BURNARD, L. (eds.) (1994), *Guidelines for Electronic Text Encoding and Interchange*, draft version 3, TEI document P3. Chicago & Oxford: ACH-ACL-ALLC.