Corrector ortográfico especializado para o proxecto IANUS

Eva Domínguez Noya
Centro Ramón Piñeiro para a Investigación en Humanidades

Xesús M. Mosquera Carregal
Servizo de Normalización Lingüística da Universidade da Coruña

As orixes do proxecto destinado á elaboración dun corrector ortográfico especializado para o ámbito da sanidade en lingua galega hai que procuralas na coincidencia temporal de dous proxectos ou momentos planificadores, por así dicir, trazados por parte da administración galega que resultan de especial relevancia en cadanseu ámbito de actuación: o sistema sanitario e a política lingüística. É o caso, por unha banda, do Plan xeral de normalización da lingua galega[1], aprobado polo Parlamento galego en 2004, e, pola outra, do impulso que por volta do ano 2006 se lle imprimiu desde a Consellaría de Sanidade ao plan de tecnoloxías e sistemas de información da sanidade galega, coñecido so a denominación de IANUS.

Tocante ao primeiro, lembremos que supuxo a materialización, desde hai anos desexada e reclamada polos sectores máis comprometidos coa revitalización social da lingua galega, dunha planificación lingüística sectorializada que servise asemade de punto de partida e de folla de ruta de utilidade e validez para os vindeiros anos. Para o que nos ocupa agora interésanos chamar a atención sobre o último dos cinco obxectivos xerais que establece (PXNLG: 39).

> Dotar o galego dos **recursos lingüísticos e técnicos** necesarios que o capaciten para vehicular a vida moderna.

e sobre dous dos sectores transversais sobre os que se estenden os obxectivos xerais (PXNLG: 51, 53):

[1] Xunta de Galicia (2004): *Plan xeral de normalización da lingua galega*. Dispoñible en: <http://www.xunta.es/linguagalega/arquivos/PNL22x24_textointegro%29.pdf>

Novas tecnoloxías
Fomentar a presenza da lingua galega nas **novas tecnoloxías**.
Lograr unha oferta ampla e competitiva de **produtos e recursos informáticos en galego**.
Potenciar a presenza da lingua galega en internet.
Potenciar a investigación da tradución automática, o recoñecemento e a síntese de voz, e outras novas técnicas que faciliten a opción positiva no mercado da información e da comunicación, e que aseguren a libre circulación do galego nos **sistemas avanzados** da vida moderna.

Implementación do corpus
Pór ao alcance dos cidadáns e dos **sectores profesionais** os medios formativos, didácticos, técnicos, lingüísticos e terminolóxicos suficientes que lles aseguren unha completa capacitación lingüística e un emprego doado do galego nas súas **actividades persoais e profesionais**.

As novas tecnoloxías aparecen no PXNLG como un novo parámetro que clasifica as linguas en función da súa presenza ou ausencia, mentres que a implementación do corpus se dirixe cara á consecución dunha maior seguranza lingüística por parte dos usuarios finais da lingua, aos cales se lles han de proporcionar os recursos e a formación necesarios para a correcta realización das súas actuacións en lingua galega.

Á súa vez, estes dous sectores transversais esténdense ao longo de sete sectores verticais ou ámbitos sociais entre os cales está incluída a sanidade. A análise sociolingüística que se fai do sector sanitario caracterízase por un forte ocultamento da lingua por parte dos seus usuarios e mais por unha moi insuficiente correspondencia entre a lingua do paciente e a do persoal sanitario (nomeadamente, na nosa opinión, a do persoal médico habería que engadir). En todo caso o que nos interesa é detérmonos na mención dalgúns dos puntos febles que se sinalan no Plan (PXNLG: 168-169):

- Non se emprega maioritariamente o galego para os rexistros de datos das historias clínicas.
- Os programas informáticos utilizados polo persoal sanitario nas áreas periféricas están maioritariamente en castelán.
- Insuficientes recursos lingüísticos e terminolóxicos.
- As receitas do Sergas veñen maioritariamente cubertas en castelán.

Se, alén do dito, se repara en que un dos obxectivos xerais que se estableceron para o sector que nos ocupa é o de «Potenciar o emprego da lingua galega na área sanitaria» (PXNLG; 171) semella máis que xustificado e pertinente o proxecto que a seguir presentaremos.

Verbo de IANUS trátase dun ambicioso e complexo proceso que no marco das chamadas tecnoloxías da información e comunicación consiste, basicamente, nunha ferramenta que integra nun único sistema de información clínico-asistencial os historiais clínicos dos e das pacientes para dar lugar, así, á chamada historia clínica electrónica (HCE). Tales son as dimensións e as mudanzas que a súa implantación plena comporta que o seu alcance resulta transversal ao abranguer a todos os axentes implicados na atención sanitaria, incluída a rede farmacéutica (a chamada receita electrónica). Segundo a información consultada repercute, entre outras cousas, nos profesionais sanitarios naquilo que ten a ver coa normalización e homoxeneización da documentación clínica, así como na maior conexión que permite entre a atención especializada e a atención primaria (as interconsultas entre servizos); no persoal farmacéutico ao favorecer a súa integración na rede sanitaria e ofrecerlle máis datos sobre o paciente e para o paciente; e neste último, por lle permitir *a priori* o acceso a unha mellor, máis clara e segura información[2].

Para alén do dito, tamén cómpre citar senllas disposicións normativas, de distinto rango, mais con claros efectos na cuestión que nos ocupa para o ámbito sanitario: a primeira é a Instrución 11/05[3] verbo do emprego do galego nos documentos da Consellaría de Sanidade e o Servizo Galego de Saúde asinada polas secretarías xerais de ambos os entes. A instrución quinta di literalmente:

> Os impresos normalizados ou aqueles que resulten do emprego ou aplicación de técnicas informáticas ou medios electrónicos, informáticos ou telemáticos, de carácter repetitivo, que deban causar os seus efectos tanto dentro como fora da Comunidade Autónoma de Galicia, deberán estar dispoñibles nas linguas galega e castelá.

[2] Para quen quixer ampliar a información ao respecto pode consultar os dous volumes publicados polo Sergas baixo o título *Documentación clínica electrónica nos hospitais de Galicia. Ianus* e mais o folleto *A dirección única de área e as novas tecnoloxías: un novo modelo*. Ambas as referencias son do ano 2006 e pódese acceder a elas desde a sección de Publicacións da páxina web do Servizo Galego de Saúde <http://www.sergas.es/MostrarContidos_N2_T01.aspx?IdPaxina=60020>

[3] Pódese consultar unha versión escaneada do documento orixinal en <http://www.cig-saude.org/AGMEDICA/html/Instruccion-emprego-do-galego.pdf> [Data da consulta: 6/6/2011].

A segunda correspóndese coa Lei8/2008, do 10 de xullo, de saúde Galicia publicada no *Diario Oficial de Galicia* o 24 de xullo de 2008. O seu artigo 9.8 establece dentro dos dereitos relacionados coa confidencialidade e a información o seguinte:

> Para garantir a mellor información sobre a historia clínica do ou da paciente, tendo en conta as novas tecnoloxías, os datos desta estarán dispoñibles en tres idiomas (galego, castelán e inglés), facendo para iso as adaptacións técnicas necesarias.

Así pois, con base nos dous programas de actuación e para dar cumprimento ás disposicións normativas establecidas, a Consellería de Sanidade e a Secretaría Xeral de Política Lingüística asinaron xa no ano 2007 un convenio que tiña como finalidade a elaboración dun lexicón computacional pertencente ao ámbito médico destinado a alimentar un sóftware de corrección ortográfica que, unha vez implementado no sistema IANUS, servise de axuda e estímulo para mellorar a calidade e incrementar o uso da lingua galega no ámbito médico. Para que así fose, non había dúbida de que, para alén do léxico común, cumpría dispormos de corpus textuais especializados onde podermos documentar a terminoloxía específica, de aí que o convenio establecese que o Servizo Galego de Saúde achegaría un conxunto inicial de historiais clínicos en formato XML sobre o que realizar os correspondentes baleirados terminolóxicos. Previíase, así mesmo, un fluxo de comunicación permanente entre o Centro Ramón Piñeiro para a Investigación en Humanidades (CIRP) –onde se realizou todo o traballo lingüístico– e o Sergas conforme fosen avanzando os traballos: desde o CIRP enviando primeiramente un paquete inicial de palabras (ao que lle habían seguir outros máis) que servise para xerar os arquivos con que implementar o corrector no sistema IANUS e desde o Sergas conforme se ía alimentando o sóftware de corrección remitindo novos textos e mesmo novas palabras propostas polo propio persoal usuario do sistema que resultasen descoñecidas para o corrector en probas.

Naquela altura desde o Centro Ramón Piñeiro para a Investigación en Humanidades estábase en disposición de proporcionar a listaxe de formas flexionadas pertencentes ao léxico común galego dado que eses datos forman parte dun recurso lingüístico imprescindible para desenvolver un dos seus proxectos: a *Etiquetación e lematización do Corpus de Referencia do Galego Actual*. Porén, polo que respecta ás formas técnicas o CIRP non posuía esa listaxe mais a súa elaboración encaixaba perfectamente no lexicón XIADA (véxase máis adiante), en concreto na creación dun módulo, paralelo ao do léxico xeral, que fose específico para o ámbito médico.

Contextualización da listaxe de formas: a etiquetación e lematización do Corpus de Referencia do Galego Actual (CORGA)

Que relación garda a listaxe de formas flexionadas e conxugadas do léxico común do galego coa etiquetación e lematización do Corpus de Referencia do Galego Actual? Procuraremos ser breves mais é preciso que situemos o lexicón do que extraeremos a listaxe de formas comúns nun proxecto moito máis amplo e complexo.

Como sabe quen nalgún momento consultou o *Corpus de Referencia do Galego Actual* (CORGA)[4], este é un corpus documental que na súa versión 2.5 chega case aos 26 millóns de formas (sen contarmos as cifras) das cales 389 917 son formas diferentes. Está integrado por distintos tipos de textos –xornais, semanarios, revistas, ensaios, e textos de ficción (novela, relato curto e teatro)– e abrangue temporalmente desde o ano 1975 até a actualidade. Os textos que contén están codificados no estándar XML (*eXtensible Markup Language*) mais a codificación que se practica no CORGA só está relacionada coa información bibliográfica e a estruturación interna do documento. Ademais, cada documento está clasificado tematicamente, o que permite realizar buscas, con ou sen comodíns, por expresión, palabra exacta ou *booleana* segundo diversos criterios que poden combinarse en función das necesidades do usuario: período temporal, área temática, tipo de documento, parte do documento etc.

Non obstante, para facer buscas máis avanzadas é imprescindible que os textos do CORGA estean lematizados e etiquetados, ou sexa, é preciso unha etiquetaxe entendida esta como o proceso mediante o cal a cada unidade léxica se lle asigna unha etiqueta e un lema específicos para cada caso. Logo de varios anos de traballo nesta dirección o resultado tanxible, a finais de 2010, é a versión 2.4 do Corpus de Referencia do Galego Actual etiquetado[5] (CORGAetq), un subcorpus do CORGA formado por algo máis de 360 000 palabras procedentes do ámbito xornalístico, etiquetado automaticamente e desambiguado á man por un lingüista, que coincide, de momento, co corpus de adestramento do *Etiquetador e lematizador do galego actual*[6].

Para asignarlles ás formas do CORGA unha etiqueta e un lema, primeiro houbo que tomar unha serie de decisións caso das etiquetas que se ían empregar, as clases

[4] <http://corpus.cirp.es/corga>

[5] <http://corpus.cirp.es/corgaetq>

[6] <http://corpus.cirp.es/xiada>

morfolóxicas que se ían utilizar ou os atributos que as ían caracterizar. Noutras palabras houbo que delimitar un etiquetario ou, tal e como adoito se denomina en lingüística de corpus, un *tagset*.

O etiquetario que se configurou para XIADA pódese consultar na páxina do proxecto[7] e a través del accédese á seguinte información: unha primeira columna que recolle as categorías gramaticais consideradas, seguida de cadansúa columna para os atributos e valores pertinentes segundo for o caso. Vexamos o dito cun par de exemplos:

								p-presente			
							a-acusativo	i-copretérito	i-indicativo		
			m-masculino			1-primeira	d-dativo	l-antepretérito	s-subxuntivo		
			f-feminino	s-singular		2-segunda	n-nominativo	f-futuro	m-imperativo		
		c-coordinante	a-masculino/feminino	p- plural	c-comparativo	3-terceira	p-preposicional	e-pretérito	f-infinitivo	s-singular	n-non determinante
		s-subordinante	n-neutro	a- singular/plural	s-superlativo	a-primeira/terceira	f- formante léxico	c-pospretérito	x-xerundio	p-plural	d-determinante
		0- non aplica	0- non aplica	0- non aplica	0-non aplica	0- non aplica	o-outros casos	0-non aplica	p-participio	a-singular/plural	a-determinante/non determinante
Categoría	**Tipo**	**Subtipo**	**Xénero**	**Número**	**Grao**	**Persoa**	**Caso**	**Tempo verbal**	**Modo**	**Posuidor**	**Valor**
Substantivo (S) 1	c-común p-propio 2		m, f, a, 0 3	s, p, a, 0 4							

Da lectura desta táboa obtense que un substantivo como *análises* para o sistema é Scfp (substantivo común feminino plural), *páncreas* é Scma (substantivo común masculino invariable respecto ao número) ou Vigo, por citarmos tres exemplos, é Sp00 (substantivo propio sobre o que non se aplica nin o xénero nin o número[8].

[7] <http://corpus.cirp.es/xiada/etiquetario.html>

[8] O número que aparece na parte inferior dereita das celas indica a posición que o atributo ocupa na etiqueta final.

								p-presente			
							a-acusativo	i-copretérito	i-indicativo		
			m-masculino			1-primeira	d-dativo	l-antepretérito	s-subxuntivo		
			f-feminino	s-singular		2-segunda	n-nominativo	f-futuro	m-imperativo		
		c-coordinante	a-masculino/feminino	p- plural	c-comparativo	3-terceira	p-preposicional	e-pretérito	f-infinitivo	s-singular	n-non determinante
		s-subordinante	n-neutro	a- singular/plural	s-superlativo	a-primeira/terceira	f- formante léxico	c-pospretérito	x-xerundio	p-plural	d-determinante
		0- non aplica	0- non aplica	0- non aplica	0-non aplica	0- non aplica	o-outros casos	0-non aplica	p-participio	a-singular/plural	a-determinante/non determinante
Categoría	**Tipo**	**Subtipo**	**Xénero**	**Número**	**Grao**	**Persoa**	**Caso**	**Tempo verbal**	**Modo**	**Posuidor**	**Valor**
Verbo (V) 1			m, f, 0 5	s, p, 0 6		1ª, 2ª, 3ª, a, 0 4		p, i, l, f, e, c, 0 2	i, s, m, f, x, p 3		

Para a categoría gramatical do verbo xa entran en xogo máis atributos de tal forma que a un xerundio como *refacendo* lle corresponde a etiqueta V0x000 (forma verbal non temporal, xerundio, sobre que o non se aplica nin persoa, nin xénero nin número) e unha forma persoal como *xurdiu* represéntase como Vei30s (forma verbal do pretérito de indicativo na súa terceira persoa de singular sen que aplique o xénero).

Cabe engadir que o número de etiquetas posibles que resulta do etiquetario anda por volta das 400 (concretamente 383) e con elas é posible dar conta, na teoría e na práctica, de calquera palabra pertencente á lingua galega independente de cal for o seu contexto de uso.

Outro recurso fundamental para a etiquetaxe de corpus é a existencia dun lexicón, xa que aínda que os etiquetadores tentan adiviñar a etiqueta dunha forma que non estea presente no lexicón, canto maior for este máis posibilidades de acerto haberá, pois en todo lexicón se almacenan, entre outros datos, a información de etiqueta e o lema que lle corresponde a cada palabra. Deste recurso en concreto é de onde se extraeu parte da listaxe de formas tanto do léxico común como do técnico que se lle entregou ao Servizo Galego de Saúde. Só parte porque no lexicón unicamente se introducen as unidades que poden aparecer illadas nun texto, é dicir, no lexicón está introducida a primeira forma do artigo determinado mais non a segunda (o recoñecemento desta, ao igual que o das contraccións ou o das formas verbais con pronomes enclíticos, prodúcese por outras vías). É por iso que antes de actuar o etiquetador entra en acción o preprocesador que se ocupa de segmentar o texto que se vai etiquetar en unidades léxicas. As unidades multigramaticais, aquelas que se corresponden cunha única forma gráfica, pero que conteñen varios elementos lingüísticos, tal é o caso das

contraccións ou das formas verbais con clíticos, son desagregadas en compoñentes, aos que se lles asigna unha etiqueta e un lema. Para realizar o seu labor o preprocesador conta cun lexicón de formas verbais que poden levar enclíticos e con información detallada de cales son as secuencias de clíticos que poden engadírselles; coñece tamén cales son as secuencias contractas e todas as súas posibles análises; e dispón, finalmente, dunha serie de regras lingüísticas que o axudan a segmentar e etiquetar, mais que tamén son determinantes para a xeración da conxugación verbal cando se constrúe, se for o caso, coa segunda forma do artigo ou con enclíticos. A outra parte da listaxe de formas para IANUS procede destoutros recursos.

Estrutura do lexicón de XIADA

Antes de máis nada debemos aclarar que o lexicón de XIADA, a pesar de clasificarse entre os lexicóns morfolóxicos que mencionan Schiller e Karttunen (1999: 135-147), non está constituído por unha mera relación ou listaxe de formas gráficas á que se lles proporciona a etiqueta e o lema correspondentes. En realidade non hai formas, senón que estas son xeradas a partir da información que contén cada entrada tanto para o léxico común como para o técnico. Por exemplo, a unidade *facultativos* (substantivo e adxectivo) non aparece como tal en ningún lugar do lexicón, mais o etiquetador recoñécea porque é xerada polo sistema a partir da entrada «facultativ», cuxo lema e subetiquetas son, respectivamente, «facultativo» e Sc (substantivo común) e A0 (adxectivo, grao non aplica), e se remite ao grupo de derivación G1 onde figura unha desinencia *«os»* coa caracterización morfolóxica de masculino plural.

O que se perseguía era que a introdución e xestión dos datos fose práctica e nos permitise aforrar tempo e espazo en disco, polo que se estimou que maioritariamente as entradas do lexicón tiñan que ser caracterizadas morfoloxicamente desagregándoas en lema, raíz, subetiqueta e grupo de derivación. Con esta estrutura evitouse a introdución, por exemplo, de 70 formas verbais por cada un dos verbos introducidos no lexicón aínda que, en contrapartida, esixiu un estudo formal pormenorizado dos integrantes de cada unha das clases gramaticais variables para agrupalos en modelos e, deste xeito, construír a gramática formal que permitise analizar e tamén xerar as formas flexionadas e conxugadas do galego moderno. Así, por exemplo, nas categorías variables, exceptuando o verbo, foi necesario o establecemento de 46 grupos de derivación que permitisen integrar os distintos modelos flexivos de xénero e número en combinación coas diferentes variacións gráficas que se producen nas raíces. Repárese en que houbo que clasificar formalmente todas as posibles variacións non só á hora de flexionar o xénero ou o número, senón tamén tendo en conta as modificacións gráfi-

cas que sofre a raíz, sexan estas acentuais (*afebril* vs. *afebrís*), sexan alternancias nos caracteres (*bifidez* vs. *bifideces*). No anexo I achégase unha táboa que recolle os grupos derivacionais válidos tanto para o lexicón xeral como para o médico, e as súas características: o representante, as terminacións, as subetiquetas morfolóxicas que as caracterizan e a indicación da normatividade para cada unha das terminacións.

Así pois, todas e cada unha das entradas constan de lema, raíz, subetiqueta e, maioritariamente nas pertencentes ás clases flexivas, de grupo de derivación. Indo por partes cómpre dicir:

1) O lema é o representante das distintas formas flexionadas ou conxugadas que se integran nun paradigma e, obrigatoriamente, cada entrada do lexicón debe remitirse cando menos a un lema. Nótese que a clase morfolóxica é un elemento constitutivo do lema polo que todo lema está asociado a unha clase de palabra, de tal xeito que para unha unidade como *médico* haberá dous lemas: un, asociado á categoría substantivo, e outro á de adxectivo.

2) *Grosso modo* pódese considerar que a raíz é o lema sen morfemas gramaticais. Por exemplo, ao lema «médico» correspóndelle a raíz «médic», mais, polo xeral, raíz e lema coinciden nas entradas das clases invariables e nas dos substantivos e adxectivos que só presentan flexión no número[9].

3) O termo subetiqueta emprégase para referírmonos á caracterización morfolóxica básica da entrada no lexicón. Isto é, continuando a exemplificación con *médico,* as subetiquetas correspondentes ás entradas como substantivo e adxectivo son, respectivamente, Sc (substantivo común) e A0 (adxectivo, grao non aplica).

4) Finalmente, os grupos de derivación constitúen conxuntos de terminacións para as que se proporcionan uns valores. Por exemplo, as dúas raíces «médic», pertencentes aos lemas «médico» substantivo e «médico» adxectivo, remítense ao grupo de derivación G1 en que se interpreta que se a entrada

remata en	*-o*,	a caracterice como	→	*masculino singular*
“	*-a*,	“	→	*feminino singular*
“	*-os*,	“	→	*masculino plural*
“	*-as*,	“	→	*feminino plural*

[9] En ningún caso se pode identificar esta raíz co lexema da gramática tradicional posto que para o que nos ocupa só se teñen en conta os morfemas flexivos e non os derivativos.

Así se explica que para todos os substantivos e adxectivos que teñen flexión de xénero e forman o plural engadindo *-s*, (*neno, médico, facultativo...*) só sexa preciso introducir o lema e a raíz correspondentes, sen necesidade de ter que inserir todas as formas do paradigma.

Con todo, en XIADA non só se pretende lograr a identificación e caracterización morfolóxica plena de calquera unidade léxica galega actual, senón que entre os seus obxectivos tamén está contar con información sobre a procedencia do lema, o ámbito léxico a que pertence e a indicación sobre a normativa oficial actual. Obviamente estes datos non son imprescindibles para etiquetar corpus, mais o feito de dispor deles sempre permite atoparlles unha utilidade como pode ser a xeración do léxico común galego. Ademais, facelo *a posteriori* da construción do lexicón sería altamente custoso e suporía perder a oportunidade de contar cun lexicón amplo e completo desde os seus inicios.

Dito isto, no lexicón de XIADA, para alén do lema, a raíz, a subetiqueta, a categoría gramatical e o grupo de derivación, cada entrada responde aos parámetros seguintes:

a) Normativa oficial. Para un correcto recoñecemento das formas existentes nos textos galegos contemporáneos é preciso que o lexicón non só estea integrado por lemas normativos. Pénsese que de non ser así moitos dos documentos do CORGA non poderían ser etiquetados ou só o serían a medias. É fundamental, daquela, darlles cabida tamén a dialectalismos, hiperenxebrismos, castelanismos etc. polo simple feito de estes apareceren nos textos. Non obstante, para os poder diferenciar introduciuse este criterio na caracterización de lemas, formas[10] e incluso algunha desinencia.

Ora ben, cómpre aclarar que cando se indica que un lema ou unha forma son normativos non implica que esten avalados polas autoridades competentes nesta materia, é dicir a Real Academia Galega. A pretensión con que se inclúe este campo no lexicón de XIADA, e polo tanto tamén na súa versión descargable[11], é facilitar información sobre a corrección ou incorrección de cada unha das entradas. Así, marcamos con «s» tanto os lemas que non aparecen condenados nas distintas obras lexicográficas actuais como aqueles que só se documentan en corpus e se entende que

[10] Repárese en que unha forma verbal como *fosemos* era considerada como correcta até xullo do ano 2003 e remitíase aos lemas «ser» ou «ir» normativos. Non obstante, per se é un *token* que non responde á normativa actual.

[11] <http://corpus.cirp.es/xiada/descargas.html>

a súa formación –mediante prefixación, sufixación, composición– se axeita ao estándar léxico do galego. Por exemplo, en ningún dicionario dos manexados (véxase a relación máis adiante) se documenta o adxectivo *funcionarial*, mais abonda con partir dunha base correcta *funcionario* e dun sufixo *-al* que indica 'relativo ou pertencente a', ao igual que sucede no par *empresario/empresarial* para o considerar como unha forma normativa. Así se explica, en definitiva, que determinados lemas se consideren normativos no noso sistema malia non contaren (aínda) co aval da RAG.

b) Ámbito léxico. Desde o principio entendeuse como útil diferenciar na estrutura do lexicón módulos segundo o tipo de léxico que contiveren e que poidan ser combinados. Para iso dispúxose no deseño da estrutura, e así se implementou na construción do lexicón, dun campo máis en que se caracteriza o lema como pertencente ao léxico común ou ao técnico-científico, e se especifica neste último o ámbito no que se clasifica: administración, economía, medicina etc.

Polo momento, o sistema conta soamente co lexicón *xeral*, no que se inclúe o léxico común da lingua galega, e mais o lexicón *médico*, en que se introduciu o vocabulario específico do ámbito da medicina tirado do material proporcionado polo Servizo Galego de Saúde.

c) Fonte. Relacionado en parte coa epígrafe *a)* ocorre que nos textos actuais se presentan formas que non aparecen nos dicionarios e/ou vocabularios existentes da lingua galega ben porque son termos técnicos para os que se acaba de propoñer unha denominación, ben porque se documentan cun uso categorial distinto, ben polo simple feito de que non todos os lemas dunha lingua se recollen nos dicionarios. Non obstante, a súa introdución no lexicón resulta imprescindible para un correcto recoñecemento e unha completa caracterización dos textos. Isto explica a decisión de incluírmos a fonte para así achegar asemade información sobre a posible documentación do lema e a fiabilidade da forma. Concretamente en XIADA esta información aparece recollida na última columna da entrada. De seguido a táboa 1 cos códigos que nela poden aparecer e as obras a que fan referencia dispostas cronoloxicamente en orde decrecente.

Recapitulando, o lexicón de XIADA, ademais de conter os datos que posibilitan a identificación e caracterización morfolóxica plena de calquera unidade léxica galega actual, proporciona información sobre a normativa oficial, o ámbito léxico e a procedencia da forma. Todo isto significa que contamos cunha ferramenta lingüística apropiada para, a partir dela, xerar a lista de formas, tanto comúns como técnicas, que cumpran co requisito de ser correctas en galego.

Táboa 1. Códigos e referencias bibliográficas

Código	Obra ou fonte
corga	Centro Ramón Piñeiro para a Investigación en Humanidades (2010): *Corpus de Referencia do Galego Actual, versión 1.5*, <http://corpus.cirp.es/corga>
xerais_2009	Carballeira Anllo, X. M. (coord.) (2009): *Gran Dicionario Xerais da Lingua* (Vigo: Xerais)
sergas	Material proporcionado polo Sergas procedente de historiais clínicos
volga_2004	González González, M. e A. Santamarina Fernández (coords.) (2004): *Vocabulario ortográfico da lingua galega (VOLGa)* (A Coruña-Santiago de Compostela: Real Academia Galega-Instituto da Lingua Galega)
drag_2004	García, C. e M. González González (dirs.) (2004): *Dicionario castelán-galego da Real Academia Galega* (A Coruña: Fundación Pedro Barrié de la Maza)
olg_2004	González Rei, B. (2004): *Ortografía da lingua galega* (A Coruña: Galinova)
irindo_2004	Ledo Cabido, Bieito (dir.) (2004): *Dicionario de galego* (Vigo: Ir Indo)
normas_2003	Real Academia Galega / Instituto da Lingua Galega (2003): *Normas ortográficas e morfolóxicas do idioma galego* (A Coruña-Santiago de Compostela: Real Academia Galega-Instituto da Lingua Galega)
dgtm_2002	Real Academia de Medicina e Cirurxía de Galicia (2002): *Diccionario galego de termos médicos* (Santiago de Compostela: Dirección Xeral de Política Lingüística)
xerais_2000	Carballeira Anllo, X. M. (coord.) (2000): *Gran Diccionario Xerais da Lingua* (Vigo: Xerais)
drag_1997	García, C. e M. González González (dirs.) (1997): *Diccionario da Real Academia Galega* (A Coruña: Real Academia Galega)

XIADA: lexicón xeral

A descrición que acabamos de realizar sobre o lexicón fai patente o seu grao de complexidade estrutural e informativa, o cal posibilita a extracción de datos en distintos niveis de profundidade en función da utilidade que se persiga: desde a extracción dunha simple listaxe de formas até a conxugación dun ou máis verbos que nel figuraren (con ou sen pronomes enclíticos), pasando pola etiquetación automática de formas. A xeración da lista de formas que cumpran co criterio de integrarse no lexicón xeral e ser normativas é un dos resultados máis sinxelos.

Débese reparar en que as entradas do lexicón xeral son suficientes para cubrir as necesidades lingüísticas que se lle poidan requirir ao sistema de IANUS: o feito de implementar nel o lexicón coas entradas do *Vocabulario Ortográfico da Lingua Galega* xunto coas 100 000 formas máis frecuentes do CORGA, dá como resultado algo máis de 55 000 lemas. Entendeuse, por tanto, que desde o Centro Ramón Piñeiro se satisfacía amplamente a demanda do Servizo Galego de Saúde en canto ao vocabulario de uso cotián.

Así, a finais do 2006, cando se lle entregou unha listaxe inicial e provisoria de formas á administración sanitaria constituída por algo máis de 4 millóns de entradas pertencentes ao léxico común, non era posible desagregar con exactitude, mais como queira que desde aquela os cambios realizados no sistema non foron moitos, pódese fixar como referencia fiable o estado actual do lexicón: arestora en XIADA (versión 2.4[12]) é posible saber cantas entradas hai no lexicón en función da súa categoría gramatical, etiqueta e subetiqueta, e grupo de derivación. Velaquí a táboa que así o reflicte:

Táboa 2

Subetiqueta	Raíces	Lemas	Formas	Formas únicas	Total lemas por categoría
Sc [substantivos comúns]	30 668	30 509	75 389	68 574	30 509
Sp [substantivos propios]	0	0	0	0	
A0 [adxectivos, grao non aplica]	14 162	14 129	66 649	46 693	14 141
As [adxectivos superlativos]	13	34	140	132	
Ac [adxectivos comparativos]	7	7	42	14	
V [verbos]	9455	6903	573 653	392 735	6903

[12] http://corpus.cirp.es/xiada/descargas.html

Subetiqueta	Raíces	Lemas	Formas	Formas únicas	Total lemas por categoría
Wa [adverbios nuclear/modificador]	1774	0	0	0	1962
Wn [adverbios nucleares]	193	1957	1967	1967	
Wm [adverbios modificadores]	4	1774	1777	1777	
Wr [adverbios relativos]	4	4	4	4	
Wg [adverbios interrogativos]	7	4	7	7	
Dd [artigos determinados]	3	2	5	5	3
Di [artigos indeterminados]	1	1	4	4	
E [demostrativos]	9	3	30	18	3
Cc [conxuncións coordinantes]	9	9	9	9	9
Cs [conxuncións subordinantes]	34	34	34	34	34
G [interrogativos-exclamativos]	8	4	82	16	4
I [indefinidos]	47	38	378	116	38
La0 [locucións adverbiais]	403				
Lp0 [locucións prepositivas]	95				
Lcc [locucións conxuntivas coordinantes]	3				
Lcs [locucións conxuntivas subordinantes]	84				
M [posesivos]	11	10	84	34	10
Nc [numerais cardinais]	93	49	323	63	79
No [numerais ordinais]	30	30	240	120	
P [preposicións]	52	52	52	52	52
Q [signos de puntuación]	20	20	20	20	20
Rt [pronomes tónicos]	21	17	86	25	25
Ra [pronomes átonos]	16	8	47	12	
T [relativos]	5	5	32	12	5
Y [interxeccións]	91	91	91	91	91
Za [abreviaturas]	875				
Zg [siglas]	2172				
Zs [símbolos]	146				

Dous apuntamentos máis:

1. O total de entradas é de 721 073 das cales son normativas 660 098 sen ter en conta ningún tipo de unidade gráfica multigramatical (contraccións, formas verbais con enclíticos, representacións da segunda forma do artigo, alomorfos pronominais etc.).

2. Das denominadas categorías periféricas (locucións, símbolos, abreviaturas e siglas) só é posible presentar, de momento, o número de entradas.

XIADA: lexicón médico. Elaboración do corrector ortográfico

A elaboración dun corrector ortográfico, especializado ou non, enmárcase no que se deu en chamar tecnoloxías da lingua, entendidas estas como unha orientación máis da lingüística computacional. En proxectos deste estilo adóitase establecer unha tripla clasificación que abrangue:

- As ferramentas, que engloban os sistemas informáticos que desenvolven ou resultan de utilidade para as aplicacións.
- Os recursos, que se corresponden cos datos lingüísticos con que se constrúen as aplicacións.
- As propias aplicacións resultantes, que se corresponden coa finalidade práctica proposta (no caso que nos ocupa un corrector ortográfico especializado).

Vista xa a ferramenta sobre a que desenvolver o traballo, corresponde agora tratar os outros dous elementos restantes. Os datos lingüísticos, como xa foi apuntado, tiráronse de historias clínicas reais procedentes do sistema sanitario galego, é dicir, supuña traballarmos con texto real o cal nos ía facilitar *a priori* o achegamento e a análise dos termos no seu contexto, observar a súa frecuencia, precisar a súa semántica, albiscar posibles relacións entre eles... Concretamente, desde o Servizo Galego de Saúde achegáronse nun principio informes de alta, de radioloxía e de anatomía patolóxica, ademais de comentarios de enfermaría. Porén, o feito de que o contido das historias clínicas estea protexido por lei ao máis alto nivel imposibilitou, tal e como sería desexable, levar á práctica calquera metodoloxía propia dun proxecto de lingüística de corpus, pois desde o Sergas remitíronse as historias clínicas reducidas informaticamente a simples listaxes de palabras.

Esta circunstancia sen ser un impedimento forte e sen minguar ou alterar a adecuación do produto final ás necesidades e prácticas lingüísticas das persoas destinatarias,

si se revelou como un pequeno hándicap que enlenteceu algún labor como pode ser a correcta identificación morfolóxica das formas por non se presentaren no seu contexto (pénsese, por exemplo, naqueles substantivos comúns con que se nomean principios activos e ao mesmo tempo son nomes comerciais de produtos), impediu o rexeitamento directo dalgunhas grallas e incorreccións (ex.: **urilización, *quedasin*), imposibilitou clasificarmos os termos segundo á súa especialidade sanitaria ou coñecer o estilo e/ou persoa en que están redactadas (de interese, por exemplo, para o tratamento e a inclusión de formas e tempos verbais) etc.

De resto, pense por un intre quen ler estas liñas na estrutura e mais no funcionamento dun corrector ortográfico. A primeira consiste basicamente na dispoñibilidade de dúas bases de datos ou lexicóns: un dicionario principal –o punto de partida dispoñible no CIRP resultado do proceso de etiquetaxe e lematización do Corga– e mais un dicionario secundario e/ou especializado que ou ben o propio usuario vai elaborando con base na súa competencia lingüística e a partir da experiencia que se deriva de traballar con textos dunha temática ou tipoloxía concreta, ou ben se acrecenta canto un paquete ou módulo integrado tal e como era a finalidade deste proxecto.

Con base nesta estrutura, o funcionamento dun corrector é relativamente sinxelo: cotexa cada unha das agrupacións de grafemas presentes nun texto con aquelas que ten dispoñibles nos seus dicionarios dando por boas todas aquelas que recoñece como propias e deténdose ou sinalando aquelas que lle resultan descoñecidas, para as cales deberá estar en disposición de achegar un número variable de alternativas. Por tanto, detrás dun corrector operan tres conceptos fundamentais: a palabra, o seu significante e os inventarios de palabras. Concretamente, a palabra tómase no sentido máis externo ou formalista como unha sucesión de grafemas, susceptibles de recibiren segmentos suprasegmentais (acentuación) e cuxas lindes veñen marcadas polos espazos ou signos de puntuación correspondentes. Dela interesa, xa que logo, unicamente o seu significante gráfico e fica á marxe toda a información unicamente recuperable e perceptible grazas ao contexto en que se usar (faltas de concordancias, desviacións semánticas etc.).

Aceptado isto, o corrector que se pretendía elaborar para o ámbito sanitario, ao igual que calquera outro, debía ser capaz de detectar os seguintes tipos de erros:

a) Os correspondentes ao ámbito da ortografía da palabra:

- Grafemas permutados
**gastorrrafia* → *gastrorrafia*
**hemilamincetomía* → *hemilaminectomía*
**inmunporoteína* → *inmunoproteína*

- Grafemas conmutados
**carvinoide → carcinoide*
**fleboestracción → fleboextracción*
**nasosexunal → nasoxexunal*
- Grafemas engadidos
**histeriopatía → histeropatía*
**postparto → posparto*
**subrrotuliano → subrotuliano*
- Grafemas suprimidos
**multitrasfundida → multitransfundida*
**nistámico → nistágmico*
**ostétrico → obstétrico*

b) Os correspondentes ao ámbito dos suprasegmentos gráficos, quer por adición incorrecta, quer por omisión, quer por incorrecta colocación do acento gráfico ou til e da diérese:

**agrafe → ágrafe*
**bucolingüais → bucolinguais*
**intraabdomináis → intraabdominais*
**lutéinico → luteínico*
**pinguecula → pingüécula*
**sublinguite → sublingüite*

c) Os correspondentes ao emprego das maiúsculas e das minúsculas:

**caliectAsia → caliectasia*
**eustaquio → Eustaquio*
**mingazzini → Mingazzini*

d) Os correspondentes á integridade formal da palabra, isto é, incorrectas fusións de palabras ou falsas segmentacións:

**unhacampimetría → unha campimetría*
**cervico tomía → cervicotomía*

A maiores, co fin de asegurar e acrecentar o valor normalizador do proxecto e, asemade, dotalo dunha «intelixencia lingüística» que o fixese ir alén da simple achega de alternativas á forma incorrecta con base en algoritmos de proximidade ortográfica, prevíronse dous tipos máis de correccións:

i) Un tratamento especial para aquelas palabras pertencentes ao ámbito médico (non necesariamente de carácter especializado) alleas á lingua galega, mais fortemente instaladas por influencia do castelán e cuxas formas gráficas en ambas as linguas difiren substancialmente. Sería o caso de termos pertencentes ao ámbito da anatomía, as posturas corporais, elementos protésicos, léxico propios de estados ou accións que describen unha sintomatoloxía dada etc.[13]

**ayuno → xaxún*
**blanquecino → esbrancuxado*
**calambrazo → descarga*
**cabestrillo → estribeira*
**chasquido → estalo*
**cuclillas → crequenas*
**desgarro → esgazadura*
**empeine → empeña*
**enrojecimiento → avermellamento*
**escozor → proído*
**esguince → escordadura*
**estornudar → esbirrar*
**hoyuelo → focha*
**mentón → queixelo*
**meñique → maimiño*
**pantorrilla → coxa*
**picor → proído*
**quejumbroso → queixoso*
**telilla → membrana*
**tobillera → nocelleira*

Constitúe unha vía que tendería a superar a vertente ortográfica do corrector para camiñar tamén cara á corrección morfolóxica (pénsese naqueles castelanismos cuxa equivalencia en galego varía de xénero).

ii) Quer por estarmos nun ámbito especializado e porque moita da documentación médica non sobarda os lindes da comunicación entre profesionais, quer polas propias circunstancias en que se xera a documentación (alto volume de documentos, necesidade de aforrar tempo...) o certo é que o traballo co corpus de termos resultantes revelou un amplo emprego de siglas[14]. A proposta que se ideou pasaba polo recoñecemento como sigla que o corrector debía facer daquelas sucesións de grafemas dispostos en maiúscula e en número variable (de 1 a 5 caracteres) con que se atopase e ofrecer o(s) seu(s) desenvolvemento posible(s). Os efectos prácticos que se perseguían eran dous:

- Tender a reducir o número das siglas sen desenvolver e, dalgún xeito, provocar unha reflexión acerca do seu uso/abuso. Repárese en que as máis delas xa non só é que teñan máis dun significado,

[13] O exposto mesmo tamén era unha vía para o tratamento dos anglicismos para os cales se acordase unha proposta en firme: *boobing ocular, clapping...*
[14] Para o seu tratamento resultou de enorme utilidade o volume *Diccionario de siglas médicas* publicado polo Ministerio de Sanidade e Consumo no ano 2003 e da autoría de Yetano Laruga e Alberola Cuñat. Dispoñible en: <http://www.msps.es/estadEstudios/estadisticas/docs/diccionarioSiglasMedicas.pdf>

TA:	temperatura ambiente
	tensión arterial
	tratamento actual
	traumatismo abdominal
APT:	alimentación parenteral total
	amnesia postraumática
	anxioplastia percutánea transluminal

senón que en ocasións a polisemia se produce dentro dunha mesma especialidade e/ou opera sobre un mesmo referente

IU:	infección urinaria
	insuficiencia urinaria
ACV:	accidente cerebrovascular
	accidente cardiovascular
AMI:	arteria mamaria interna
	arteria mesentérica inferior

- Contribuír a camiñar cara á elaboración dunha linguaxe médica verdadeiramente galega. Chegará o día en que se documenten siglas como as seguintes?

AX:	anestesia xeral?
BAX:	bo aspecto xeral?
IVE:	insuficiencia ventricular esquerda?
LSE:	lóbulo superior esquerdo?
MIE:	membro inferior esquerdo?
MIXE:	menisco interno do xeonllo esquerdo?

O material proporcionado polo Servizo Galego de Saúde, logo de lle aplicar o etiquetador de XIADA que actuou como corpus de exclusión, ficou reducido a unha listaxe de máis de 40 000 ítems sobre a que aínda houbo, antes de máis nada, que suprimir aqueles correspondentes a signos de puntuación, grallas, castelanismos, termos non pertencentes ao ámbito médico, pseudopalabras, falsas segmentacións etc. nas condicións sinaladas anteriormente.

De resto, chegados a este punto o procedemento foi moi semellante ao de calquera labor lexicográfico e/ou terminolóxico, fóra de non ter que realizar as correspondentes definicións, e condicionado polas fontes de consulta dispoñibles desde o principio e mais por aquelas que se ían descubrindo de xeito paralelo ao avance dos traballos.

Concretamente, o corpus de voces clasificouse en tres grupos segundo as posibilidades de documentalas en obras de referencia galegas e o grao de dificultade á hora de definir o seu significante correcto.

1. Termos presentes nas obras lexicográficas de consulta e, especialmente, no *Diccionario galego de termos médicos* publicado no ano 2002 pola antiga Dirección Xeral de Política Lingüística, avalado pola Real Academia de Medicina e Cirurxía de Galicia e coa colaboración do Seminario de Lexicografía da Real Academia Galega.

fenitoína	*ferrocinética*	*fibroadenoma*
fenotiacina	*ferropenia*	*fibroadiposo*
fenoxibenzamina	*ferroterapia*	*fibrocalcificado*
feocromocitoma	*festinación*	*fibrolipoma*

2. Termos non documentados na mencionada obra (nin no seu lemario nin no corpo dos artigos lexicográficos) mais cuxa estrutura morfolóxica se considerou plenamente correcta e acorde aos procedementos e regras de formación de palabras da lingua galega.

cinestésico	*citoplásmico*	*monocítico*
cirróxeno	*citoxenético*	*monohidrato*
citohistolóxico	*...*	*mucopurulento*
citolítico	*miomatoso*	*multifocal*
citopático	*monoclonal*	*multinodular*
citoplasmático	*monocorial*	*...*

3. Termos de dificultade media ou alta sobre os que se intensificou o traballo terminográfico puntual de estudo e contraste con outras linguas, ben porque a información sobre eles era máis ben escasa ou confusa, ben porque a información dispoñible nas fontes de consulta resultaba contraditoria, ben porque se trataba de anglicismos para os que se procurou unha proposta alternativa: *adstrinxente/astrinxente, decorticación/descorticación, diploe/díploe/diplóe, estadiamento/estadiaxe/estadificación, liposclerose/lipoesclerose,* prefixos *lumb-/lomb-* etc.

Chegados a este punto, cómpre especificar cales foron as fontes de consulta e de documentación para a fixación do corpus:

a) Para alén das obras lexicográficas de consulta habitual (que coinciden *grosso modo* coas sinaladas na táboa 1), traballouse arreo coas seguintes obras terminolóxicas:

Currás Fernández, C. / Dosil Maceira, B. (1999): *Diccionario de psicoloxía e educación* (Santiago de Compostela: Xunta de Galicia).

Daviña Facal, L. (2000): *Diccionario das ciencias da natureza e da saúde* (A Coruña: Deputación Provincial da Coruña).

Fraga García, H. (2004): *Glosario etimolóxico de termos anatómicos.* Edición bilingüe galego-castelán, castellano-gallego (A Coruña: Galinova Editorial).

Reyes Oliveros, Fco. / García González, C. (dirs.) (2003): *Diccionario galego de termos médicos* (Santiago de Compostela: Xunta de Galicia).

b) Totalmente dispoñible na rede, o Servizo Galego de Saúde dispón dun amplo e diverso catálogo de publicacións (vid. nota 2), maioritariamente en galego, estruturado en distintas coleccións ou epígrafes e con maior ou menor grao de especialización en función dos seus destinatarios. Sirva de mostra a seguinte escolma de títulos:

Braquiterapia no cancro de próstata localizado

Boletín de avaliación farmacoterapéutica de novos medicamentos: paliperidona

Boletín de farmacovixilancia: enfermidade de Alzheimer: reaccións adversas asociadas ó seu tratamento

Desfibrilación semiautomática externa

Estado actual da enfermidade celíaca: Guía de diagnóstico e tratamento de atención primaria a especializada

Guía de autoaxuda para deixar de fumar

Guía rápida. Técnico en transporte sanitario

Guía de práctica clínica da hipertensión arterial para atención primaria

Guía de práctica clínica. Cancro de pulmón e mesotelioma

Manual de boas prácticas para profesionais de tatuaxe, micropigmentación e piercing

Manual de cirurxía menor

Manual de intervención nas autoescolas. Alcohol, outras drogas e conducción

Guía de bo uso de absorbentes de incontinencia urinaria	*Plan galego de coidados paliativos*
Guía do usuario da unidade de hemodiálise	*Protocolo de diagnóstico e tratamento das infeccións de transmisión sexual*
Guía técnica do proceso de atención á fibromialxia	*Procedemento de xestión de gases medicinais no medio sanitario*

c) Consultáronse co Servizo de Terminoloxía de Galicia uns 150 termos, previamente traballados, e dos cales se obtivo unha cumprida resposta acerca de cal debe ser a súa forma gráfica en lingua galega. Exemplos: *banding ocular, puñopercusión, *quirofanillo, *reganancia, *steppage, *tenting...*

d) Internet: Ante a escaseza de fontes escritas e como non podía ser doutro xeito, a rede tornouse nunha ferramenta imprescindible e de consulta cotiá ao longo de todo o traballo, xa non só para a simple procura e documentación dos termos en motores de busca, senón tamén para se ir fornecendo dun soporte bibliográfico e de consulta que foi medrando conforme avanzaban os traballos. Estamos a falar de bases de datos terminolóxicas, artigos de referencia no ámbito da tradución e a linguaxe médica, páxinas oficiais con glosarios e listaxes oficiais de voces etc. Velaquí unha escolma de ligazóns:

- Axencia Española de Medicamentos e Produtos Sanitarios: <http://www.aemps.es/>
- Buscatermos: <http://www1.usc.es/buscatermos> (banco de datos terminolóxicos multilingüe da Universidade de Santiago de Compostela)
- Cercaterm:
- Clasificación Internacional de Doenzas: <http://eciemaps.mspsi.es/>
- Diccionario Ilustrado de Términos Médicos (Medciclopedia): <http://www.iqb.es/diccio/diccio1.htm>
- Dicionário Digital de Termos Médicos: <http://www.pdamed.com.br/diciomed/pdamed_0001_aa.php>
- InterActive Terminology for Europe (IATE): <http://iate.europa.eu> (base de datos terminolóxica multilingüe da Unión Europea)
- Vademecum.es: <http://www.vademecum.es/>

e) Artigos da revista *Panace@* da Asociación Internacional de Tradutores e Redactores de Medicina e Ciencias Afíns, e a monumental obra de Fernando Navarro baixo o título *Diccionario crítico de dudas inglés-español de medicina* (McGraw-Hill).

Con esta metodoloxía e estas fontes de consulta chegouse a un inventario de formas destinado a alimentar o sóftware de corrección que conta co seguinte número de formas:

A. Categorías léxicas

Substantivo [3894 lemas]	*Grupo*	*Raíces*	*Exemplo*
	G1	27	hemofílico, dixestólogo
	G6	10	rasurador, ex-fumador
	G29	3657	abdominoplastia, cavernoma
	G30	71	adefovir, biomarcador
	G32	72	bambuterol, antimitocondrial
	G33	4	centil, captopril
	G34	3	bifidez, inmadurez
	sen grupo	48	córtex, neurosífilis, fundus

Adxectivo [2819 lemas]	*Grupo*	*Raíces*	*Exemplo*
	G1	1593	afáquico, braquicefálico
	G6	42	citorredutor, metabolizador
	G29	266	acaricida, bacteriforme
	G30	248	bulbomedular, infrahiliar
	G32	660	abstinencial, drenábel
	G33	2	afebril, infantoxuvenil
	G34	8	fotomotriz, obturatriz
	sen grupo	2	isogrupo, cuádriceps

Verbo [133 lemas]	*Grupo*	*Raíces*	*Exemplo*
	V1	77	anticoagular, estenosar
	V2	1	cardioverter
	V3	8	multitransfundir, reexpandir
	Vi1	40	epitelizar, vasectomizar
	Vi2	1	subxacer
	Vi3	1	abducir
	Vi6	2	resituar, reavaliar
	Vi7	2	extruír, protuír
	Vi37	1	reintervir

B. Categorías periféricas

- 1974 siglas, que inclúen as efectivamente documentadas na listaxe de traballo inicial, así como as sinaladas como máis frecuentes por un facultativo con base na obra xa mencionada publicada polo Ministerio de Sanidade. Exemplos: *CAA, CAD, CAE, CAI, CAP, CAM...*
- 2580 nomes propios, que inclúen 363 antropónimos (os denominados epónimos médicos, isto é, termos médicos acompañados dun nome propio pertencente a un médico, investigador ou estudoso con que se nomean doenzas, síndromes, técnicas, probas, partes da anatomía etc.): *Creutzfeld-Jakob, Creutzfeld, Jakob, Crohn, Cronwall, Cushing, Dandy-Walker...;* e 2217 nomes de medicamentos: *Cafergot, Cafiaspirina, Cafinitrina, Calcial...*
- 62 abreviaturas: *Men., Mov., Oft., Pac., Pen., Pto...*
- 42 símbolos: CO_2, D_2, D_3, *Dx...*

Algunhas reflexións finais

1. O feito de que nun proceso de renovación e mellora tecnolóxica e procedemental das dimensións do proxecto IANUS se prevese a súa ligazón co emprego da lingua galega non pode deixar de ser considerado máis que positivo, afeitos como estamos a que non sempre a cuestión lingüística se inclúa dentro das prioridades ou das variables que cómpre atender en procesos de cambio como o que nos ocupa. Supuña, por unha banda, dar cumprimento ás previsións que se planificaran e aprobaran con anterioridade e, pola outra, tomaba como base un proxecto de lingüística aplicada xa en marcha con que poder xerar unha ferramenta práctica –un corrector ortográfico especializado– para o uso do persoal sanitario. A experiencia acumulada con ferramentas deste tipo revela a súa utilidade e a boa acollida por parte dos usuarios. Así, nos procesadores de texto que adoito manexamos os correctores ortográficos que neles actúan cumpren a dupla misión de serviren de garantes da corrección ortográfica dos escritos ao sinalaren os sempre posibles erros de mecanografado e, asemade, de contribuíren á formación lingüística de quen os emprega ao lle proporcionaren alternativas a aquelas formas que considera incorrectas. Mesmo se podería engadir unha terceira función pois, cando actúan por exceso, suscitan na persoa que está a escribir unha reflexión metalingüística que a obriga a ter que decidir se se fía dos seus coñecementos ou do que lle alerta o corrector (ex.: topónimos estranxeiros, neoloxismos, formas derivadas...).

Así pois, cabe pensar que implementar un corrector ortográfico para o seu emprego por parte do persoal sanitario que conte coa terminoloxía que con certeza se está a empregar nos centros hospitalarios e de saúde ha ser unha medida que só pode traer beneficios para dispor dunha documentación clínica en galego e de calidade. Porén, dito isto entendemos que sería iluso confiarmos unicamente na elaboración dun corrector o aumento do uso do galego e a mellora da formación lingüística do persoal. Cumpriría, por exemplo, acompañalo da correspondente formación (por exemplo, canda a que se deu e está a dar na Fundación da Escola Galega da Administración Sanitaria respecto ao proxecto IANUS) e dunhas normas lingüísticas de uso na liña das recollidas na circular 11/2005.

Polo demais, lamentablemente co transcorrer do tempo as previsións iniciais que semellaban apuntar na liña dunha introdución da lingua galega parella á da implantación de IANUS no sistema sanitario galego non se están a cumprir. A retroalimentación entre o Sergas e o equipo lingüístico deste proxecto non funcionou desde o principio por causas descoñecidas e só imputables á primeira parte, de modo que anos despois de iniciado o proxecto este non só non conta aínda co corrector en galego implementado no sistema, senón que aínda non se puideron facer as primeiras e pertinentes probas sobre o seu funcionamento nin comprobar a adecuación do corpus ofrecido nin perfilar algunhas das decisións tomadas[14].

Como queira que a implantación de IANUS non está aínda rematada por seren de tal magnitude os cambios que comporta no traballo diario do/a profesional, cabe agardar que á vista das numerosas deficiencias –non só técnicas– que se lle sinalan (Veras Castro) a posibilidade de vermos o corrector en funcionamento poida algún día chegar a ser unha realidade.

2. Con base nestes tres exemplos

gl	*frémito*	gl	*linfanxite*	gl	*meniscectomía*
pt	*frémito, frêmito*	pt	*linfangite*	pt	*meniscectomia*
es	*frémito*	es	*linfangitis*	es	*meniscectomía*

[14] Pénsese, por exemplo, que á hora de incluír as formas verbais se adoptou a determinación de achegar todo o paradigma verbal, o cal seguramente resulte excesivo e recargue innecesariamente o sistema. Da conxugación verbal resultan en xeral 83 formas por modelo que se corresponden con 74 etiquetas, das cales son formas únicas 57; 23 combinacións cun só clítico producen 1311 formas normativas; mentres que as 58 combinacións de 2 clíticos producen por verbo 3306 formas.

ca	*fremitus, freminent*	ca	*linfangitis*	ca	*meniscectomia*
fr	*frémissement*	fr	*lymphangite*	fr	*méniscectomie*
it	*fremito*	it	*linfoangite, linfangite*	it	*meniscectomia*
en	*fremitus, thrill*	en	*lymphangitis*	en	*meniscectomy*

e mais lembrándonos daquela solicitude de probas radiolóxicas, localizada nun hospital de Vigo, que circulou pola rede e na que se solicitaba a súa tradución ao castelán[15], entendemos que a seguinte pregunta que cómpre facérmonos é se os baixos índices de emprego da lingua galega no ámbito sanitario obedecen exclusivamente a un problema de corpus? É realmente un problema de falta de terminoloxía ou de que esta resulte moi complicada ou confusa? Repárese en que boa parte da terminoloxía médica está integrada por formantes cuxas orixes están na lingua latina e, sobre todo, na grega e cuxa presenza non é nin tan sequera panrománica, senón internacional.

Sen negarmos que son necesarias obras como as presentes nesta mesa de traballo e o propio corrector polas razóns xa expostas, sería interesante non esquecermos e pórmos en valor o substrato común presente en infinidade de linguas. Fan falta vocabularios (lonxe estamos de ter aínda un completo atlas anatómico ou un inventario de material cirúrxico en lingua galega por citarmos unicamente dous exemplos), precísanse obras de consulta que resulten de utilidade para formar os termos, estratexias con que combater os anglicismos crecentes que dan nome a novas técnicas, dispositivos e achados, mais fai falta tamén a formación do persoal sanitario. Unha decidida introdución da lingua galega no ámbito sanitario debería ser aproveitada coa intención de a mellorar, de facela cumpridora dos principios de precisión, clareza, exactitude e corrección dos que agora, segundo estudosos da materia desde a perspectiva do castelán, non cumpre (velaí está a experiencia que nos proporciona a linguaxe administrativa, en que a adopción por parte das máis das administracións galegas da lingua galega como lingua de relación co administrado e inter-administracións foi parella á do proceso modernizador e renovador da linguaxe administrativa).

Paralelamente, sería necesario un efectivo cumprimento das previsións regulamentarias previstas, ben como unha concienciación lingüística do persoal sanitario, para o cal entendemos o Plan xeral de normalización da lingua galega como un bo punto de partida apenas aínda transitado.

[15] Véxase <http://www.vieiros.com/nova/61933/algun-problema-coa-lingua-na-consulta-medica>.

Anexo I: os grupos derivacionais do sistema de XIADA

identificador e representante	terminación grupo	subetiqueta	seguinte	normativa 2003
G1 [NEN-O]	o	ms		si
	a	fs		si
	os	mp		si
	as	fp		si
G2 [IRM-ÁN]	án	ms		si
	á	fs		si
	áns	mp		si
	ás	fp		si
	anas	fp		non
	ana	fs		non
G3 [PILLAB-ÁN]	ás	fp		non
	án	ms		si
	ana	fs		si
	áns	mp		si
	anas	fp		si
	á	fs		non
G4 [LANCAR-AO]	ao	ms		si
	á	fs		si
	aos	mp		si
	ás	fp		si
G5 [INGL-ÉS]	és	ms		si
	esa	fs		si
	eses	mp		si
	esas	fp		si
G6 [AQUEL]		ms		si
	a	fs		si
	es	mp		si
	as	fp		si
G7 [LAMB-ÓN]	ón	ms		si
	ona	fs		si
	óns	mp		si
	onas	fp		si
G8 [CAMPI-ÓN]	ón	ms		si
	oa	fs		si
	óns	mp		si
	oas	fp		si
	ona	fs		non
	onas	fp		non

G9 [ALCALD-E]	es	mp		si
	esas	fp		si
	e	ms		si
	esa	fs		si
G10 [BAILAR-ÍN]	ín	ms		si
	ina	fs		si
	íns	mp		si
	inas	fp		si
G11 [PROFET-A]	a	ms		si
	isa	fs		si
	as	mp		si
	isas	fp		si
G12 [XOGRAR]		ms		si
	esa	fs		si
	es	mp		si
	esas	fp		si
G13 [EST-E]	e	ms		si
	a	fs		si
	es	mp		si
	as	fp		si
G14 [CABR-ÓN]	ón	ms		si
	a	fs		si
	óns	mp		si
	as	fp		si
G15 [XUD-EU]	eu	ms		si
	ía	fs		si
	eus	mp		si
	ías	fp		si
G16 [GAL-O]	o	ms		si
	iña	fs		si
	os	mp		si
	iñas	fp		si
G17 [R-EI]	ei	ms		si
	aíña	fs		si
	eis	mp		si
	aíñas	fp		si
G18 [TSAR]		ms		si
	ina	fs		si
	es	mp		si
	inas	fp		si
G19 [HERO-E]	e	ms		si
	ína	fs		si
	es	mp		si
	ínas	fp		si

G20 [GR-OU]	ou	ms		si
	úa	fs		si
	ous	mp		si
	úas	fp		si
G21 [ESPAÑO-L]	l	ms		si
	la	fs		si
	is	mp		si
	las	fp		si
G22 [RAPA-Z]	z	ms		si
	za	fs		si
	ces	mp		si
	zas	fp		si
G23 [AV-Ó]	ó	ms		si
	oa	fs		si
	ós	mp		si
	oas	fp		si
G24 [SACERDOT-E]	e	ms		si
	isa	fs		si
	es	mp		si
	isas	fp		si
G25 [VAMPIR-O]	o	ms		si
	esa	fs		si
	os	mp		si
	esas	fp		si
G26 [BAR-ÓN]	ón	ms		si
	onesa	fs		si
	óns	mp		si
	onesas	fp		si
G27 [EMPERA-DOR]	dor	ms		si
	triz	fs		si
	dores	mp		si
	trices	fp		si
G28 [ACT-OR]	or	ms		si
	riz	fs		si
	ores	mp		si
	rices	fp		si
G29 [ALEGRÍA]		s		si
	s	p		si
G30 [MAR]		s		si
	es	p		si
G31 [CORT-ÉS]	és	s		si
	eses	p		si
G32 [XENIA-L]	l	s		si
	is	p		si

G33	il	s		si
[ED-IL]	ís	p		si
G34	z	s		si
[AUDA-Z]	ces	p		si
G35	ás	s		si
[COMP-ÁS]	ases	p		si
G36	ús	s		si
[AUTOB-ÚS]	uses	p		si
G37	ás	p		non
[RE-AL]	al	s		si
G38	as	fp		si
[AMB-OS]	os	mp		si
G39	unha	fs		si
[ALG-ÚN]	unhas	fp		si
	ún	ms		si
	úns	mp		si
G40		ms		si
[BO]	a	fs		si
	s	mp		si
	as	fp		si
G41		ms		si
[UN]	ha	fs		si
	s	mp		si
	has	fp		si
G42	os	ma		si
[TRESCENT-OS]	as	fp		si
G43	ísim	s	G1	si
G44	u	ms		si
[CR-U]	us	mp		si
	úa	fs		si
	úas	fp		si
G45	eu	ms		si
[T-EU]	eus	mp		si
	úa	fs		si
	úas	fp		si
G46	n	s		si
[CA-N]	ns	p		si
	s	p		non

Bibliografía

Álvarez, R. / Regueira, X. L. / Monteagudo, H. (1989[2]): *Gramática Galega* (Vigo: Galaxia).

Álvarez, R. / Xove, X. (2002): *Gramática da lingua galega* (Vigo: Galaxia).

Álvarez, R. (1994): «Secuencias de pronomes átonos en galego moderno» en Lorenzo, R. (ed.), *Actas do XIX Congreso de lingüística e filoloxía románicas,* vol. VI*:* 247-265 (A Coruña: Fundación Pedro Barrié de la Maza).

Ariza García, A. / Tapia Poyato, A. M.ª (1997): «El corrector ortográfico y la presentación del texto escrito» en *Cauce, Revista de Filología y su Didáctica,* 20-21: 375-412. Dispoñible en: <http://cvc.cervantes.es/literatura/cauce/pdf/cauce20-21/cauce20-21_22.pdf> [Última consulta: 6/6/2011].

Bray, T. / Jean, P. / Sperberg-McQueen, C. M. (eds.): *Extensible Markup Language (XML)* 1.0 (Fifth Edition). Dispoñible en: <http://www.w3.org/TR/REC-xml> [Última consulta: 6/6/2011].

Burnage, G. / Dunlop, D. (1993): «Encoding the British National Corpus», en Aarts, J. / Haan, P. de / Oostdijk, N. (eds.): *English Language Corpora: Design, Analysis and Exploitation* (Amsterdam: Rodopi).

Burnard, L. (1997): «The Text Encoding Initiative's. Recommendations for the Encoding of Language Corpora: Theory and Practice» Dispoñible en: <http://users.ox.ac.uk/~lou/wip/Soria/> [Última consulta: 6/6/2011]

Burnard, L. / Bauman, S. (eds.) (2009): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Dispoñible en: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> [Última consulta: 6/6/2011]

Carballeira Anllo, X. M. (coord.) (2000): *Gran Diccionario Xerais da Lingua* (Vigo: Edicións Xerais de Galicia).

— (2009): *Gran Dicionario Xerais da Lingua.* 2 volumes (Vigo: Edicións Xerais de Galicia).

Chacón Calvar, R. / Rodríguez Alonso, M. (1993): *Diccionario crítico de dúbidas e erros da lingua galega* (Sada: Ediciós do Castro).

Costa Casas, X. X. / González Refoxo, M.ª A. / Morán Fraga, C. C. / Rábade Castiñeira, X. C. (1988): *Nova Gramática para a aprendizaxe da língua* (A Coruña: Vía Láctea).

Department of Computer Science (Vassar College) / Equipe Langue et Dialogue (LORIA/CNRS) (2008): *Corpus Encoding Standard for XML*. Dispoñible en: <http://www.xces.org> [Última consulta: 6/6/2011].

Díaz Regueiro, M. (1992): *Os verbos galegos* (Santiago de Compostela: Consellería de Educación e Ordenación Universitaria. Dirección Xeral de Política Lingüística).

Domínguez Noya, E. (2008): «O Corpus de Referencia do Galego Actual *(CORGA)*: presente e futuro» en González Seoane, E. / Santamarina, A. / Varela Barreiro, X. (eds.), *A lexicografía galega moderna: Recursos e perspectivas:* 139-151 (Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega).

Domínguez Noya, E. / Barcala Rodríguez, Fco. M. / Molinero, M. A. (2009): «Avaliación dun etiquetador automático estatístico para o galego actual: Xiada» en *Cadernos de Lingua*, 30/31: 151-193.

Ferreiro, M. (1999): *Gramática histórica galega. I. Fonética e Morfosintaxe* (Santiago de Compostela: Edicións Laiovento).

Freixeiro Mato, X. R. (2000): *Gramática da lingua galega. II. Morfosintaxe* (Vigo: A Nosa Terra).

García, C. e M. González González (dirs.) (1997): *Diccionario da Real Academia Galega* (A Coruña: Fundación Pedro Barrié de la Maza / Real Academia Galega).

Gómez Guinovart, X. (2006): «Tecnoloxías da lingua galega e normalización lingüística» en VV. AA. *Xornadas sobre Lingua e Usos. Lingua e Investigación:* 79-91 (A Coruña: Servizo de Normalización Lingüística-Universidade da Coruña). Dispoñible en: <http://www.udc.es/snl/documentospdf/Libro_Lingua_Investigacion.pdf> [Última consulta: 6/6/2011].

González González, M. e A. Santamarina Fernández (coords.) (2004): *Vocabulario Ortográfico da Lingua Galega* (Santiago de Compostela-A Coruña: Instituto da Lingua Galega-Real Academia Galega).

González Rei, B. (2004): *Ortografía da lingua galega* (A Coruña: Galinova Editorial).

Instituto da Lingua Galega / Real Academia Galega (1993[11]; 1997[16]; 2003[18]): *Normas ortográficas e morfolóxicas do idioma galego* (Santiago de Compostela: ILG / RAG).

Ledo Cabido B. (dir.) (2004): *Dicionario de galego* (Vigo: Ir Indo).

Leech, G. / Wilson, A. (1994): «Morphosyntactic Annotation», Draft-Work In Progress, EAGLES Document EAG-CSG/IR.T3.1. en N. Calzolari e J. M. McNaught (eds.), *EAGLES Interim Report* EAG-EB-IR-2.

— (1996): *Recommendations for the Morphosyntactic Annotation of Corpora.* EAG-TCWG-MAC/R. Dispoñible en: <http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html> [Última consulta: 6/6/2011]

Monachini, M. / Calzolari, N. (eds.) (1996): *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*. EAG-CLWG-MORPHSYN/R. Dispoñible en: <http://www.ilc.cnr.it/EAGLES96/morphsyn/morphsyn.html> [Última consulta: 6/6/2011].

Monachini, M. / Calzolari, N. (1999): «Standardization in the Lexicon» en Halteren, H. van (ed.) *Syntactic wordclass tagging:* 149-174 (Dordrecht: Kluwer Academic Publishers).

Nicolás Rodríguez, R. (1993): *Diccionario dos verbos galegos* (Vigo: Edicións do Cumio).

Rojo Sánchez, G. /López Martínez, M. / Domínguez Noya, E. / Barcala Rodríguez, Fco. M. (2010): *Léxico de XIADA: Etiquetador/Lematizador do Galego Actual*. Dispoñible en: <http://corpus.cirp.es/xiada/descargas.html> [Última consulta: 6/6/2011].

Veras Castro, R. (2010): «IANUS: Luces e sombras. Así non vai ben» en *Cadernos de atención primaria,* 17, vol. 1: 87-88. Dispoñible en: <http://www.agamfec.com/pdf/CADERNOS/VOL17/vol_1/13_Espazo_para_o_debate_2.pdf> [Última consulta: 6/6/2011].

Schiller, A. / Karttunen, L. (1999): «Lexicons for Tagging» en Halteren, H. van (ed.) *Syntactic wordclass tagging:* 135-147 (Dordrecht: Kluwer Academic Publishers).