Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega

José Luis Aguirre Moreno <jlagui@uvigo.es>
Nuria Andión Rodríguez <andion@uvigo.es>
Xavier Gómez Guinovart <sli@uvigo.es>
Seminario de Lingüística Informática
Universidade de Vigo

Resumen. En este trabajo se examinan algunos aspectos del etiquetado lingüístico de un corpus técnico de informática en lengua gallega, en lo que respecta a cuestiones ortográficas, léxicas y morfosintácticas. En primer lugar, presentamos la características del corpus analizado y algunas de las aplicaciones de su procesamiento. A continuación, mostramos las técnicas empleadas en su anotación morfosintáctica, centrándonos en la discusión de nuestra propuesta de etiquetario y en el esquema de codificación. Por último, presentamos una aproximación a los problemas específicos que plantea la anotación léxica, terminológica y ortográfica del corpus.

1. Introducción

En esta comunicación presentaremos algunos aspectos del etiquetado lingüístico basado en TEI/SGML del CLUVI ("Corpus Lingüístico da Universidade de Vigo"), un corpus textual versátil de lengua gallega contemporánea oral y escrita, diseñado y compilado con el fin de realizar diversos estudios de Lingüística Aplicada en los ámbitos de la Filología Gallega, de la Filología Inglesa, de la Traducción, de la Lingüística Computacional y de la Sociolingüística. En su estado actual de desarrollo, contiene alrededor de 4.000.000 de palabras y está constituido por cinco subcorpus específicos interrelacionados: dos orales y tres escritos. La sección oral del CLUVI está formada por un corpus de habla espontánea bilingüe gallego-castellano de unas 500.000 palabras (Rodríguez Yáñez et al. 2001), y por un corpus de lengua oral de los medios de comunicación audiovisuales gallegos de una extensión semejante. La sección escrita, por su parte, consta del corpus paralelo TECTRA de textos literarios inglés-gallego (Álvarez Lugrís 2001) y de dos corpus técnicos, el corpus XIGA de textos sobre informática en galego y el corpus LEGA de textos jurídico-administrativos en gallego, estos tres de alrededor de 1.000.000 de palabras cada uno. Cada subcorpus del CLUVI posee sus propias características lingüísticas y exige un determinado grado de procesamiento lingüístico-computacional específico y adaptado a sus rasgos lingüísticos distintivos. Sin embargo, el desarrollo coordinado de las técnicas de análisis necesarias para la explotación lingüística de los distintos subcorpus que componen el CLUVI permite un aprovechamiento racional de los recursos de investigación, favorece el desarrollo de técnicas transversales de análisis y potencia la explotación pluridisciplinar de los corpus de trabajo, que se lleva a cabo conjuntamente por el Seminario de Lingüística Informática[1] y por el Seminario de Sociolingüística[2] de la Universidade de Vigo.

2. Características del corpus XIGA

En esta exposición, examinaremos algunos de los problemas de la anotación lingüística del CLUVI, en lo que respecta a cuestiones ortográficas, léxicas y morfosintácticas, mostrando las soluciones adoptadas con referencia al corpus XIGA, aunque extensibles en su mayor parte al resto de subcorpus. El corpus XIGA, compilado en el Seminario de Lingüística Informática de la Universidade de Vigo, es un corpus textual de gallego que recoge material escrito de diversos ámbitos y registros en el campo de la informática, como artículos periodísticos, manuales y textos de ayuda, menús y mensajes de programas, discusiones y noticias ("news") en grupos de Internet, y textos académicos, escolares e informativos. Desde un punto de vista terminográfico, la explotación del XIGA

[1] http://www.uvigo.es/webs/sli
[2] http://www.uvigo.es/webs/ssl

facilita la elaboración de propuestas terminológicas normativas ajustadas a la realidad lingüística plasmada en los textos (Gómez Guinovart et al. 2001). Por otra parte, desde un punto de vista lingüístico descriptivo, el corpus XIGA permite la realización de investigaciones lingüísticas fundamentadas sobre el uso social de las propuestas terminológicas, ya que muestra de manera empírica la dirección de la selección léxica, es decir, las vacilaciones y preferencias actuales en la selección de términos informáticos en la lengua gallega escrita contemporánea (Gómez Guinovart & Lorenzo Suárez 2001). Por último, desde el punto de vista del procesamiento del lenguaje, el corpus XIGA ofrece un material textual de gallego muy adecuado para la extracción automática monolingüe de terminología y para la caracterización lingüístico-cuantitativa del lenguaje de especialidad propio de esta disciplina.

3. Etiquetado morfosintáctico

Un primer paso, imprescindible para la explotación del CLUVI, consiste en analizar los textos e incorporar en ellos información lingüística. La anotación morfosintáctica consiste en indicar para cada palabra su categoría gramatical junto con sus rasgos gramaticales más relevantes.

3.1. Etiquetario

Antes de abordar la anotación morfosintáctica de un corpus se debe decidir el conjunto de etiquetas que se podrán asignar a las palabras, es decir, se ha de diseñar un etiquetario o conjunto de marcas que describan los fenómenos lingüísticos relevantes para las lenguas que se estudian, por lo que respecta a las categorías gramaticales o partes de la oración (Morel et al. 1997). Los criterios que seguimos en la elaboración de este etiquetario tratan de hallar un compromiso entre tres aspectos, a menudo contradictorios: adecuación teórica de las categorías propuestas, exhaustividad y posibilidad de implementación en el etiquetado automático (o semi-automático). Otro aspecto fundamental en la elaboración de recursos lingüísticos computacionales es el de su posible reutilización, de modo que se pueda compartir información con proyectos actuales y futuros. Para ello, resulta imprescindible seguir las directrices de los estándares más aceptados en el campo de la anotación de corpus. En nuestro trabajo adoptamos el estándar TEI/SGML (Sperberg-McQueen & Burnard 1994) por lo que respecta al lenguaje de marcación y las propuestas de EAGLES (1996) en lo relativo a las categorías gramaticales y rasgos morfosintácticos que conviene distinguir.

Para la creación del etiquetario, nos basamos principalmente en la descripción gramatical de la lengua gallega recogida en Álvarez et al. (1986), complementándola en algunos aspectos con la ofrecida en Freixeiro (1998-2000). La adopción de estas descripciones gramaticales, en las que se emplea una terminología lingüística clásica, obliga a tomar en consideración algunas categorías que difieren de las propuestas por los estándares. Un ejemplo relevante es el referido al uso adjetivo o pronominal de las formas de los pronombres. Así, en Álvarez et al. (1986: 157) la categoría de pronombre es una clase semántica caracterizada por la ausencia de "significación constante, concreta e determinada", y que puede realizar, en la frase nominal, funciones de núcleo o de adyacente. En la práctica del etiquetado, esto implica que un demostrativo puede recibir siempre la etiqueta de "Pronombre demostrativo", tanto si aparece en función pronominal como si lo hace en función adjetiva, lo que en ciertas aplicaciones puede resultar poco adecuado.

En otras ocasiones, por el contrario, debemos incluir categorías que, aunque no existen de manera explícita en las gramáticas de la lengua, son necesarias para cubrir todos los fenómenos con los que nos encontramos en los textos reales, así como para dar cuenta de algunos elementos que resultan de gran utilidad para la anotación automática. Así, por ejemplo, tomamos de EAGLES (1996) categorías como "Puntuación" o como "Residual", que incluye las siguientes anotaciones:

REX	Palabra extranjera
RFO	Fórmula
RSI	Símbolo
RAC	Acrónimo
RAB	Abreviatura
RSC	Sin clasificar

3.2. Esquema de codificación

Junto con el conjunto de etiquetas, es preciso documentar los criterios que se siguen para la adscripción de las palabras a una u otra

categoría, en lo que se suele denominar *esquema de codificación* o manual del codificador. Los estándares mencionados no especifican la manera en que se ha de elaborar este esquema, pues no se puede decidir de antemano, pero lo ideal es que la explicación sea lo suficientemente exhaustiva y detallada como para que si lo consultan varias personas y aplican el esquema a un mismo texto asignen las mismas etiquetas a todas las palabras analizadas. Para ello, tratamos de delimitar claramente las fronteras entre las distintas categorías y de dar soluciones explícitas a los fenómenos que pueden presentar problemas por encontrarse en un punto intermedio entre dos o más categorías. Asimismo, confeccionamos listas lo más exhaustivas posible con las palabras integrantes de categorías semi-cerradas, es decir, de aquellas que están integradas por un número limitado de palabras, como adverbios, preposiciones y conjunciones. En estos casos, hemos considerado de utilidad relacionar aquellas formas que pueden pertenecer a más de una de estas tres categorías, junto con ejemplos y explicaciones de su uso que pueden servir de ayuda para la anotación de los textos, pues este aspecto se presta a confusión.

En nuestro etiquetario distinguimos las siguientes categorías genéricas: Nombre, Adjetivo, Pronombre, Verbo, Artículo, Interjección, Conjunción, Adverbio, Preposición, Puntuación y Residual. Describiremos a continuación las características generales del etiquetario.

En los nombres se distingue entre comunes y propios, y en cada uno de ellos se tienen en cuenta los rasgos de género y número. No se incluye ninguna característica de tipo semántico (si son de persona, de medida, etc.), ni se etiquetan los nombres propios compuestos (por ejemplo, "Santa Euxenia de Ribeira") como tales, sino que cada palabra dentro del nombre llevará su etiqueta correspondiente. Tampoco prevemos la existencia de nombres sin género o sin número (aunque al etiquetar nos encontremos palabras como "París" o "Internet", que resultan difíciles de clasificar).

En cuanto a la subespecificación léxica de género o número, es decir, aquellas palabras que formalmente no presentan oposición de género (*cantante*) o de número (*martes*), no reciben ninguna etiqueta que las distinga de las demás. Este criterio es aplicable a todas las categorías. Simplemente se les asigna el género o número que presenten en cada contexto. En este sentido nuestro etiquetario tiene una clara orientación a la anotación de corpus, frente a la marcación de información léxica. De manera análoga, no incluimos etiquetas que representen información sobre género y número para las palabras de las clases gramaticales cerradas que no tengan formas diferenciadas. Por ejemplo, en algunos etiquetarios hay una marca diferente para *yo* en "yo soy bueno" y en "yo soy buena" (además de otra para cuando no está especificado, como en "yo soy taxista"). De este modo, se multiplica el número de etiquetas necesarias, y, sobre todo, se generan ambigüedades difíciles de resolver por los sistemas de etiquetado automático, mientras que la información gramatical añadida consideramos que no es relevante.

En los pronombres personales distinguimos formas rectas, oblicuas, ligadas y átonas. Aunque estos rasgos no existen en los estándares internacionales, preferimos esta descripción porque se ajusta más a las necesidades de la gramática gallega. Más adelante, cuando ponemos en relación el etiquetario del gallego con el del inglés, para su aplicación al corpus TECTRA, a cada una de las etiquetas de las formas no rectas le asignamos el (los) posible(s) caso(s): acusativo, dativo u oblicuo[3].

Hay algunas características propias del gallego que se deben tener en cuenta a la hora de afrontar la anotación morfosintáctica de corpus en esta lengua. Una de ellas es el gran número de posibles contracciones de las llamadas palabras gramaticales (o de categorías cerradas). La primera consecuencia de ello es que el número de etiquetas del repertorio aumenta en comparación con los de otras lenguas. En cuanto a su tratamiento, en nuestro esquema de codificación mantenemos las contracciones sin dividir en la segmentación, y asignamos una etiqueta a la palabra ortográfica (secuencia de caracteres delimitada por espacios en blanco). Estas marcas están expresamente descritas en el etiquetario, y

[3] Una parte de nuestra investigación actual, que no desarrollamos en esta comunicación, consiste en hallar un sistema de codificación que potencie la explotación de los corpus paralelos y, en particular, del corpus TECTRA de traducciones inglés-gallego, mediante la correspondencia entre las anotaciones morfosintácticas basadas en distintos conjuntos de etiquetas.

además siempre se forman de la misma manera: "etiqueta de la primera palabra", seguida de "&" y seguida de "etiqueta de la segunda palabra". Por ejemplo, a la contracción de la preposición *en* y el pronombre personal masculino singular de 3ª persona *el (nel)* le corresponde la etiqueta PREP&PPMS3. Ello es posible porque todas las contracciones se dan entre palabras de categorías cerradas.

Otro fenómeno por el que no coinciden las palabras léxicas y las ortográficas es el de los enclíticos, es decir, los pronombres que aparecen ligados con los verbos: *díxome* ("dixo" + "me"), *cóntase* ("conta" + "se"). En gallego es mucho más frecuente que, por ejemplo, en castellano, porque todas las personas de todos los tiempos aceptan enclíticos; no es infrecuente la aparición de dos enclíticos en una forma: *acercóuselle* ("acercou" + "se" + "lle"). A su vez, los enclíticos pueden estar contraídos: *díxomo* ("dixo" + "me" + "o"). Como podemos ver, las combinaciones que surgen son muy numerosas. Por ello, en nuestro etiquetario no están relacionadas expresamente todas, sino que adoptamos la misma convención para la creación de la etiqueta que para las contracciones, es decir, la etiqueta de una forma con enclítico(s) está compuesta de la etiqueta de la palabra principal más las correspondientes a las palabras ligadas, con signos "&" para separarlas. Por ejemplo, a *acercóuselle* le correspondería la marca VIPES3&PPS3AR&PPS3AD (verbo – indicativo – pretérito perfecto - singular – 3ª persona & pronombre – personal – singular – 3ª persona – átono – reflexivo & pronombre – personal – singular – 3ª persona – átono – dativo).

Un fenómeno gramatical parecido es el de la segunda forma del artículo determinado. Consiste en la contracción de ciertas palabras que acaban en *-r* o en *-s* (entre las que se incluyen los infinitivos y las formas verbales de la segunda persona) y los artículos *o / a / os / as*, que se convierten en los alomorfos *lo / la / los / las*, y se unen a la anterior palabra mediante un guión haciendo desaparecer su última consonante: *bebe-lo leite* ("beber" + "lo") (cast. "beber la leche"). La etiqueta que asignamos a esta palabra es la correspondiente al artículo seguida de "-2": la forma *bebe-lo* se etiquetaría VINF&ARDMS-2. La segunda forma del artículo puede aparecer junto con otros enclíticos: en *gústalle-lo leite* ("gusta" + "lles" + "lo") (cast. "les gusta la leche"), *gústalle-lo* recibe la marca VIPRS3&PPP3AD&ARDMS-2 (verbo – indicativo – presente – singular – 3ª persona & pronombre – personal – plural – 3ª persona – átono – dativo & artículo – determinado – masculino – singular – 2ª forma).

También es importante el carácter flexivo de la lengua gallega. Varían los nombres y adjetivos en género y número, y de los verbos recogemos los rasgos modo, tiempo, número y persona. La flexión verbal genera un total de 84 etiquetas de formas personales, más 6 de formas impersonales (este número difiere de otras descripciones, como la recogida en Vilares et al. (1998), porque en nuestro etiquetario incluimos una marca distinta para la segunda persona de cortesía de singular y de plural de cada tiempo verbal, y para las cuatro combinaciones de género y número de los participios).

Hay otros fenómenos que hacen que no coincidan las palabras ortográficas con las palabras léxicas, pero de manera opuesta a los anteriores, de modo que una categoría gramatical está formada por más de una palabra. Es el caso de las locuciones. En nuestro esquema contemplamos principalmente las locuciones prepositivas, conjuntivas y adverbiales. En su codificación optamos por una solución en la línea de lo propuesto por Sampson (1995). Mantenemos la separación de las palabras en la segmentación; si la palabra forma parte de una locución, se le asigna la etiqueta correspondiente a su categoría si apareciera de manera aislada, más el signo "&", la categoría de la locución, el número de palabras que la integran y el número de orden de ese elemento dentro de ella. Por ejemplo, en la locución prepositiva *antes de que*, *antes* aparecerá como ADV&PREP31, *de* como PREP&PREP32, y *que* como CONX&PREP33.

En la siguiente tabla (Tabla 1) se puede observar el número y distribución de las etiquetas morfosintácticas que proponemos para la lengua gallega. Distinguimos en este resumen las contracciones expresamente descritas de las otras formas. No incluimos las combinaciones generadas por las locuciones, ni las generadas por los enclíticos:

CATEGORÍAS		ETIQUETAS	
		Sin contrac.	Contrac-ciones
Nombres		8	
Adjetivos		12	
Pronombres	Formas tónicas	14	
	Formas oblicuas	6	
	Formas ligadas	5	
	Contracciones de pronombres personales		10
	Formas átonas	24	
	Contracciones de 2 pron. Personales átonos		24
	Demostrativos	18	
	Contracciones de pronombres demostrativos		18
	Posesivos	32	
	Indefinidos	5	
	Contracciones de pronombres indefinidos		2
	Numerales	12	
	Relativos e interrogativos	2	
Artículos		8	
Contracciones de artículos			24
Interjecciones, conjunciones, adverbios y preposiciones		4	
Puntuación		19	
Residual		6	
TOTALES		175	78
Verbos. Formas personales		84	
Verbos. Formas impers.		6	
TOTAL sin contracciones		265	
TOTAL con contracciones		343	

Tabla 1. Resumen de etiquetas morfosintácticas propuestas.

4. Etiquetado léxico y ortográfico

Otro tipo de información lingüística que nos interesa codificar en el corpus XIGA es la información léxica y terminológica propia del dominio de la informática. La anotación de las unidades terminológicas se lleva a cabo en dos etapas: en la primera, se localizan y anotan los términos del corpus que forman parte de un repertorio terminológico estándar de términos informáticos (Gómez Guinovart & Lorenzo Suárez 1994, Hermida 2001); en la segunda, se aplican técnicas clásicas de extracción automática de terminología basadas en patrones morfosintácticos e índices de frecuencia (Justeson & Katz 1995).

También se organiza en dos fases la identificación en el corpus de nombres propios, entre los que destacan los correspondientes a empresas y productos de marca registrada. En una primera etapa, se identifican y anotan convenientemente los nombres recogidos en una lista elaborada manualmente; mientras que, en una segunda fase, se emplean técnicas de reconocimiento de nombres características del ámbito de la extracción de información, basadas en expresiones regulares y patrones ortográficos (Grishman 1997).

La información ortográfica no puede quedar al margen del interés de nuestro trabajo, ya que el corpus XIGA incluye textos escritos en gallego en normativas ortográficas distintas a la oficial. A esta variabilidad ortográfica internormativa, producto de una realidad sociolingüística gallega no normalizada y que tiene aún pendiente la consolidación definitiva de su estándar, hay que añadirle la variabilidad resultante de los errores ortográficos producidos por los *lapsus calami* o por el desconocimiento de la norma (un desconocimiento agravado, en muchos casos, por la confusión generada por la coexistencia de diversas normativas).

Tanto para la variabilidad internormativa, como para los errores ortográficos, hemos optado por mantener en el corpus las formas originales, incorporando en el etiquetado las acalaraciones ortográficas pertinentes para poner en relación la forma gráfica atestiguada con su forma gráfica correspondiente en la normativa oficial en vigor (RAG/ILG 1982, 1995; Real Academia Galega 1997).

Para la variabilidad internormativa utilizamos la etiqueta <ORIG>, indicando en su atributo "reg" la forma ortográfica normativa de referencia. En cuanto a las formas léxicas con errores ortográficos (es decir, en cuanto a las formas gráficas que no pertenecen a la normativa utilizada en el texto y que no coinciden con la normativa oficial), empleamos la etiqueta <SIC> para señalarlas e indicamos en su atributo "corr" la forma normativa oficial correspondiente. Así mismo, decidimos incorporar la etiqueta ortográfica <VAR> para anotar las variantes léxicas menos usuales admitidas por la normativa oficial en diversos niveles lingüísticos (ortográfico, léxico y morfológico), con vistas a facilitar el procesamiento de los datos, sin falsear las formas atestiguadas. Por ejemplo, dentro de la

normativa oficial vigente, tan correcto es en gallego escribir *variable* como *variábel*, si bien se recomienda (y es más frecuente) el uso de la primera forma. En estos casos, utilizamos la etiqueta <VAR> para señalar la variante y el atributo "usu" para incorporar la forma normativa más usual.

Ilustraremos estas diferentes anotaciones léxicas y ortográficas con un ejemplo, donde las etiquetas <TERM> se reservan para la terminología informática y las etiquetas <NAME> incluyen un atributo "type" que especifica su clase semántica.

```
<P> En <DATE value="1991"> 1991 </DATE>
<NAME type="person"> Linus Torvalds </NAME>
decide <ORIG reg="fabricarse"> fabricar-se
</ORIG> un núcleo similar <VAR usu="ó"> ao
</VAR> de <NAME type="trademark"> Unix
</NAME>, <VAR usu="ó"> ao </VAR> que vai
chamar <NAME type="trademark"> Linux
</NAME>, co obxecto de <ORIG reg="podelo">
pode-lo </ORIG> <TERM> executar </TERM> no
seu <TERM> computador </TERM>, un <TERM> PC
<VAR reg="compatible">compatíbel </VAR>
</TERM>. O proxecto <NAME type="group"> GNU
</NAME> pola <ORIG reg="súa"> sua </ORIG>
banda xa estaba a traballar nun núcleo do
seu sistema, o <NAME type="trademark"> Hurd
</NAME>, mais <NAME type="person"> Linus
</NAME> vai desenvolver <ORIG reg="moito">
muito </ORIG> máis rápido o seu núcleo,
aproveitando un <ORIG reg="medio"> meio
</ORIG> en expansión: a <NAME
type="element"> Internet </NAME>, por onde
lle <ORIG reg="chegaría"> chegaria </ORIG>
axuda tanto de programadores como de <ORIG
reg="probadores"> provadores </ORIG> do
<ORIG reg="producto"> produto </ORIG>. Por
outra banda o <NAME type="trademark"> Linux
</NAME> encaixa <ORIG reg="perfectamente">
perfeitamente </ORIG> co <TERM> software
</TERM> da <NAME type="group"> GNU </NAME>,
de feito foi desenvolvido usando o <TERM>
compilador </TERM> <NAME type="trademark">
gcc </NAME>. Xuntando o <NAME
type="trademark"> Linux </NAME> coas
ferramentas <NAME type"group"> GNU </NAME> e
<SIC corr="algún"> algun </SIC> outro <TERM>
software </TERM>, máis ou menos libre, como
o <TERM> sistema de xanelas </TERM> <NAME
type="trademark"> XFree86 </NAME>, <ORIG
reg="xorde"> xurde </ORIG> un <TERM> sistema
operativo </TERM> libre, o <NAME
type="trademark"> GNU/Linux </NAME>.
```

En este fragmento del artículo "Computadores galegófonos" de Ramón Flores[4], pueden observarse diversas diferencias ortográficas con respecto a la normativa oficial, entre otras razones, por divergencias en el uso del guión con los enclíticos ("fabricar-se" por "fabricarse", "pode-lo" por "podelo", etc.), por discrepancias sobre las normas de acentuación ("sua" por "súa", "chegaria" por "chegaría", etc.) o por la selección particular de formas léxicas ("meio" por "medio", "xurde" por "xorde", "muito" por "moito"). Junto a estas variantes inter-normativas, se pueden apreciar también errores ortográficos intra-normativos ("algun" por "algún"), y formas aceptadas por la normativa oficial como variantes menos recomendables ("ao" por "ó", "compatíbel" por "compatible"). En todos estos casos, indicamos en el atributo correspondiente la forma gráfica estándar de referencia.

[4] Publicado en Internet como http://mx4.xoom.com/ramonflores.1/PARADISO/Computadores-galegofonos.html.

5. Conclusiones

En este artículo hemos planteado algunos de los problemas generales que plantea el etiquetado lingüístico (ortográfico, léxico y morfosintáctico) de corpus en lengua gallega, como son el diseño de un etiquetario gramatical adaptado a las características lingüísticas específicas de esta lengua y la constatación en los textos de formas gráficas en normativas distintas de la oficial. En ambos casos, hemos tratado de explicar y argumentar las soluciones que hemos adoptado en el procesamiento del corpus XIGA de textos del ámbito de la informática en gallego. Una vez consolidadas, estas soluciones serán la base de próximos desarrollos en el procesamiento del CLUVI. Deseamos que la utilización sistemática de los estándares propuestos por TEI y EAGLES permita que otros grupos y proyectos compartan nuestro trabajo.

Bibliografía

Álvarez, R.; Regueira, X. L. & Monteagudo, H. (1986). *Gramática galega*. Vigo: Galaxia.

Álvarez Lugrís, A. (2001). *Estilística comparada da traducción: Proposta metodolóxica e aplicación práctica ó estudio do corpus TECTRA de traduccións do inglés ó galego*. Vigo: Servicio de Publicacións da Universidade de Vigo.

EAGLES (1996). *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages*. Pisa: ILC-CNR.

Freixeiro Mato, X. R. (1998-2000). *Gramática da lingua galega*. Vigo: A Nosa Terra.

Gómez Guinovart, X. & Lorenzo Suárez, A. M. (1994). *Vocabulario de informática (galego, inglés, castelán)*. Vigo: Servicio de Normalización Lingüística da Universidade de Vigo.

Gómez Guinovart, X. & Lorenzo Suárez, A. (2001). "Neología informática del gallego y equivalencias terminológicas en el corpus XIGA/SLI". *Actas del XIX Congreso de la Asociación Española de Lingüística Aplicada*.

Gómez Guinovart, X.; Lorenzo Suárez, A. M. & Araya, R. (2001). "Aspectos da elaboración dun léxico de informática en lingua galega". *Cadernos de lingua*.

Grishman, R. (1997). "Information extraction: techniques and challenges". En Pazienza, M. T. (ed.), *Information extraction: a multidisciplinary approach to an emerging information technology*, 10-27. Berlín: Springer.

Hermida, Ana (2001). *Glosario de termos da Internet (galego, inglés, castelán, portugués)*. Vigo: Seminario de Lingüística Informática.
[URL http://www.uvigo.es/webs/sli/glinternet/].

Justeson, J. & Katz, S. (1995). "Technical terminology: some linguistic properties and and algorithm for identification in text". *Natural Language Engineering* 1(1): 9-27.

Morel, J.; Torner, S.; Vivaldi, J. & Cabré, M. T. (1997). "El Corpus de l'IULA: etiquetaris". IULA/INF018/97. Barcelona: Universitat Pompeu Fabra.

RAG/ILG (1982). *Normas ortográficas e morfolóxicas do idioma galego*. Vigo, Real Academia Galega/Instituto da Lingua Galega (12ª edición revisada: 1995).

Real Academia Galega (1997). *Diccionario da Real Academia Galega*. A Coruña: Real Academia Galega.

Rodríguez Yáñez, X. P.; Lorenzo Suárez, A. M. & Ramallo, F. (2001). "Corpus informatizado de habla bilingüe gallego-castellano de la Universidad de Vigo". *Actas del XIX Congreso de la Asociación Española de Lingüística Aplicada*.

Sampson, G. (1995). *English for the computer*. Oxford: Clarendon Press.

Sperberg-McQueen, C. M. & Burnard, L. (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange: TEI P3*. Chicago-Oxford: ACL / ACH / ALLC.

Vilares, M.; Graña, J.; Araujo, T.; Cabrero, D. & Diz, I. (1998). "A tagger environment for Galician". *Workshop on Language Resources for European Minority Languages*. Granada.