Tecnoloxías da lingua galega e normalización lingüística

Xavier Gómez Guinovart
Seminario de Lingüística Informática
Universidade de Vigo. http://sli.uvigo.es

1. AS TECNOLOXÍAS DA LINGUA

As tecnoloxías da lingua (TL), xunto coa enxeñaría lingüística, o procesamento da linguaxe natural e boa parte das industrias da lingua, forman parte da orientación máis aplicada da lingüística computacional. O seu obxectivo central é o desenvolvemento de aplicacións informáticas de uso xeral para facilitar o uso da lingua, a súa tradución, o seu estudo e a súa aprendizaxe. Con todo, o desenvolvemento de aplicacións de TL supón unha sólida base previa de recursos e ferramentas. Por ferramentas de TL entendemos os sistemas informáticos orientados ao desenvolvemento de aplicacións de TL, mentres que os recursos de TL constitúen os datos lingüísticos sobre os que se constrúen as aplicacións. Os recursos e a investigación lingüística constitúen os alicerces das ferramentas e das aplicacións e, asemade, as ferramentas e as aplicacións son de grande utilidade cando se trata de ampliar ou de mellorar os recursos lingüísticos adquiridos. As ferramentas, os recursos e as aplicacións son os compoñentes que integran os sistemas de TL. Podemos observar a súa interrelación na seguinte figura, empregando como elemento de comparación o nivel de análise lingüística con que se vincula cada elemento (cf. Sarasola, 2000)

<table>
<tr><td rowspan="8">Aplicacións</td><td colspan="5">A8. sistemas de diálogo</td></tr>
<tr><td colspan="5">A7. sistemas de aprendizaxe de idiomas</td></tr>
<tr><td colspan="4">A6. tradución automática e asistida</td><td></td></tr>
<tr><td colspan="4">A5. recuperación da información, extracción da información, sistemas de pregunta-resposta, resumo e catalogación</td><td></td></tr>
<tr><td colspan="3">A3. corrector gramatical e do estilo</td><td></td><td>A4. conversión texto-voz</td></tr>
<tr><td>A1. corrector ortográfico</td><td colspan="3"></td><td>A2. conversión voz-texto</td></tr>
<tr><td></td><td>Léxico</td><td>Morfoloxía</td><td>Sintaxe</td><td>Semántica</td><td>Fala</td></tr>
</table>

Ferramentas				F6. desambiguación semántica no nivel léxico	
					F5. procesamento da fala no nivel de frase
			F4. analizador sintáctico		
		F3. etiquetador morfolóxico e lematizador			
		F1. análise e xeración morfolóxica			F2. procesamento da fala no nivel de palabra
	Léxico	Morfoloxía	Sintaxe	Semántica	Fala

Recursos	R9. córpora anotados semanticamente e cos sentidos das palabras desambiguados				
	R8. base de coñecementos léxico-semánticos				
	R7. córpora de textos anotados sintacticamente				
	R6. listaxe de lemas con información morfo-sintáctica e con expresións plurilexemáticas				
	R5. córpora de textos anotados categorialmente e lematizados				
	R4. listaxe de lemas con información morfosintáctica				
	R2. córpora de textos crus ou sen anotar				R3. córpora orais
	R1. listaxe de palabras flexionadas				
	Léxico	Morfoloxía	Sintaxe	Semántica	Fala

1.1. Aplicacións das tecnoloxías da lingua

O abano de aplicacións xeradas no eido das TL é realmente amplo, aínda que o podemos presentar dunha maneira didáctica agrupando os seus produtos en catro categorías: os sistemas de comprensión e xeración de textos, as aplicacións das tecnoloxías da fala, as aplicacións para o procesamento documental e as aplicacións para o procesamento plurilingüe.

1.1.1. Comprensión e xeración de textos

Os sistemas informáticos de comprensión da linguaxe permiten que o computador, a partir dun enunciado oral ou escrito, sexa quen de construír unha representación do seu significado, permitindo por tanto que o computador obedeza ordes humanas expresadas en lingua natural. Por outra banda, os programas informáticos de xeración de textos logran que os computadores poidan elaborar un texto a partir dun significado, facilitando que os computadores expresen os datos que coñecen en forma de enunciados dunha lingua natural humana. A combinación axeitada das técnicas informáticas de comprensión e de xeración de enunciados permite estabelecer diálogos en lingua natural entre un computador e unha persoa (A8). Na maioría dos casos, a través dese diálogo, o ser humano procurará resolver un problema axudándose da potencia de cálculo do computador ou tentará coñecer unha información almacenada na súa memoria; noutras circunstancias, nos ámbitos da robótica e da domótica, a persoa indicará mediante enunciados lingüísticos as accións físicas que debe realizar a máquina ou o robot controlado polo computador.

1.1.2. Tecnoloxías da fala

As tecnoloxías da fala ocúpanse do procesamento dos sons da linguaxe co obxectivo de permitir a comunicación oral entre as persoas e os computadores. Cando a dirección da comunicación é da persoa ao computador empréganse aplicacións de recoñecemento da fala, que consiste no procesamento da percepción acústica ou conversión de voz a texto (A2). As técnicas de recoñecemento da fala son cada vez máis populares grazas aos programas de ditado. Os sistemas de ditado máis recentes para computadores persoais permiten ditar texto ao computador utilizando fala continua cunha precisión bastante elevada. Os mellores produtos de ditado, en condicións estándar, poden obter máis do 90 por cento de fiabilidade despois dun mínimo adestramento. Outras aplicacións importantes do recoñecemento da fala son o control vocal do computador, de grande relevancia para persoas con certas discapacidades motoras, e o acceso vocal a servizos automatizados por vía telefónica. Por outro lado, cando a dirección da comunicación é do computador á persoa, utilízanse as técnicas de síntese da fala, que consisten no procesamento da produción fonética ou conversión de texto a voz (A4). As aplicacións máis coñecidas da síntese da fala son os programas de lectura, que permiten escoitar con voz sintetizada calquera texto que apareza na pantalla do computador. Estas aplicacións son imprescindíbeis para profesionais que teñen que saber o que pon nunha pantalla, pero non poden desviar a súa mirada do traballo que están a realizar (como cando se está a pilotar un avión ou a realizar

unha cirurxía médica) e, por suposto, para as persoas cegas ou con outro tipo de discapacidades visuais que manexan un computador.

1.1.3. Procesamento documental

As aplicacións de procesamento documental están concibidas para o seu uso durante a elaboración, a xestión e a revisión de documentos textuais. Estas constitúen algunhas das actividades máis comúns na informática persoal e, por tanto, un dos ámbitos de maior relevancia para as TL. Dentro desta gama de aplicacións, pódense citar os dicionarios electrónicos e os sistemas de verificación automática da ortografía (A1), a sintaxe e o estilo (A3) integrados nos procesadores de texto, os programas de xeración automática de resumos, os sistemas de extracción de información, os sistemas de recuperación da información textual, os sistemas de pregunta-resposta e os programas de catalogación documental automatizada (A5). A xeración automática de resumos permite presentar a información dos documentos de maneira sinóptica, o que posibilita realizar unha avaliación visual rápida da súa pertinencia en relación a unha necesidade específica de información. No canto diso, as aplicacións de extracción da información serven para localizar un conxunto de datos concretos nun texto, por exemplo, para incorporar eses datos a unha base de datos. A diferenza das aplicación de comprensión da linguaxe natural e de xeración de resumos, a extracción da información non pretende representar toda a información dun texto, senón unicamente a información seleccionada, coa finalidade de ofrecer unha vía eficiente de consulta aos grandes volumes de datos escritos en linguaxe natural nas noticias dos xornais, nos informes médicos hospitalarios ou nas sentenzas xudiciais. Pola súa banda, a recuperación da información textual traballa con conxuntos amplos de documentos e, tipicamente, co conxunto de documentos accesíbeis na web. O seu obxectivo é seleccionar, a partir dese conxunto, os documentos máis relevantes en relación cunhas necesidades concretas de información expresadas nunha consulta en forma de termos chave. O resultado da consulta é a lista seleccionada dos documentos que o sistema de recuperación da información (tipicamente, o buscador de Internet) considera relevantes para satisfacer a consulta. Outro tipo distinto de aplicacións para a procura de información textual en grandes volumes de documentos son os sistemas de pregunta-resposta. Estes sistemas están deseñados para responder a preguntas formuladas en linguaxe natural, baseándose na información contida en grandes coleccións de textos e, tipicamente, de novo, no conxunto de documentos accesíbeis na web. Dada unha consulta en forma de pregunta en lingua natural, un sistema de pregunta-resposta, en primeiro lugar, vai tratar de comprender a pregunta e, despois, non se vai limitar a confeccionar unha listaxe cos enderezos dos documentos máis relevantes, como había facer un sistema de recuperación da información, senón que vai mirar de atopar entre os documentos consultados a resposta concreta a esa pregunta. Por último, as aplicacións de catalogación documental determinan automaticamente o contido xeral dos textos analizados e clasifícanos dentro dunha tipoloxía semántica preestabelecida. Por exemplo, as axencias de novas usan filtros de catalogación documental automática para recoller mensaxes de todo o mundo, analizalas,

e clasificalas por categorías, de xeito que as mensaxes relevantes poidan ser reenviadas inmediatamente aos clientes apropiados.

1.1.4. Procesamento plurilingüe

As principais aplicacións para o procesamento plurilingüe están orientadas cara ao ensino de linguas ou cara á tradución. No primeiro grupo, situariamos as aplicacións didácticas da informática, como os métodos de aprendizaxe de idiomas (A7) asistida por computador e os programas de creación de exercicios para clase. No segundo, as aplicacións para a tradución (A6), como os programas de tradución automática e os sistemas integrados de tradución asistida. Estes sistemas integrados combinan nun único produto informático un procesador de textos especialmente deseñado para traducir, un conxunto de dicionarios bilingües, ferramentas para a xestión terminolóxica (para a creación e o mantemento de glosarios, a consulta automática de glosarios durante a tradución, e a extracción automática de terminoloxía), e unha utilidade de tradución asistida denominada *memoria de tradución*. A memoria de tradución consiste nunha base de datos onde se almacena a versión orixinal e traducida de cada unha das frases que se traducen no marco da aplicación. Cando se está traducindo unha frase, o programa detecta automaticamente se esa mesma frase ou outra similar xa fora traducida con anterioridade, de xeito que se poida reutilizar a tradución vella coas modificacións que se consideren oportunas, e non cumpra reescribila completamente de novo. Un último tipo particular de aplicacións para o procesamento plurilingüe é o constituído polos programas de identificación automática da lingua, os cales tentan determinar de maneira automática, no menor tempo posíbel e con distintas finalidades, o idioma que está sendo utilizado pola persoa que interacciona co computador.

1.2. Recursos lingüísticos e ferramentas de desenvolvemento

Porén, como xa indicamos na introdución deste apartado, o desenvolvemento de aplicacións de TL para unha lingua depende dunha maneira crucial dos recursos e ferramentas de TL ao seu dispor e, de certo, os recursos máis necesarios na construción de aplicacións de TL son os dicionarios e os córpora lingüísticos. Os dicionarios poden consistir en simples listas de palabras flexionadas (R1), nunha listaxe de lemas con información morfosintáctica (R4), na listaxe anterior ampliada con expresións plurilexemáticas (R6), en bases de coñecementos léxico-semánticos (R8), ou en calquera dos anteriores formatos complementados con información léxica plurilingüe. Os córpora tamén poden amosar unha grande variedade e complexidade, desde os córpora de textos crus ou sen anotar (R2), aos córpora de textos anotados categorialmente e lematizados (R5), os córpora de textos anotados sintacticamente (R7) e os córpora anotados semanticamente e cos sentidos das palabras desambiguados (R9), alén da posíbel versión plurilingüe de cada un destes formatos, con textos comparábeis ou paralelos, e dos córpora orais (R3).

As diversas características dos córpora están en función do seu contido textual e das aplicacións para as que foron deseñados. Por exemplo, moitas aplicacións das tecnoloxías da fala (como as axendas por voz, os sistemas de ditado ou as utilidades de

control oral do teléfono móbil) basean os seus algoritmos de identificación de palabras en córpora orais de aprendizaxe constituídos por enunciados gravados en condicións acústicas controladas. Moi distintos destes córpora orais son os córpora compostos por textos escritos orientados, por exemplo, á elaboración de algoritmos para a clasificación de documentos en aplicacións como o filtrado do correo electrónico lixo ou o direccionamento selectivo de textos. Dentro desta categoría de córpora escritos é onde hai que situar os córpora escritos plurilingües, constituídos por textos escritos en dous ou máis idiomas, e creados para satisfacer diversas necesidades relacionadas co procesamento informático do multingüismo. Os córpora paralelos representan unha especialización dos córpora plurilingües, na que os textos recompilados son traducións os uns dos outros. En calquera caso, os córpora crus, isto é, os córpora que conteñen exclusivamente as palabras e os signos de puntuación dos textos orixinais, teñen unha utilidade bastante limitada para as TL. Para que os córpora comecen a ser verdadeiramente útiles no desenvolvemento de aplicacións, débese enriquecer o texto dos documentos orixinais con diversos tipos de información lingüística, se é posíbel, de xeito automático ou semiautomático, mediante unha ferramenta específica de desenvolvemento de TL. Estes córpora enriquecidos con información lingüística reciben a denominación de córpora etiquetados, xa que a información engadida ao texto orixinal incorpórase contida dentro de etiquetas, en xeral, empregando algunha especificación da linguaxe XML para a anotación de textos.

As ferramentas de desenvolvemento de TL, pola súa banda, son programas informáticos que realizan tarefas lingüísticas moi concretas e especializadas sobre as que se constrúen as aplicacións máis xerais de TL de uso común. De acordo coa tarefa específica á que estean destinadas, podemos agrupar as ferramentas básicas de TL en programas de análise e xeración morfolóxica (F1), programas para o procesamento da fala no nivel de palabra (F2), programas de desambiguación morfolóxica e lematización (etiquetadores) (F3), programas de análise sintáctica (F4), ferramentas para o procesamento da fala no nivel de frase (F5), e programas de desambiguación semántica no nivel léxico (F6). Na actualidade, a anotación morfosintáctica dos córpora etiquetados (isto é, a incorporación dunha indicación da súa categoría gramatical a cada palabra do corpus) pode efectuarse cun alto grao de automatización, polo que resulta o tipo de anotación que se atopa con máis frecuencia nos córpora etiquetados. Os programas etiquetadores analizan categorialmente as palabras do corpus e devólvenas acompañadas da súa etiqueta coa categoría morfosintáctica máis probábel no seu contexto, cunha taxa de acerto moi elevada. Algúns destes etiquetadores logran tamén estabelecer o lema (ou entrada de dicionario) de cada palabra etiquetada con moito acerto. Os resultados dos programas etiquetadores constitúen o punto de partida para os programas de análise sintáctica, que se encargan de construír representacións sintácticas das frases analizadas seguindo, en xeral, as especificacións dalgunha gramática computacional elaborada para tal efecto. Os programas de desambiguación semántica no nivel léxico, finalmente, analizan cada palabra polisémica do texto para indicar cal dos seus posíbeis significados é o máis probábel no seu contexto de aparición.

2. AS TECNOLOXÍAS DA LINGUA GALEGA

No caso da lingua galega, os principais axentes implicados na produción de TL son os centros de investigación pública, fundamentalmente universitarios, e un contado conxunto de empresas de capital privado. Entre os primeiros, os grupos máis activos e con máis experiencia nesta disciplina son o Grupo de Tratamento do Sinal da Universidade de Vigo (http://www.gts.tsc.uvigo.es), o Instituto da Lingua Galega da Universidade de Santiago de Compostela (http://www.usc.es/~ilgas/), e o Seminario de Lingüística Informática da Universidade de Vigo (http://sli.uvigo.es); e, xa fóra do ámbito universitario, o Centro Ramón Piñeiro para a Investigación en Humanidades, dependente da Xunta de Galicia (http://www.cirp.es). No sector privado, as empresas que desenvolven actividades máis relacionadas coas TL son Imaxin Software (http://www.imaxin.com) e Dimensiona (http://www.dimensiona.com), as dúas situadas en Santiago de Compostela.

2.1. Panorama das tecnoloxías lingüísticas do galego

Recollemos a continuación algunhas das aplicacións e dos recursos de TL máis relevantes para a lingua galega dispoñíbeis a través de Internet neste momento (inicios de 2006), seguindo a clasificación exposta no apartado anterior, co obxectivo de ofrecer, aínda que sexa dun xeito moi sucinto, unha visión panorámica do estado da cuestión neste ámbito de traballo da lingüística aplicada do galego.

A1. *Corrector 2.Mil3*. Última versión do corrector ortográfico de galego de Imaxin Software para Microsoft Office, baseado no VOLG (Santamarina, 2003). Inclúe as normas aprobadas pola Real Academia Galega en 2003. A descarga do corrector é gratuíta desde a páxina web de Imaxin Software, no enderezo http://www.imaxin.com.

A2. *Transcrigal*. Sistema de conversión voz-texto para a transcrición automática de programas de noticias desenvolvido polo Grupo de Tecnoloxías do Sinal da Universidade de Vigo (García Mateo, 2003). Hai unha versión de demostración na web no enderezo http://www.gts.tsc.uvigo.es/Transcrigal.

A4. *Cotovía*. Sistema de conversión texto-voz desenvolvido polo Grupo de Tecnoloxías do Sinal da Universidade de Vigo (García Mateo, 2003). Pódese probar unha versión de demostración na web no enderezo http://www.gts.tsc.uvigo.es/cotovia/.

A7. *Tradutor español <-> galego*. Tradutor automático OpenTrad/Apertium español-galego e galego-español de código aberto desenvolvido polo Seminario de Lingüística Informática da Universidade de Vigo (Corbí Bellot e outros, 2005). O seu uso é libre e gratuíto a través da nosa web no enderezo http://sli.uvigo.es/tradutor.

A8. *é-galego*. Curso multimedia de galego de nivel intermedio dirixido a falantes de castelán, elaborado polo Instituto da Lingua Galega da Universidade de Santiago de Compostela (Rodríguez Malmierca e Gromaz Campos, 2002). De seguimento libre e gratuíto na web no enderezo http://e galego.cesga.es.

F3-F4. *Analizador lingüístico de galego.* Demostración das ferramentas FreeLing (Atserias e outros, 2006) de análise morfolóxica, desambiguación, lematización e análise sintáctica da lingua galega desenvolvidas polo Seminario de Lingüística Informática da Universidade de Vigo. O seu uso é libre e gratuíto a través da nosa web no enderezo http://sli.uvigo.es/lingua.

R2. *Tesouro Medieval Informatizado da Lingua Galega (TMILG)*. Corpus diacrónico do galego, de máis de nove millóns de palabras, desenvolvido no Instituto da Lingua Galega (Varela Barreiro, 2004). Contén a totalidade das obras non notariais publicadas da Galicia medieval (literarias, históricas, relixiosas, xurídicas e técnicas) e o 80% das obras notariais publicadas. De libre consulta, tras a alta no sistema, no enderezo http://corpus.cirp.es/tmilg.

R2. *Corpus de Referencia do Galego Actual (CORGA)*. Corpus de textos do galego contemporáneo (desde 1975 á actualidade) de 17500000 palabras, desenvolvido polo Centro Ramón Piñeiro para a Investigación en Humanidades (López Martínez, 2005) e accesíbel vía internet, tras a alta no sistema, no enderezo http://corpus.cirp.es/corga.

R3. *A Nosa Fala*. Corpus oral representativo de todas as variedades diatópicas do galego desenvolvido no Instituto da Lingua Galega (Fernández Rei e Hermida Gulías, 2003), e dispoñíbel libremente na web no enderezo http://www.consellodacultura.org/arquivos/asg/anosafala.php.

R5. *Corpus Lingüístico da Universidade de Vigo (CLUVI)*. Corpus paralelo, de traducións ao galego ou do galego, de 14 millóns de palabras, desenvolvido polo Seminario de Lingüística Informática da Universidade de Vigo (Gómez Guinovart e Sacau Fontenla, 2004) e de libre consulta no enderezo http://sli.uvigo.es/CLUVI.

R5. *Corpus Técnico do Galego (CTG)*. Corpus etiquetado e lematizado de rexistros especializados do galego contemporáneo nos ámbitos do dereito, da economía, da informática e as telecomunicacións, e da ecoloxía e medio ambiente, que están a realizar o Seminario de Lingüística Informática e o Observatorio de Neoloxía da Universidade de Vigo. De libre consulta no enderezo http://sli.uvigo.es/CTG.

R5. *Tesouro Informatizado da Lingua Galega (TILG)*. Corpus etiquetado e lematizado do galego moderno de 20 millóns de palabras desenvolvido no Instituto da

Lingua Galega. Contén case a totalidade das obras en galego entre 1612 e 1980. De libre consulta na web no enderezo http://www.ti.usc.es/TILG.

R6. *Dicionario CLUVI inglés-galego (CLIG)*. Dicionario bilingüe inglés-galego elaborado a partir do corpus paralelo CLUVI, que conta cunhas 10 000 acepcións ilustradas con exemplos documentados nos textos en inglés traducidos ao galego recompilados no CLUVI (Gómez Guinovart e Sacau Fontenla, 2005). De acceso libre e gratuíto na web no enderezo http://sli.uvigo.es/CLIG.

R6. *Dicionario da Real Academia Galega (DRAG)*. Versión electrónica do dicionario de referencia da RAG (García Pérez e González González, 1997), que conta cunhas 20 000 entradas. De libre consulta no enderezo http://www.edu.xunta.es/diccionarios.

R6. *Dicionário e-Estraviz*. Versión electrónica actualizada do dicionario de Estraviz (Estraviz, 1995), que conta cunhas 90 000 entradas redactadas en normativa reintegracionista. De libre consulta no enderezo http://www.agal-gz.org/estraviz/.

2.2. Liñas de traballo

Este panorama que acabamos de presentar permítenos observar algunhas das insuficiencias das tecnoloxías da lingua galega que cumpriría emendar a curto prazo. Con certeza, unha das eivas máis importante dáse no ámbito da semántica computacional, no que se constata a inexistencia de compoñentes funcionais en todos os niveis de desenvolvemento. No nivel básico dos recursos lingüísticos, non dispomos nin dunha base de coñecemento léxico-semánticos do galego, nin de córpora anotados desambiguados semanticamente e, paralelamente, no nivel das ferramentas, tampouco dispomos aínda de programas informáticos para a desambiguación semántica das palabras dun texto. No nivel das aplicacións, tanto a inexistencia de sistemas complexos de diálogo baseados na comprensión automática dos enunciados, coma a inexistencia de aplicacións de extracción de información en galego que aproveiten o potencial do procesamento semántico (como os sistemas de pregunta-resposta), están motivadas en grande medida por esta ausencia de recursos e ferramentas sobre os que fundamentar a análise semántica imprescindíbel para o funcionamento destas importantes tecnoloxías.

No campo dos recursos lingüísticos, cómpre seguir traballando na elaboración e na dispoñibilidade en Internet de repertorios lexicográficos plurilingües para distintas combinacións lingüísticas co galego para as que non dispomos de ningún recurso electrónico hoxe en día. De certo, unha das combinacións con maior demanda social sería a combinación galego-español-galego, aínda que tamén serían dunha grande utilidade para as didácticas das linguas e para a tradución profesional as combinacións galego-francés-galego, galego-alemán-galego e outras. Outra das eivas salientábeis no campo dos recursos lexicográficos cuxa solución consideramos prioritaria son os dicionarios electrónicos de sinónimos e de ideas afíns, tan necesarios nas aplicacións de escritura asistida por

ordenador facilitadas polos procesadores de texto, coma nos sistemas avanzados de recuperación de información que os utilizan para expandir os termos chave da consulta.

Finalmente, no ámbito das aplicacións, hai que traballar arreo en tres vertentes de grande relevancia. Por un lado, para desenvolver sistemas de corrección gramatical e do estilo que, integrados nos procesadores de textos máis populares, actúen como asistentes na redacción de documentos en lingua galega. Por outro, na elaboración de aplicacións operativas para a recuperación, a extracción, a consulta, o resumo e a catalogación da información escrita na nosa lingua. Por último, no desenvolvemento dun tradutor automático inglés-galego e galego-inglés que, con todas as limitacións da tradución automática, pero tamén con todas as súas vantaxes de inmediatez e gratuidade, nos permita acceder en galego aos contidos orixinalmente redactados en inglés e, ao mesmo tempo, nos permita presentar en lingua inglesa os textos redactados orixinalmente no noso idioma.

3. TECNOLOXÍAS DA LINGUA E NORMALIZACIÓN LINGÜÍSTICA

A dirección actual da investigación e desenvolvemento en tecnoloxías da lingua sinala dous obxectivos prioritarios: por unha banda, a superación das barreiras lingüísticas que, especialmente no ámbito de Internet, impiden un acceso igualitario aos contidos e, de maneira moito máis relevante para a industria, dificultan a implantación efectiva dun comercio electrónico realmente global; por outra banda, a superación das «barreiras arquitectónicas» de acceso ás novas tecnoloxías mediante a implementación de interfaces de interacción persoa-computador en linguaxe natural oral (Gómez Guinovart, 2000). As tecnoloxías chave que nos haberán de permitir o uso da propia lingua para acceder a todas as posibilidades de información, comunicación e consumo postas ao dispor dos habitantes do primeiro mundo polas novas tecnoloxías son a identificación automática da lingua, a tradución automática, o recoñecemento e síntese da fala, a recuperación interlingüística da información, os sistemas de diálogo e a web semántica. Porén, para poder acceder en igualdade de condicións a esta infraestrutura social e económica na que se configura a nosa sociedade da información, unha comunidade lingüística deberá contar con estas tecnoloxías e deberá estar representada dentro desa infraestrutura. Pola súa grande incidencia social, a presenza dunha lingua neste ámbito ha ser determinante para acadar ou para conservar o estado de lingua normalizada.

Estimamos que a presenza do galego nas tecnoloxías da lingua, e a través delas en Internet, resulta tamén determinante para a súa normalización lingüística porque, por unha banda, Internet é o medio de comunicación actual de maior importancia estratéxica, e a presenza do galego nos medios de comunicación é imprescindíbel para a súa normalización; por outra banda, porque o incremento da percepción social do prestixio e utilidade do galego é fundamental para a recuperación da súa transmisión lingüística familiar, e a presenza do galego en Internet garante o incremento necesario desa percepción de prestixio e utilidade; e, finalmente, porque o galego precisa dunha identidade e unha presenza in-

ternacional recoñecida e respectada, e a presenza do galego en Internet asegura a difusión axeitada para o seu coñecemento e recoñecemento no nivel global (Ramos i Armengol, 1997; Gómez Guinovart, 2003).

En último termo, pensamos que a presenza do galego nas tecnoloxías da lingua, e a través delas na informática, é importante para a súa normalización porque pode contribuír a aumentar o seu uso, o seu prestixio e a súa utilidade, ademais de constituír un factor simbólico importante na percepción propia e allea da nosa lingua.

4. CONCLUSIÓNS

Aínda que o estado actual das tecnoloxías da lingua galega pon de manifesto o inmenso traballo que se ten feito, o certo é que existen algunhas eivas importantes que cumpriría emendar a curto prazo. Consideramos que as tarefas de normalización da lingua galega deben incluír entre os seus obxectivos o de dotar a sociedade do conxunto necesario de tecnoloxías da lingua galega. Estas tecnoloxías constitúen o requisito imprescindíbel para que a comunidade lingüística galega poida acceder en pé de igualdade e na lingua de seu a todos os bens e servizos da sociedade da información. A responsabilidade principal na promoción e xestión do desenvolvemento destas tecnoloxías corresponde ás institucións públicas galegas e, moi concretamente, á Secretaría Xeral de Política Lingüística e á Consellaría de Innovación e Industria, que deben realizar un importante esforzo para situar en as tecnoloxías da lingua galega ao nivel das linguas normalizadas do noso contorno, contando co compromiso necesario das entidades públicas e privadas con implicación social e procurando a optimización dos investimentos públicos. Neste sentido, cumpriría garantir que a obtención dos produtos vaia sempre vinculada á formación de persoal investigador propio e á creación de tecido económico galego, e cumpriría, asemade, asegurar a divulgación social dos resultados, fomentando que os produtos realizados grazas aos investimentos públicos en investigación, desenvolvemento e innovación sexan realmente públicos e reutilizábeis, libres, de código aberto, e compartidos sen restricións innecesarias.

Referencias

Atserias, J., B. Casas, E. Comelles, M. Gonzàlez, L. Padró e M. Padró (2006). «FreeLing 1.3: Syntactic and semantic services in an open-source NLP library», en *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Xénova (Italia).

Corbí Bellot, A.M., M.L. Forcada, S. Ortiz Rojas, J. A. Pérez Ortiz, G. Sánchez Ramírez, F. Sánchez Martínez, I. Alegria, A. Mayor e K. Sarasola (2005). «An open-source shallow-transfer machine translation engine for the Romance languages of Spain». En *Proceedings of the European Associtation for Machine Translation, 10th Annual Conference*. Budapest (Hungría).

Estraviz, I.A. (1995). *Dicionário da língua galega*. Santiago de Compostela: Sotelo Blanco.

Fernández Rei, F. e C. Hermida Gulías (2003). *A nosa fala: bloques e áreas lingüísticas do galego*. Santiago de Compostela: Consello da Cultura Galega.

García Mateo, C. (2003). «Tecnologías del habla y lenguas minoritarias», en *Procesamiento del lenguaje natural*, 31: 381-384.

García Pérez, C. e M. González González (dir.) (1997). *Diccionario da Real Academia Galega*. A Coruña: Real Academia Galega.

Gómez Guinovart, X. (2000). «Perspectivas de la lingüística computacional» en *Novática: Revista de la Asociación de Técnicos de Informática*, 145: 85-87.

Gómez Guinovart, X. (2003). «A lingua galega en Internet», en A. Bringas e B. Martín (eds.), *Nacionalismo e globalización: lingua, cultura e identidade*: 71-88. Vigo: Servizo de Publicacións da Universidade de Vigo.

Gómez Guinovart, X. e E. Sacau Fontenla (2004). «Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo)», en *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. Lisboa (Portugal).

Gómez Guinovart, X. e E. Sacau Fontenla (2005). «Técnicas para o desenvolvemento de dicionarios de tradución a partir de córpora aplicadas na xeración do Dicionario CLUVI Inglés-Galego», en *Viceversa: Revista Galega de Traducción*, 11: 159-171.

López Martínez, M.S. (2005). «El Corpus de Referencia del Gallego Actual (Corga). Problemas de configuración y anotación», en C.D. Pusch, J. Kabatek e W. Raible (eds.), *Romanistische Korpuslingustik II: Korpora und diachrone Sprachwissenschaft*: 281-292. Tübingen: Gunter Narr Verlag.

Ramos i Armengol, R. (1997). «Internet com a eina de dinamització lingüística»: en *Llengua i ús*, 9: 13-20.

Rodríguez Malmierca, M.J. e M. Gromaz Campos (2002). «É-galego: aprendiendo gallego en la red»: en *Ágora Digital*, 3.

Santamarina, A. (2003). «O Vocabulario ortográfico da lingua galega (VOLG): criterios de elaboración», en M. Álvarez de la Granja e E. González Seoane (eds.), *A estandarización do léxico*: 53-92. Santiago de Compostela: Consello da Cultura Galega.

Sarasola K. (2000). «Strategic priorities for the development of language technology in minority languages», en *Proceedings of Workshop «Developing language resources for minority languages: re-useability and strategic priorities». Second International Conference on Language Resources and Evaluation (LREC 2000)*. Atenas (Grecia).

Varela Barreiro, X. (2004). «Un proxecto do ILG no abalo da gramática histórica da lingua galega», en R. Álvarez, F. Fernández Rei e A. Santamarina (eds.), *A lingua galega: historia e actualidade*, vol. 2: 649-684. Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega.