A investigación ao servizo da normalización da lingua galega na sociedade da información

Manuel González González
Real Academia Galega. Universidade de Santiago

1. ALGUNHAS REFLEXIÓNS AO REDOR DO CUME MUNDIAL SOBRE A SOCIEDADE DA INFORMACIÓN

Como saben todos vostedes, hai poucos días finalizou o Cume Mundial sobre a Sociedade da Información na cidade de Tunes, evento que tivera a súa primeira parte en Xenebra no mes de decembro de 2003. Hai algunhas ideas expostas repetidamente neste foro, que me gustaría recordar aquí, ben porque fan fincapé na importancia que certas tecnoloxías relacionadas coa investigación lingüística teñen na sociedade da información e no desenvolvemento universal, ben porque insisten na necesidade de desenvolvemento das tecnoloxías da información e da comunicación por parte de cada pobo particular. Velaquí as que quero salientar, extraídas de dous documentos fundamentais deste cume: un deles, que leva por título *Declaración de principios*, foi aprobado o 12 de decembro de 2003 en Xenebra durante a primeira fase do cume; o outro, coñecido como *Compromiso de Tunes*, foi discutido na segunda fase, que tivo lugar en Tunes, do 16 ao 18 de novembro de 2005.

1.1. A sociedade da información debe estar ao servizo do desenvolvemento das persoas e dos pobos

> «Declaramos o noso desexo e compromiso comúns de construírmos unha sociedade da información centrada na persoa, incluínte e orientada ao desenvolvemento, na que todos poidan crear, consultar, utilizar e compartir a información e o coñecemento, para que as persoas, as comunidades e os pobos poidan desenvolver o seu pleno potencial na promoción do seu desenvolvemento duradeiro e mellorar a súa calidade de vida, consonte os obxectivos e principios da Carta das Nacións Unidas e respectando e defendendo plenamente a Declaración Universal de Dereitos Humanos» (*Declaración de principios*, punto 1)

1.2. As TIC inciden en todos os aspectos da vida e son fundamentais para o desenvolvemento

> «As tecnoloxías da información e a comunicación (TIC) teñen inmensas repercusións en practicamente todos os aspectos das nosas vidas. O rápido progreso destas tecnoloxías brinda oportunidades sen precedentes para acadar niveis máis elevados de desenvolvemento» (*Declaración de principios*, punto 8)

1.3. A fenda dixital: perigo de que moitos pobos queden á marxe nas tecnoloxías da información

> «Somos plenamente conscientes de que as vantaxes da revolución da tecnoloxía da información están na actualidade desigualmente distribuídas entre os países desenvolvidos e en desenvolvemento, así como nas sociedades. Estamos plenamente comprometidos en facer desta fenda dixital unha oportunidade dixital para todos, especialmente para aqueles que corren perigo de quedaren atrás e mesmo marxinalizados» (*Declaración de principios*, punto 10).

A este respecto convén que recordemos o atraso que presenta Galicia con respecto á maioría das sociedades do mundo occidental en canto á utilización de Internet, probablemente por razóns de índice cultural, pero tamén sen dúbida pola existencia doutras realidades tanxibles que pexan o desenvolvemento normal de Internet en Galicia, como son os custos elevados das liñas, ou a dispersión da poboación que dificulta, por exemplo, a chegada do cabo ao conxunto da poboación, e que dificulta tamén a xeneralización do acceso á banda ancha.

1.4. As TIC poden ser unha grande axuda para colectivos con discapacidades e grupos desfavorecidos

> «Debe promoverse o emprego das TIC a todos os niveis na educación, a formación e o perfeccionamento dos recursos humanos, tendo en conta as necesidades particulares das persoas con discapacidades e os grupos desfavorecidos e vulnerables» (*Declaración de principios*, punto 30).
>
> «Esforzarémonos sen pausa en promover o acceso universal, ubicuo, equitativo e alcanzable ás TIC, e especialmente o deseño universal e as tecnoloxías auxiliares para todos, con atención especial aos discapacitados, en todas as partes, con obxecto de garantir unha distribución máis uniforme dos seus beneficios entre as sociedades e dentro de cada unha delas» (*Compromiso de Tunes*, punto 18).

1.5. Necesidade de desenvolvemento das TIC por parte de cada pobo

> «Para acadar un desenvolvemento duradeiro da sociedade da información, deben reforzarse as capacidades nacionais en materia de investigación e desenvolvemento de TIC» (*Declaración de principios*, punto 33).

O desenvolvemento das tecnoloxías da información e da comunicación non pode quedar só en mans de certos estados nin deben ser ideadas para utilización exclusiva de determinadas linguas de maior peso socioeconómico no planeta. En primeiro lugar, porque a exclusividade no desenvolvemento destas tecnoloxías por parte de determinadas sociedades o que fai é acentuar a fenda entre sociedades ricas e pobres e perpetuar a situación de dependencia tecnolóxica e, o que é máis grave, tamén ideolóxica dos países menos capaces con respecto aos máis capaces. En segundo lugar, aínda que unha parte destas tecnoloxías son teoricamente utilizables por parte de calquera sociedade e aplicables a calquera lingua, existe outra parte moi importante que non é aplicable sen desenvolvementos específicos para cada unha das linguas, que esixen un gasto de enerxía e mesmo un investimento económico que para os estados dominantes non lles resultan rendibles. Por iso é fundamental o reforzamento da capacidade de cada nación en materia de investigación e desenvolvemento de TIC.

E neste aspecto é inevitable facer unha pequena reflexión sobre a situación en Galicia neste ámbito, onde o investimento en I+D é deficiente no sector público e moito máis aínda no sector privado. Mentres que en España o 52,4% do investimento en I+D procede da iniciativa privada, en Galicia o 60% procede das institucións públicas. Visto así podería parecer que en Galicia existe unha forte aposta do sector público pola investigación, pero nada máis lonxe da realidade: o compromiso do sector público foi nos últimos tempos moi deficitario; pero a aposta seria da iniciativa privada en Galicia, salvo algunhas excepcións moi contadas, foi practicamente inexistente.

1.6. Necesidade de implicación dos poderes públicos para que o desenvolvemento das TIC non quede só ao azar do mercado

> «Os poderes públicos deben intervir, segundo proceda, para corrixir os fallos do mercado, manter unha competencia leal, atraer investimentos, fomentar o desenvolvemento de infraestruturas e aplicacións TIC, para aumentar ao máximo os beneficios económicos e sociais e atender as prioridades nacionais» (*Declaración de principios*, punto 39).

Xa comentamos que Galicia ten certas características estruturais, como a dispersión da poboación, que dificultan unha xeneralización dalgún tipo de tecnoloxías da comunicación, pero este déficit que pode retraer nalgún caso o investimento da empresa privada debe estar compensado polos poderes públicos. O que non se pode manter por máis tempo, se queremos unha Galicia de futuro, é a situación actual de acusado atraso

non só con respecto a Europa, senón tamén con respecto á media do Estado español. Preocúpanos que só o 38% dos fogares galegos teñan un ordenador, fronte ao 48,1% dos de España (10 puntos menos con relación á media estatal); preocúpanos que só o 19% dos fogares galegos teñan acceso a Internet (fronte ao 30,9% do conxunto do Estado); preocúpanos que nunha sociedade onde Internet se converteu no medio máis dinámico de información e comunicación, só o 29,4% dos galegos sexa usuario deste medio (fronte ao 37,5% no conxunto do Estado español). Do escaso dinamismo das novas posibilidades de transaccións comerciais que ofrece Internet dá idea que só o 3,5% dos galegos utilizase algunha vez o comercio electrónico (fronte ao tamén irrisorio 5% no conxunto de España). E as porcentaxes son igualmente preocupantes se examinamos a situación no mundo da empresa, onde a empresa galega no seu conxunto (claro que existen notables excepcións) presentan un retraso considerable en utilización das TIC con respecto á media estatal, e enorme se establecemos como referencia os países máis desenvolvidos de Europa.

1.7. As TIC deben estar adaptadas a cada idioma e respectar cada cultura

> «As aplicacións deben ser fáciles de utilizar, accesibles para todos, alcanzables, adaptadas ás necesidades nacionais en materia de idioma e cultura, e favorables ao desenvolvemento duradeiro» (*Declaración de principios*, punto 51)

> «A diversidade cultural é o patrimonio común da humanidade. A sociedade da información debe fundarse no respecto da identidade cultural, a diversidade cultural e lingüística, as tradicións e as relixións e estimular ese respecto, ademais de promover un diálogo entre as culturas e as civilizacións. O fomento, a afirmación e preservación dos diversos idiomas e identidades culturais, tal como se consagran nos correspondentes documentos acordados polas Nacións Unidas, incluída a Declaración Universal da UNESCO sobre a Diversidade Cultural, contribuirán a enriquecer aínda máis a sociedade da información» (*Declaración de principios*, punto 52)

Cómpre facer un esforzo de localización de produtos informáticos ao galego, para garantir o dereito dos nosos cidadáns a poder utilizar a súa lingua en todo tipo de actividades ordinarias, pero tamén polo valor simbólico negativo que ten ante os usuarios da lingua e o conxunto da sociedade o percibir que a lingua galega non está presente nos produtos que se identifican coa utilidade, modernidade e o progreso. Evidentemente, existe un problema real económico, derivado da limitación do mercado. O mercado en lingua galega é un mercado relativamente pequeno, que ás veces non convida ao empresario a afrontar un gasto que lle vai reportar uns beneficios escasos. Pero este é un problema común a moitas outras linguas entre as que se atopan tamén linguas non minorizadas, que esixen nas súas lexislacións a localización dos produtos para poderen ser comercializados dentro do país. Probablemente unha lei exclusivamente galega nesta dirección, que obrigue as empresas á comercialización dunha versión en galego no territorio da Comunidade

Autónoma, non tería os efectos apetecidos, porque a unha empresa grande, como Microsoft, un mercado coma o galego pode resultarlle ata certo punto marxinal. Neste aspecto sería moito máis efectiva unha lexislación estatal, que esixise a localización de todo produto informático, e outros similares, a todas as linguas oficiais do Estado español, como condición para poder ser comercializado e distribuído en calquera punto do Estado. E o mercado de todo o Estado xa ten suficiente peso como para obrigar as empresas a que fagan un esforzo neste sentido.

1.8. A sociedade da información debe basearse na solidariedade mundial e na comprensión entre os pobos e nacións

> «Temos a firme convicción de que estamos entrando colectivamente nunha nova era que ofrece inmensas posibilidades, é dicir, a era da sociedade da información e a expansión da comunicación humana. Nesta sociedade incipiente é posible xerar, intercambiar, compartir e comunicar informacións e coñecementos entre todas as redes do mundo. Se tomamos as medidas necesarias, axiña todos os particulares poderán colaborar para construír unha nova sociedade da información baseada no intercambio de coñecemento e asentada na solidariedade mundial e unha mellor comprensión entre os pobos e as nacións. Confiamos en que estas medidas abran unha vía cara ao futuro desenvolvemento dunha verdadeira sociedade do coñecemento» (*Declaración de principios*, punto 67).

1.9. As TIC deben permitir unha maior dispoñibilidade de información e unha maior accesibilidade a esta

> «Instamos aos gobernos a que, utilizando o potencial das TIC, creen sistemas públicos de información sobre leis, regulamentos, contribuíndo a unha maior xeneralización do acceso público e a unha maior dispoñibilidade desta información» (*Compromiso de Tunes*, punto 17).

1.10. Necesidade de deseñar estratexias para a conservación da información dixital

> «Recoñecemos que o acceso equitativo e sostible á información esixe a implementación de estratexias para a conservación a longo prazo da información dixital que se está creando» (*Compromiso de Tunes,* punto 27).

2. DOUS PUNTOS NO PXNL IMPORTANTES PARA O GALEGO NA SOCIEDADE DA INFORMACIÓN

Sería imposible achegármonos a todos os aspectos da investigación sobre a lingua galega relacionada coa sociedade da información. Por iso voume limitar a analizar dous

puntos recollidos nos sectores transversais do *Plan xeral de normalización da lingua galega* que, como saben, foi aprobado hai un ano por unanimidade no Parlamento de Galicia:

a) As novas tecnoloxías, e, particularmente, as chamadas tecnoloxías da fala.
b) A implementación do corpus.

2.1. As novas tecnoloxías

2.1.1. Por que é necesario o desenvolvemento das novas tecnoloxías?

O das novas tecnoloxías é un sector prioritario pola súa eficacia e polo seu simbolismo. O investimento en novas tecnoloxías é, probablemente, o de maior rendemento e debe ser prioritario. É imprescindible se queremos facilitar que se incorporen e se manteñan instalados en galego sectores como o comercial, o informativo, o xuvenil e o mundo urbano.

Da presenza do galego nas novas tecnoloxías vai depender en boa parte a imaxe que os galegos e os non galegos teñan desta lingua; e da imaxe da lingua depende o uso: se a imaxe é boa, o uso medrará de seu (e mesmo sen subvencións); se a imaxe é cativa, o uso minguará por moito que a prol desta lingua se fagan esforzos políticos e económicos.

Ademais hoxe xa non se pode considerar de ningún xeito que a demanda de recursos, de produtos e de servizos relacionados coas chamadas novas tecnoloxías se restrinxa a unha determinada elite ou sector dominante da sociedade. Hoxe son xa moitos os sectores que reclaman recursos técnicos e informáticos para usar o galego nunha sociedade decididamente urbana e moderna. Pero esta demanda incrementarase de xeito espectacular nos vindeiros anos, e hai que estar preparados para lle dar resposta eficaz.

2.1.2. Obxectivos relacionados coas novas tecnoloxías

No *Plan xeral de normalización da lingua galega* recóllense explicitamente os seguintes obxectivos neste ámbito:

– Fomentar a presenza do galego 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 en tradución automática, recoñecemento e 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.

2.2. A implementación do corpus

Sobradamente é coñecido o fenómeno histórico do afastamento da lingua galega dos ámbitos máis formais da comunicación, cunha situación persistente ao longo de séculos

dunha diglosia funcional ao lado dunha diglosia de adscrición. Esta situación persistente durante tanto tempo xerou unha inseguridade de moitos sectores á hora de utilizaren a lingua galega na súa actividade.

Debemos aspirar a que a utilización do galego non supoña un esforzo complementario para quen se decide a abrir a súa actividade persoal ou profesional á lingua galega. É necesario, pois, superar as secuelas deste déficit histórico. Como facelo? Parece imprescindible adoptar, entre outras, medidas como:

– unha formación axeitada do persoal, que o faga sentir seguro e competente en lingua galega.
– unha oferta suficiente de recursos léxicos e terminolóxicos que lles faciliten o labor.

O déficit de recursos terminolóxicos é unha realidade palpable e preocupante. Nas enquisas sobre o uso de galego en sectores especializados, aparece case sempre visible o problema da falta de terminoloxía en galego estable e fiable; pero tamén é verdade que raramente este déficit é considerado como a causa principal da non utilización da lingua galega nestes ámbitos. Parece evidente que na sociedade actual, na que as transformacións, os avances e os descubrimentos se producen a unha velocidade de vertixe, o traballo terminolóxico debe pasar a ocupar un lugar importante en todas as linguas de maneira estable. Pero no caso da lingua galega, temos que facer un esforzo suplementario nos próximos anos, non só para atender as necesidades terminolóxicas das novas creacións, senón para cubrir o déficit histórico na nosa lingua en amplos campos da ciencia e da técnica que temos aínda a ermo. É urxente ofrecer á sociedade recursos terminolóxicos fiables, adaptados á nosa maneira de ver o mundo e á nosa cultura.

É necesaria a elaboración sistemática dos recursos terminolóxicos de cada sector e de cada área, que cubra dunha maneira folgada as necesidades coas que os usuarios se van atopar.

A terminoloxía técnico-científica e das linguas de especialidade en xeral debe emanar dunha institución fiable, que unifique as solucións, e que xere confianza entre os usuarios.

Pero esta terminoloxía hai que difundila e implantala. A terminoloxía carece de sentido se non chega con fluidez aos seus destinatarios. Cómpren canles de difusión e medios de implantación da terminoloxía que sexan eficaces:

– vocabularios sistemáticos plurilingües das linguas de especialidade;
– folletos específicos con vocabularios máis reducidos, destinados a colectivos concretos e que cubran as necesidades destes;
– incorporación deste léxico ás ferramentas e aos programas informáticos que se utilizan habitualmente nos distintos sectores;

– elaboración de modelos de documentación de circulación frecuente en lingua galega;
– un servizo de consultas, que resolva nun prazo o máis breve posible as dificultades que o profesional atope no exercicio do seu labor en galego etc.

Por estas razóns o *Plan xeral de normalización da lingua galega* recolle como un dos seus obxectivos (o C.1):

> *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*.

3. NOVAS TECNOLOXÍAS

Destes dous puntos recollidos no PXNL, aos que me referín, permítanme que me deteña un pouco máis en analizar a situación das novas tecnoloxías, dentro das que se encadran a tradución automática, a síntese de voz, o recoñecemento de voz, os clasificadores e resumidores automáticos, a identificación de locutores, a diarización etc.

3.1. A tradución automática

3.1.1. A fiabilidade e nivel de calidade

A tradución automática experimentou un avance considerable nos últimos anos. A tradución deste tipo entre linguas afíns, como pode ser entre o galego e o castelán, ou o francés, ou o italiano, ofrece hoxe xa unha fiabilidade bastante alta, de xeito que o resultado ofrecido pódese considerar case un produto final aceptable, aínda que sempre necesitado dunha revisión final. Na actualidade, por exemplo, hai xornais que se traducen automaticamente, cunha achega de revisión realmente reducida.

Pero este nivel de calidade non se acada aínda cando se trata de linguas tipoloxicamente distanciadas, como pode ser o caso do galego e o inglés ou o galego e o alemán. Con este tipo de linguas a tradución automática utilízase polo de agora máis ben como unha ferramenta auxiliar de accesibilidade a contidos para persoas que descoñecen a lingua fonte, pero non como un produto de calidade lingüística contrastada.

3.1.2. Desconfianza inxustificada e prexuízos respecto á TA

Non diría toda a verdade se non me fixese eco de certa desconfianza por parte de moitos sobre os resultados ofrecidos polas ferramentas que permiten a versión dunha lingua a outra sen intervención humana.

Esta desconfianza, como en xeral calquera estado de opinión, ten unha etioloxía que convén analizar, porque algunhas das razóns que a produciron responden a certas realidades, aínda que estas sexan percibidas dunha maneira non totalmente obxectiva.

Cales son as causas que deron lugar a esta desconfianza? Eu sinalaría, entre outras, as seguintes:

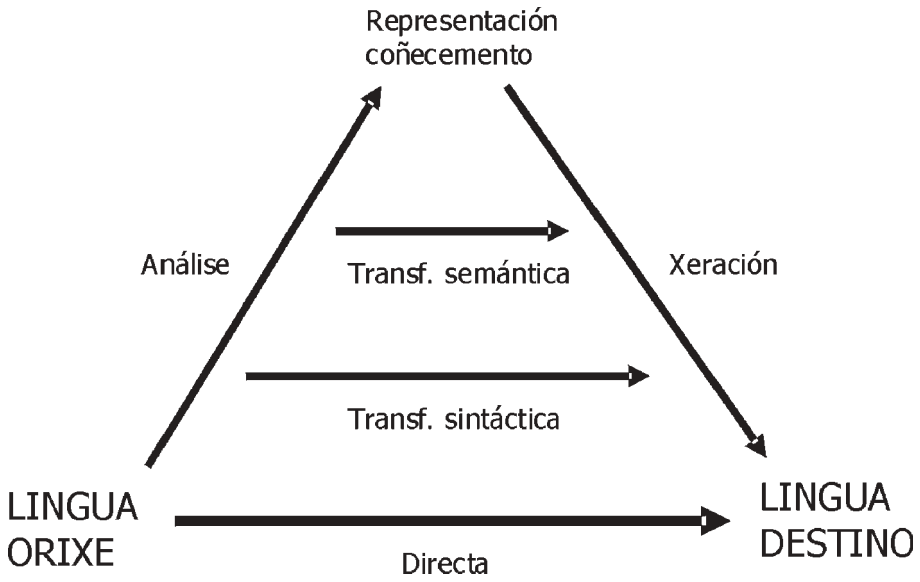
a) Non dicir sempre a verdade sobre o que ofrece a tradución automática. A tradución automática abre unhas enormes posibilidades, aforra moitísimo traballo, pero, sobre todo, entre linguas de distintas familias, ofrece resultados interesantes, pero non óptimos.
b) Outro motivo que xerou mesmo escepticismo sobre as posibilidades destes sistemas foi a proliferación de produtos que acadaron unha notable difusión, pero dun nivel de calidade realmente baixo.
c) Tamén se lle apón á tradución automática, que produce textos sen valor literario, e isto é verdade. Este tipo de tradución baséase en automatismos que nunha mesma situación xeran un mesmo resultado; aínda que tamén é posible lograr certa variedade estilística, programando a xeración de resultados distintos distribuídos seguindo unha orde determinada ou dunha maneira aleatoria. E esta é unha concepción oposta á que está na base de moitos recursos literarios, que pretenden xustamente o retorcemento ou mesmo a ruptura do código. Polo tanto hai certos recursos literarios que só poderán ser plasmados nunha tradución mediante a intervención humana, e nin sequera a de calquera humano, senón a dunha persoa que teña a formación e a sensibilidade necesaria, primeiro para captalos no texto de orixe e, posteriormente, para lograr transferilos, noutro proceso de creación, á lingua de destino.

Pero isto non invalida a importancia social da tradución automática. Debemos ter en conta que só o 3% da tradución que se fai no mundo corresponde a obras literarias. O resto son textos diplomáticos, administrativos, comerciais, técnicos..., moitos deles mecánicos e rutineiros.

3.1.3. Métodos máis utilizados na tradución automática

Nos sistemas de tradución automática utilízanse fundamentalmente dous métodos: os métodos baseados en regras (ou tradución por transferencia) e os métodos baseados en modelos estatísticos. Recentemente son frecuentes outros deseños de carácter mixto, que combinan complementariamente o método por regras e o método estatístico.

O método de transferencia ou método baseado en regras é o máis utilizado nos sistemas comerciais de tradución. Fundaméntase no establecemento de correspondencias entre distintas estruturas e a diferentes niveis: a nivel de palabra (e daquela fálase de transferencia léxica), a nivel sintáctico (transferencia sintáctica) e a nivel semántico (transferencia semántica). Este método implica o establecemento de correspondencias entre diferentes niveis de análise das dúas linguas. O polígono de Vaquois móstranos estes distintos niveis de transferencia:



En todo proceso de tradución automática contamos sempre con dous produtos: un texto de entrada nunha lingua A e un texto de saída nunha lingua B:

A	B
Ayer por la mañana los niños subieron a la cumbre más alta de la montaña	Onte pola mañá os nenos subiron ao cume máis alto da montaña

Para percorrer este proceso é necesario realizar na lingua de partida: unha análise léxica, unha análise morfolóxica, unha análise sintáctica e unha análise semántica; e a continuación formular as regras que permitan establecer as correspondencias entre a lingua de partida e a lingua de chegada. Canto máis alto é o nivel de transferencia, maior é a complexidade da análise do texto que hai que realizar. A maior parte dos sistemas actuais chegan pouco máis lonxe que a unha análise sintáctica relativamente simple.

3.1.4. Ferramentas de tradución automática para a lingua galega

Voume referir a tres produtos que, ao meu xuízo, reúnen o mínimo de calidade para renderen un servizo á sociedade galega: O *Es-gl, Traduza* e *Apertium.*

3.1.4.1. O Es-gl

É un sistema de tradución desenvolvido no Centro Ramón Piñeiro para a Investigación en Humanidades, coa colaboración da empresa Incyta. É un programa que de mo-

mento só traduce textos do castelán ao galego, cun nivel de calidade realmente alto. O sistema utilizado é un herdeiro directo do que fora o sistema de tradución automática METAL. Este sistema utiliza técnicas de enxeñaría lingüística: considera toda a oración castelá no seu conxunto, fai unha análise morfolóxica, sintáctica e, na medida do posible, semántica dela, e vértea ao galego dun xeito apropiado. Leva incorporado unha serie de ferramentas:

- un dicionario monolingüe castelán, que ten información morfolóxica, sintáctica e semántica de cada palabra.
- unha gramática de análise do castelán, que detecta cál é a estrutura sintáctica da oración en castelán.
- unha gramática de transferencia castelán-galego, que converte as estruturas lingüísticas da lingua orixe (castelán) en estruturas lingüísticas propias do galego.
- un dicionario bilingüe castelán-galego, que establece as correspondencias palabra a palabra entre o castelán e o galego, pero que tamén realiza comprobacións sobre o contexto no que se atopa a voz castelá para ditaminar a acepción correcta e para examinar os trazos sintácticos e semánticos das unidades lingüísticas implicadas.
- unha gramática de xeración do galego, que se ocupa da estrutura final do texto en galego; e
- un dicionario monolingüe galego que, coma no caso do castelán, está enriquecido con información morfolóxica, sintáctica e semántica.

O sistema tamén dispón doutros programas auxiliares que se encargan da segmentación do texto en frases e do tratamento dos formatos

3.1.4.2. Traduza

É un sistema de tradución de calidade bastante boa, realizado a partir de iniciativa privada, e construído pola empresa Dimensiona de Santiago de Compostela.

3.1.4.2. Apertium

É un tradutor que acaba de presentarse hai poucos días, e é o resultado dun proxecto de investigación financiado polo programa Profit do Ministerio de Industria, Turismo e Comercio. O proxecto foi coordinado pola empresa Eleka de Euskadi, e nel participan ademais: a Fundación Elhuyar, a Universidade do País Vasco, a empresa Thera, a Universidade Politécnica de Catalunya, a Universidade de Alacante, Imaxín, e a Universidade de Vigo. Este proxecto ten por finalidade a creación dun sistema que permita a tradución automática entre todas as linguas oficiais do Estado.

A versión actual, que pode ser probada en liña, non ofrece aínda uns resultados de nivel alto, pero estamos seguros de que co paso do tempo mellorará de maneira notable.

Para min esta iniciativa resulta especialmente atractiva por diversas razóns: é un exemplo de investigación interdisciplinaria (na que participan informáticos, lingüistas e enxeñeiros de telecomunicacións), é o resultado da colaboración de grupos diversos (empresas e Universidades de distintos puntos do Estado), anúnciase que se presentará en código aberto, e é de distribución libre.

3.2. A síntese de voz en galego: *Cotovía*

Os sintetizadores de voz ou conversores texto-voz son ferramentas que permiten a conversión dun texto escrito nunha cadea oral, de xeito que a transferencia texto-voz poida ser levada a cabo cunha calidade aceptable sen a intervención directa do falante.

Para a lingua galega existen neste momento dúas ferramentas deste tipo: unha, de carácter privado, propiedade de Telefónica, e outra, construída no Centro Ramón Piñeiro para a Investigación en Humanidades, obra dun equipo de investigación interdisciplinario integrado por enxeñeiros de telecomunicacións da Universidade de Vigo (primeiro baixo a dirección de Mª Carme García Mateo, e despois de Eduardo Rodríguez Banga) e por lingüistas da Universidade de Santiago (baixo a dirección de Manuel González González). Este último sistema elaborado no Centro Ramón Piñeiro é coñecido co nome de *Cotovía*, e pode ser probado en liña no enderezo *http://www.cirp.es/res/index.html*.

Un conversor texto-voz debe cumprir unha esixencia mínima, que é a da intelixibilidade, pero será máis atractivo canto máis se achegue ao modelo de fala humana e os seus resultados sexan percibidos polo usuario como algo que se aproxima á fala natural.

Desde o punto de vista científico a investigación en certo tipo de síntese de voz é dunha grande importancia na lingüística, porque permite validar os trazos realmente pertinentes na emisión dun fonema. Pero o desenvolvemento que se produciu nos últimos anos da síntese de voz non se debe a este interese científico, senón ao feito de que se mostrou un instrumento de primordial importancia no mundo da comunicación e das relacións sociais, ao dar lugar á creación de postos de información automatizados, á instauración de sistemas automáticos integrais de recoñecemento, tradución e síntese, e un amplo elenco de funcións derivadas da lectura automática. É de salientar tamén a importancia que os desenvolvementos da conversión texto-voz ten no mundo das minusvalías, por exemplo na creación de lectores para invidentes, que lles facilitan o acceso a moitas fontes escritas que antes lles resultaban inaccesibles, a non ser que estivesen en sistema Braille.

Hoxe estanse realizando fortes investimentos na investigación de síntese e recoñecemento de voz en todas as linguas, polo papel transcendente que teñen xa na sociedade actual, pero sobre todo polo que van ter na sociedade do futuro próximo. O desenvolvemento neste eido é aínda máis transcendental para as linguas que están en proceso de normalización, porque sen dúbida o desenvolvemento ou non deste tipo de tecnoloxías para as linguas en situación delicada vai ser un factor decisivo para que os falantes consideren que a súa lingua ten capacidade para dar resposta ás súas necesidades ou non. Por iso se ten dito tantas veces que as linguas minorizadas que non sexan capaces de subir no carro das novas tecnoloxías terán graves problemas de pervivencia.

Cotovía é un sintetizador baseado na concatenación de unidades pregravadas (*difonos*). E este é hoxe o método máis utilizado, xa que é o que ofrece unha mellor relación entre complexidade e prestacións. A utilización de difonos simplifica moito os problemas derivados da coarticulación.

Hoxe baixo a etiqueta de *Cotovía* preséntanse en realidade dous prototipos diferentes: un construído sobre unidades extraídas de fala natural e outro sobre unidades extraídas de logátomos. O primeiro ofrece unha pronunciación máis distendida e natural, pero presenta certos problemas derivados sobre todo da presenza de realizacións excesivamente relaxadas. O segundo, con unidades extraídas de logátomos, posúe unha modulación máis mecánica e menos natural, pero unha maior robustez, debido a que as unidades están extraídas de contextos especialmente seleccionados e coidados, que permiten realizacións máis uniformes e unha etiquetaxe máis exacta das unidades.

Na súa presentación actual *Cotovía* é xa un sistema bastante flexible, que permite escoller para o seu funcionamento entre os dous tipos de prototipos mencionados, pero que permite tamén escoller entre voces masculinas e femininas, e permite manipular a velocidade de dicción e a frecuencia fundamental, segundo o usuario prefira unha dicción máis veloz ou máis lenta e unha voz máis grave ou máis aguda.

3.3. O recoñecemento de voz en galego

Referireime brevemente a dous proxectos de investigación relacionados co recoñecemento de voz en lingua galega: o primeiro deles co título *Desenvolvemento de sistemas de diálogo para acceso telefónico a servizos telemáticos (TelCorreo)*, e o segundo *Transcrigal.*

3.3.1. Desenvolvemento de sistemas de diálogo para acceso telefónico a servizos telemáticos (TelCorreo)

A finalidade deste proxecto era a de construír unha ferramenta que permita a consulta de correo electrónico vía telefónica sen intervención humana. Imaxinemos que un domingo imos pasar a tarde á praia, e queremos consultar o correo electrónico: con este sistema podemos facelo, marcando o número dun servidor. O sistema faranos saber automaticamente que correos recibimos, cunha información básica de cada un deles (remitente, asunto e hora de recepción), e nós poderemos pedirlle a lectura do contido daquel ou daqueles que nos interesen. Igualmente o sistema tamén nos permite responder oralmente a estes correos ou enviar oralmente un novo ao destinatario que desexemos.

Este foi un proxecto realizado fundamentalmente con fondos da Unión Europea, no que participaron a Universidade de Santiago, a Universidade de Vigo, o Centro Ramón Piñeiro para a Investigación en Humanidades, e Cesatel (unha empresa de Servizos Avanzados de Telecomunicación e Tratamento de Información). Foi desenvolvido por un equipo interdisciplinario de enxeñeiros de telecomunicación da Universidade de Vigo e de lingüistas da Universidade de Santiago, codirixido por M.ª C. García Mateo e Manuel González González.

3.3.2. Transcrigal

É un proxecto encamiñado a lograr un transcritor automático de noticias para a TV, que permita nun futuro a subtitulación automática de programas informativos de TV. O desenvolvemento deste tipo de utilidades encádrase dentro das recomendacións da UE de subtitular os programas de TV para facilitar a integración de persoas con minusvalías, coma os xordomudos.

Este proxecto tivo ata o momento un financiamento realmente escaso e totalmente insuficiente para lograr os obxectivos para os que estaba previsto. O financiamento que se obtivo procedía da CICYT, da Xunta de Galicia e da CRTVG.

Logrouse un índice de recoñecemento correcto relativamente alto nos locutores habituais dos servizos informativos, pero os resultados eran aínda baixos para as persoas que aparecían de maneira ocasional nestes informativos. Esta deficiencia debíase a que o corpus de adestramento do sistema non era o suficientemente amplo como para permitir uns niveis globais aceptables. En realidade, practicamente toda a subvención recibida dos organismos colaboradores foi destinada á conformación deste corpus de adestramento, tanto oral coma escrito.

3.4. A investigación lingüística ocupa un lugar importante no desenvolvemento destes sistemas

Alguén poderá preguntarse que relación ten isto coa investigación lingüística en galego. A resposta é moi simple: a investigación lingüística é fundamental para o desenvolvemento destes sistemas. Non hai tradución automática, síntese de voz, recoñecemento de voz, con todas as súas aplicacións derivadas, sen investigación lingüística.

A investigación lingüística aplicada ás tecnoloxías da fala é aínda moi deficitaria en galego, e presenta importantes problemas de falta de recursos básicos, de inexistencia de planificación e coordinación, en fin de desestruturación, que leva a que existan fortes lagoas en determinados ámbitos, duplicacións e triplicacións de esforzos illados que producen resultados insuficientes, e desaproveitamento de logros acadados en proxectos, que nunca foron explotados fóra do ámbito no que foron concibidos.

Neste momento necesítase con urxencia:

a) A creación de grandes bases de datos e córpora etiquetados con sistematicidade e fiabilidade, tanto orais, coma escritos.
   Actualmente existen para o galego tres córpora de carácter oral etiquetados: *SpeechDat* (que foi elaborado para o adestramento do sistema TelCorreo, e que está realizado seguindo a metodoloxía e especificacións internacionais dos sistemas SpeechDat), un corpus etiquetado de programas de noticias (que foi o utilizado para o adestramento de Transcrigal), e *Vogatel*, que é privado, propiedade de Telefónica, e que só é utilizable para proxectos da propia empresa). Cómpre a elaboración dun único corpus oral equilibrado, cun mínimo de

500 horas etiquetadas, que sexa de carácter público, e que poida ser utilizado libremente por calquera investigador.

Existen tamén córpora escritos, coma o *Tesouro informatizado da Lingua Galega (Tilga)* do Instituto da Lingua Galega, o *Corpus de referencia do galego actual (Corga)* do Centro Ramón Piñeiro para a Investigación en Humanidades, ou o *Corpus lingüístico da Universidade de Vigo (Cluvi)*. Pero son córpora, no fondo, concibidos con finalidade de explotación para o estudo do léxico, insuficientes polo seu volume, pero sobre todo porque non están dotados dun sistema de anotación completo que os faga aproveitables para o estudo da fonética, morfoloxía, sintaxe, léxico e pragmática. Non sería perder o tempo constituír un corpus escrito de varios centos de millóns de palabras, debidamente anotado, que cubrise as necesidades dun amplo número de investigadores. Doutro xeito, cada investigador verase obrigado a elaborar o seu propio corpus, que cubra as súas necesidades particulares, pero que será inútil para outros usos. Canto non se aforraría cun único corpus que cubrise as necesidades de todos ou dunha gran maioría, e ademais dunha maneira moito máis plena e eficaz?

– É necesario construír analizadores morfolóxicos e sintácticos automáticos, que sexan fiables e flexibles, de maneira que se poidan aplicar a usos e ferramentas moi diversas. Carece de sentido que se constrúa un analizador morfolóxico para tratar de anotar un corpus de maneira automática, outro distinto para aplicar a un sistema de tradución automática, un terceiro para un sintetizador de voz, un cuarto para a construción dun ditáfono e un quinto para auxiliar a un transcritor fonético automático. O lóxico sería construír un analizador robusto e flexible, que se poida integrar en todos estes sistemas.
– Son necesarios léxicos electrónicos con información semántica estruturada, tanto do galego común, como das chamadas linguas de especialidade. Sen eles dificilmente se pode avanzar en ningún dos sistemas de tecnoloxías da fala aos que nos viñemos referindo. Son igualmente urxentes léxicos electrónicos bilingües galego-inglés, galego-francés, galego-alemán, galego-ruso, galego-chinés, galego-catalán etc.; ata que dispoñamos destes utensilios non se poderá avanzar na tradución automática a estas linguas ou desde estas linguas.
– É necesaria unha organización da investigación. Somos un país pequeno e con recursos limitados, por iso hai que optimizar os esforzos:
    • Debemos fomentar a creación de equipos interdisciplinarios de carácter estable.
    • Cómpre levar a cabo unha distribución de tarefas entre grupos de investigación, sen que isto implique negar a liberdade de investigación de ninguén.

- É necesaria a modularidade. Hoxe case todos os sistemas presentan un alto grao de complexidade, no que actúan en paralelo ou no que se superpoñen distintos elementos interdependentes, que esixen unha alta especialización. Poñamos por caso un sistema de tradución automática, no que existe un módulo de análise morfosintáctica da lingua de partida, e un módulo de gramática de transferencia, ou un módulo de léxico. Por que non se vai especializar un determinado grupo nun destes módulos, de maneira que teña un coñecemento moito máis profundo dos problemas que presenta? E o mesmo poderiamos dicir na síntese de voz: pode haber un grupo especializado traballando nun analizador, e outro nun módulo prosódico, con independencia de que exista en moitos casos información que deban compartir.
- A modularidade facilita a polivalencia. Un bo silabador do galego serve tanto para aplicar a unha ferramenta de síntese ou recoñecemento de voz, como para informar sobre a silabación correcta dos lemas dun dicionario.
- Os recursos de investigación básica (como bases de datos, córpora...) deben ser de dispoñibilidade pública (polo menos para a investigación); de aí que neste tipo de recursos é totalmente necesaria a presenza do sector público, que é o único capaz de garantir que poida chegar aos destinatarios que os necesiten. O investimento en recursos básicos nunca é caro, porque a moi curto prazo repercute no aforro de tempo e de custos e evita a duplicación (ou n-plicación) de esforzos dedicados a un mesmo obxectivo.

4. CONCLUSIÓN

Internet e as novas tecnoloxías son unha arma de dobre fío para moitas linguas: poden contribuír decisivamente á normalización e á estabilización do galego, pero poden contribuír tamén á súa desaparición.

Hai datos que por si sós resultan arrepiantes, como que a metade das 6000 linguas faladas no mundo corren perigo de extinción, segundo a Unesco; ou que só o 11% da poboación mundial ten acceso a Internet. Como se puxo de manifesto no Cume Mundial sobre a Sociedade da Información, as TIC son fundamentais para o desenvolvemento das persoas e dos pobos, pero moitos pobos van quedar marxinados no desenvolvemento das tecnoloxías da información. O desenvolvemento das TIC non se pode deixar exclusivamente ao azar do mercado, é necesaria unha forte implicación dos poderes públicos que fomenten o desenvolvemento e aplicacións das TIC, e que garantan igualmente a adaptación das aplicacións a cada lingua e a cada cultura. No mundo industrializado un factor

decisivo para a supervivencia de moitas linguas vai ser a súa capacidade para incorporaren as novas tecnoloxías.

É totalmente imprescindible poñer en marcha todos os esforzos e todos os medios para consolidar unha posición digna de Galicia na sociedade actual e do futuro inmediato, na sociedade da tecnoloxía e da innovación, na sociedade da información, na sociedade do coñecemento. Se non somos quen a conseguilo, o progresivo crecemento do déficit tecnolóxico levaranos a ocupar unha posición recuada, unha posición de debilidade fronte ás sociedades máis desenvolvidas, que nos pode facer desembocar na perda da nosa identidade:

a) Por sermos incapaces obxectivamente de responder ás necesidades vitais da sociedade desde a nosa idiosincrasia e desde a nosa lingua.
b) Porque esta incapacidade xera, primeiro, falta de confianza en nós, despois aceptación acrítica dos modelos lingüísticos e doutro tipo que nos veñen de fóra, pero que nos facilitan acceso aos avances e ao benestar, e, por último, abandono e mesmo desprezo do propio por consideralo inútil e arcaico.

Estamos aínda a tempo, pero cómpre visión de futuro, esforzo, ilusión e racionalidade. Só así estaremos en condicións de construírmos a Galicia do mañá.