D2.3.gl
Language Report for Galician
(Galician version)

Version 1.0

2011-09-07

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	07-09-2011	Carmen García-Mateo, Montserrat Arza	UVIGO	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

METANET4U

D2.3.gl Language Report for Galician (Galician version)

Document METANET4U-2011-D2.3.gl
EC CIP project #270893

Deliverable
Number: D2.3.gl
Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: University of Vigo; Universitat Politècnica de Catalunya

Authors: Carmen García-Mateo, Montserrat Arza

Reviewer: Asunción Moreno

Índice de contidos

Sumario

Moitas linguas europeas corren o risco de se converter en vítimas da era dixital debido á súa escasa presenza e á falta de recursos na Internet. Moitas oportunidades de mercado local continúan sen ser explotadas por mor das barreiras lingüísticas. Se non tomamos medidas agora, moitos cidadáns europeos veranse prexudicados polo feito de falar o seu idioma nativo.

A innovadora tecnoloxía lingüística (TL) é un intermediario que permitirá aos cidadáns europeos participar nunha sociedade da información e do coñecemento igualitaria, global, e triunfante a nivel económico. A tecnoloxía lingüística plurilingüe abrirá as portas cara a unha comunicación instantánea, barata e fluída, e cara a unha interacción que supere os problemas lingüísticos.

Hoxe en día, os principais provedores de servizos lingüísticos son empresas estadounidenses. O tradutor de *Google*, un servizo gratuíto, é só un exemplo entre moitos. O éxito recente de *Watson*, un sistema informático creado por IBM que gañou unha das entregas do programa televisivo *Jeopardy* contra opoñentes humanos, demostra o inmenso potencial da tecnoloxía lingüística. Como europeos, debémonos formular certas cuestións urxentes:

- Deben depender as nosas comunicacións e infraestruturas de coñecemento de compañías monopolistas?
- Podemos confiar verdadeiramente en servizos lingüísticos que poden ser inmediatamente desactivados por outros?
- Estamos a competir dun xeito activo no mercado global pola investigación e o desenvolvemento na tecnoloxía lingüística?
- Están dispostos dende os outros continentes a abordar os nosos problemas de tradución e outros asuntos que teñan relación co plurilingüismo europeo?
- Pode a nosa bagaxe cultural europea axudar a conformar a sociedade do coñecemento ofrecendo unha tecnoloxía de gran calidade que sexa mellor, máis segura, máis precisa, máis innovadora e máis robusta?

Este libro branco sobre o idioma galego demostra que, no caso desta lingua, a investigación e a industria da tecnoloxía lingüística son bastante limitadas. A pesar de que existen tecnoloxías e recursos para o galego, o número é inferior que as que existen para o inglés. Ademais, as devanditas tecnoloxías e recursos tamén son de menor calidade.

Segundo a avaliación que se detalla neste informe, debe levarse a cabo unha acción inmediata antes de que se poidan conseguir avances para o idioma galego.

META-NET contribúe á construción dun espazo de información dixital europeo plurilingüe. Acadando este obxectivo, pode prosperar unha unión de nacións multicultural e converterse nun modelo de cooperación internacional pacífica e igualitaria. Se non se pode alcanzar este obxectivo, Europa terá que elixir entre sacrificar as súas identidades culturais ou sufrir unha derrota económica.

Un risco para as nosas linguas e un reto para a tecnoloxía lingüística

Hoxe en día somos testemuñas dunha revolución dixital comparable á invención da imprenta de Gutenberg.

Como mostran os recentes eventos sucedidos en África do Norte, estamos a presenciar unha revolución dixital que ten importantes repercusións nas comunicacións e na sociedade. Os recentes avances na tecnoloxía da comunicación dixital e en rede equipáranse ás veces coa invención da imprenta de Gutenberg. Que é o que nos di esta analoxía sobre o futuro da sociedade europea da información y das nosas linguas en particular?

Tras a invención de Gutenberg, acadáronse grandes avances na comunicación e no intercambio de coñecemento grazas a esforzos como a tradución de Lutero da Biblia á lingua común. Nos séculos posteriores, desenvolvéronse técnicas culturais para unha mellor xestión do procesamento da linguaxe e do intercambio de coñecemento:

- a normativización ortográfica e gramatical dos principais idiomas permitiu unha rápida difusión de novas ideas científicas e intelectuais;
- a evolución dos idiomas oficiais fixo posible que os cidadáns se puidesen comunicar dentro de certas fronteiras (a miúdo políticas);
- o ensino e a tradución das linguas permitiu un intercambio entre idiomas;
- a creación de directrices xornalísticas e bibliográficas asegurou a calidade e dispoñibilidade do material impreso;
- a creación de diferentes medios de comunicación como os xornais, a radio, a televisión, os libros, e outros formatos, satisfixeron as distintas necesidades comunicativas.

Nos últimos vinte anos, a tecnoloxía da información, ou informática, veu axudando á hora de automatizar e facilitar moitos procesos:

- os programas informáticos de autoedición substitúen á mecanografía e á redacción;
- Microsoft PowerPoint substitúe ás transparencias nos retroproxectores;
- o correo electrónico envía e recibe documentos habitualmente máis rápido cás máquinas de fax;
- Skype permite realizar chamadas telefónicas vía Internet e organizar reunións virtuais;
- os formatos de codificación de audio e vídeo facilitan o intercambio de contidos multimedia;
- os motores de busca proporcionan acceso ás páxinas web baseándose en palabras clave;
- os servizos en liña, como o tradutor de Google, realizan traducións rápidas e aproximadas;
- as plataformas sociais multimedia facilitan a colaboración e o intercambio de información.

A pesar de que moitas destas ferramentas e aplicacións resultan moi útiles, son suficientes para poñer en marcha unha sociedade europea

da información plurilingüe, unha sociedade moderna e global onde a información e os bens poidan circular libremente?

As fronteiras lingüísticas obstaculizan á sociedade da información europea

Non podemos saber con exactitude como será a sociedade da información futura. Á hora de debater unha estratexia enerxética común europea ou unha política exterior, é posible que queiramos que os ministros de exteriores europeos falen nos seus idiomas nativos. Tamén pode que queiramos unha plataforma na que a xente que fale idiomas diferentes e que teña diversas competencias lingüísticas poidan debater un tema concreto mentres a tecnoloxía recolle automaticamente as súas opinións e xera breves resumos. É posible tamén que queiramos falar coa asistencia dun seguro médico situado nun país estranxeiro.

A economía global e o espazo de información enfróntannos con máis idiomas, falantes e contidos.

É evidente que as necesidades comunicativas hoxe en día son moi diferentes ás de hai uns anos. Nun contorno de economía global e espazo de información, vémonos expostos a máis idiomas, falantes e contidos, e isto require que sexamos capaces de interactuar con rapidez cos novos medios de comunicación. A actual popularidade das plataformas sociais multimedia (Wikipedia, Facebook, Twitter e YouTube) e só a punta do iceberg.

Hoxe en día podemos transmitir xigabytes de texto por todo o mundo en cuestión de segundos antes de que recoñezamos que está nun idioma que non comprendemos. Segundo un informe recente solicitado pola Comisión Europea, o 57% dos usuarios de Internet en Europa mercan bens e servizos en idiomas diferentes aos seus idiomas nativos. (O inglés é a lingua estranxeira máis común, seguida do francés, o alemán e o español). O 55% de usuarios le contidos en idiomas estranxeiros, mais só un 35% empregan outra lingua para escribir correos electrónicos ou deixar comentarios na web.[1] Hai uns anos, o inglés podía ser considerado a lingua vehicular da Internet, xa que a grande maioría dos seus contidos estaban escritos neste idioma. Non obstante, a situación actual é moi diferente. A cantidade de contido que aparece na Internet noutros idiomas (especialmente idiomas asiáticos e árabes) disparouse.

Que idiomas europeos prosperarán e persistirán na información en rede e na sociedade do coñecemento?

Sorprendentemente, no debate social non se lle prestou demasiada atención á brecha dixital ubicua causada polas fronteiras lingüísticas; así e todo, isto presenta unha cuestión urxente: que idiomas europeos prosperarán e persistirán na información en rede e na sociedade do coñecemento?

As nosas linguas en perigo

A imprenta contribuíu a un intercambio moi valioso de información en Europa, pero tamén levou consigo a extinción de moitos idiomas europeos. Raras veces se imprimía nas linguas rexionais e minoritarias. Como resultado, moitos idiomas, como o córnico ou o dálmata, víronse limitados á transmisión oral, o cal restrinxiu a súa adopción, difusión e uso continuados.

[1] Dirección Xeral da Sociedade da Información e Medios de Comunicación da Comisión Europea, *User language preferences online* (Preferencias lingüísticas dos usuarios da Internet), Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

As aproximadamente 60 linguas que existen en Europa son un dos seus bens culturais máis ricos e importantes. A multitude de idiomas de Europa constitúe un elemento fundamental do seu éxito social[2]. Mentres que outros idiomas populares, como o inglés ou o chinés [FIXME: Español], manterán de seguro a súa presenza na sociedade e no mercado dixital emerxente, moitos idiomas europeos poderían verse illados a causa das comunicacións dixitais e poderían chegar a converterse en idiomas irrelevantes na sociedade da Internet. Naturalmente, tales sucesos serían pouco desexables. Por unha banda, perderíase unha oportunidade estratéxica, o cal debilitaría a posición global de Europa. Por outra banda, tales acontecementos entrarían en conflito co obxectivo da participación igualitaria de cada cidadán europeo, independentemente do seu idioma. Segundo un informe da UNESCO sobre o plurilingüismo, os idiomas son un medio imprescindible para o gozo dos dereitos fundamentais, tales como a expresión política, a educación, e a participación na sociedade.[3]

A ampla variedade de linguas en Europa é un dos seus bens culturais máis importantes e constitúe un elemento fundamental do éxito de Europa.

A tecnoloxía lingüística é unha tecnoloxía instrumental clave

No pasado, as inversións centrábanse no ensino de idiomas e na súa tradución. Por exemplo, segundo algunhas estimacións, o mercado europeo para a tradución, a interpretación, a localización de *software*, e a globalización das páxinas web era, no ano 2008, de 8,4 miles de millóns de euros, e esperábase un crecemento dun 10% anual.[4] Non obstante, estes recursos existentes non son suficientes para satisfacer as necesidades actuais e as futuras.

A tecnoloxía lingüística é unha tecnoloxía instrumental clave que axuda a protexer e a promover os idiomas europeos. A tecnoloxía lingüística permite que os cidadáns colaboren, fagan negocios, compartan coñecementos, e participen nos debates sociais e políticos independentemente das barreiras lingüísticas ou dos coñecementos informáticos. A tecnoloxía lingüística hoxe en día axúdanos en tarefas cotiás, como escribir un correo electrónico, buscar información na rede, ou reservar un voo.

A tecnoloxía lingüística permite que os cidadáns colaboren, fagan negocios, compartan coñecementos, e participen nos debates sociais e políticos empregando idiomas diferentes.

Estamos a nos beneficiar da tecnoloxía lingüística cando:

- buscamos e traducimos páxinas web;
- empregamos as opcións de corrección gramatical e ortográfica nos procesadores de texto;
- lemos as recomendacións dun produto nunha tenda *online*;
- escoitamos as instrucións verbais dunha voz sintética nun sistema de navegación;
- traducimos páxinas web cun servizo *online*.

As tecnoloxías lingüísticas detalladas neste informe constitúen unha parte fundamental das futuras aplicacións innovadoras. A tecnoloxía

A tecnoloxía lingüística pode considerarse como o sistema operativo que permite a interacción entre o usuario e o contido.

[2] Comisión Europea, *Plurilingualism: an asset for Europe and a shared commitment* (Plurilingüismo: unha vantaxe para Europa e un compromiso compartido), Bruxelas, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[3] Director Xeral da UNESCO , *Intersectoral mid-term strategy on languages and multilingualism* (Estratexia intersectorial a medio prazo sobre os idiomas e o plurilingüismo), París, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[4] Dirección Xeral de Tradución da Comisión Europea, *Size of the language industry in the EU* (Tamaño da industria lingüística na UE), Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

lingüística é, polo xeral, unha tecnoloxía instrumental que se sitúa dentro dun marco de aplicación máis amplo, como un sistema de navegación ou un motor de busca. Estes libros brancos céntranse na capacidade das tecnoloxías básicas en cada lingua.
Nun futuro próximo, necesitaremos que estea dispoñible unha tecnoloxía lingüística para todas as linguas europeas, que sexa economicamente alcanzable e que se atope perfectamente integrada dentro duns ámbitos informáticos máis amplos. Sen a tecnoloxía lingüística, non se pode acadar unha experiencia interactiva, multimedia e plurilingüe dos usuarios.

Oportunidades para a tecnoloxía lingüística

A tecnoloxía lingüística fai posible a a tradución automática, a produción de contidos, o procesamento da información e a xestión do coñecemento para todos os idiomas europeos. A tecnoloxía lingüística tamén pode potenciar o desenvolvemento de interfaces lingüísticas intuitivas para os electrodomésticos, a maquinaria, os vehículos, os ordenadores e os robots. A pesar de que xa existen moitos prototipos, as aplicacións comerciais e industriais aínda están nas primeiras fases do desenvolvemento. O ritmo actual do progreso presenta unha verdadeira oportunidade, tendo en conta que a investigación veu progresando de maneira constante durante os últimos anos. Por exemplo, a tradución automática (TA) hoxe en día é capaz de proporcionar unha precisión bastante aceptable en certos ámbitos específicos, e as aplicacións experimentais ofrecen unha xestión da información e o coñecemento plurilingüe, así como unha produción de contidos en moitas linguas europeas.

As aplicacións lingüísticas, as interfaces de usuario baseadas en voz, e os sistemas de diálogo adoitan atoparse en ámbitos altamente especializados, e a miúdo mostran un rendemento limitado. Un ámbito de investigación existente é o uso da tecnoloxía lingüística nas operacións de rescate en zonas de catástrofe. En tales situacións de alto risco, a precisión na tradución pode ser cuestión de vida ou morte. O mesmo razoamento pódese aplicar ao uso da tecnoloxía lingüística na industria sanitaria. Os robots intelixentes con capacidades lingüísticas en varios idiomas poden servir para salvar vidas.

Existen grandes oportunidades de mercado nas industrias da educación e do espectáculo para a integración das tecnoloxías lingüísticas en xogos, ofertas de lecer educativo, o ámbito da simulación ou programas de formación. Os servizos de información móbil, os programas de aprendizaxe de idiomas asistidos por ordenador, os ámbitos de aprendizaxe electrónica, as ferramentas de autoavaliación, e os programas de detección de plaxio son só uns cantos exemplos máis onde a tecnoloxía lingüística pode desempeñar un papel importante. A popularidade de aplicacións sociais multimedia, como Twitter e Facebook, suxiren unha necesidade de tecnoloxías lingüísticas sofisticadas que poidan supervisar as publicacións, resumir os debates, indicar as opinións públicas, detectar as respostas emocionais, identificar as violacións dos dereitos de autor, ou facer un seguimento de uso indebido.

A tecnoloxía lingüística representa unha grande oportunidade para a Unión Europea que non se aplica só a nivel económico senón tamén a

O plurilingüismo é a regra, non a excepción.

nivel cultural. O plurilingüismo en Europa converteuse na regra. As empresas, organizacións e escolas europeas tamén son multinacionais e diversas. Os cidadáns queren comunicarse máis aló das fronteiras lingüísticas que aínda existen no Mercado Común europeo. A tecnoloxía lingüística pode axudar a superar estas barreiras aínda existentes, defendendo á vez un uso libre e aberto da linguaxe. Ademais, unha tecnoloxía lingüística plurilingüe e innovadora para Europa pode tamén axudarnos a comunicarnos cos nosos socios globais e as súas comunidades plurilingües. As tecnoloxías lingüísticas colaboran á prosperidade das oportunidades económicas internacionais.

Os retos que afronta a tecnoloxía lingüística

A pesar de que a tecnoloxía lingüística fixo considerables avances nos últimos anos, o ritmo actual do proceso tecnolóxico e da innovación dos produtos é demasiado lento. Non podemos agardar dez ou vinte anos a que se leven a cabo melloras significantes que poidan fomentar a comunicación e a produtividade no noso ámbito plurilingüe.

O ritmo actual do progreso tecnolóxico é demasiado lento como para chegar aos programas informáticos importantes nos próximos dez ou vinte anos.

As tecnoloxías lingüísticas de uso xeneralizado, tales como as funcións de corrección gramatical e ortográfica en procesadores de textos, son habitualmente monolingües e só están dispoñibles para certos idiomas. As aplicacións para a comunicación plurilingüe requiren un certo nivel de complexidade. A tradución automática e os servizos en liña como o tradutor de Google ou o tradutor de Bing son excelentes á hora de crear unha boa aproximación aos contidos dun documento. Pero tanto estes servizos *online* como as aplicacións profesionais de tradución automática presentan dificultades varias cando se necesitan traducións moi precisas e completas. Hai moitos exemplos coñecidos de malas traducións, como por exemplo as traducións literais dos nomes Bush ou Kohl, que poñen de manifesto os retos que aínda ten que afrontar a tecnoloxía lingüística.

A aprendizaxe das linguas

Para demostrar como xestionan a linguaxe os ordenadores e por que a aprendizaxe dun idioma é unha tarefa tan complicada, botamos unha pequena ollada ao xeito en que os seres humanos aprenden un primeiro e un segundo idioma, e despois facemos un bosquexo de como funcionan os sistemas de tradución automática; é por isto que o ámbito da tecnoloxía lingüística está moi relacionado co eido da intelixencia artificial.

Os seres humanos adquiren as habilidades lingüísticas de dúas maneiras diferentes. Primeiro, un bebé aprende o seu idioma nativo a través dos exemplos. O contacto con mostras lingüísticas concretas por parte de falantes nativos do idioma, como son os pais, irmáns e outros membros da familia, axuda a que os bebés produzan as súas primeiras palabras e locucións curtas aproximadamente a partir dos dous anos. Isto só é posible grazas a unha disposición xenética especial que teñen os seres humanos para a aprendizaxe do seu primeiro idioma.

Os seres humanos adquiren as habilidades lingüísticas de dúas maneiras diferentes: aprendendo exemplos e aprendendo as regras básicas do idioma.

A aprendizaxe dun segundo idioma normalmente require moito máis esforzo. Durante a idade escolar, os idiomas estranxeiros adoitan aprenderse a través do estudo da súa estrutura gramatical, o vocabulario, e a ortografía con libros e materiais educativos que describen o

coñecemento lingüístico no que respecta ás regras abstractas, ás táboas e aos textos de exemplo. A aprendizaxe dun idioma estranxeiro require moito tempo e esforzo, e faise máis difícil coa idade.

Os dous tipos principais de sistemas de tecnoloxía lingüística adquiren o idioma dun xeito semellante aos seres humanos.

Os dous tipos principais de sistemas de tecnoloxía lingüística adquiren as capacidades lingüísticas dun xeito semellante aos seres humanos. Os enfoques estatísticos obteñen coñecementos lingüísticos das extensas coleccións de textos con exemplos concretos nun só idioma ou nos chamados textos paralelos que están dispoñibles en dous ou máis idiomas. Os algoritmos de aprendizaxe automática modelan un certo tipo de facultade lingüística que é capaz de obter pautas sobre o correcto uso de palabras, locucións curtas e frases completas nun só idioma ou traducidas dun idioma a outro. En termos numéricos, a cantidade de frases que requiren os enfoques estatísticos é enorme. A calidade do rendemento aumenta segundo aumenta o número de textos analizados. Estes sistemas adéstranse normalmente con textos que conteñen millóns de frases. Está é unha das razóns polas cales os provedores de motores de busca son tan ávidos de compilar a maior cantidade de material escrito posible. A corrección ortográfica nos procesadores de texto, na información dispoñible *online*, e nos servizos de tradución, como o servizo de busca de Google ou o tradutor de Google, baséanse nun enfoque estatístico (operan a partir de bases de datos).

Os sistemas baseados en normas son o segundo tipo máis importante de tecnoloxía lingüística. Os expertos en lingüística, en lingüística computacional e en informática codifican as análises gramaticais (normas de tradución) e crean listas de vocabulario (lexicóns). O establecemento dun sistema baseado en normas require moito tempo e un labor intensivo. Os sistemas baseados en normas tamén requiren expertos altamente especializados. Algúns dos principais sistemas de tradución automática baseados en normas levan en constante desenvolvemento dende hai máis de vinte anos. A vantaxe dos sistemas baseados en normas é que os expertos poden obter un control máis detallado sobre o tratamento da linguaxe. Isto fai que sexa posible corrixir erros sistematicamente no programa informático, así como proporcionar comentarios detallados ao usuario, especialmente cando estes sistemas baseados en normas se empregan para a aprendizaxe de idiomas. Debido ás limitacións financeiras, a tecnoloxía lingüística baseada en normas só é viable para as linguas principais.

O galego na sociedade europea da información

Datos xerais

O galego pertence á familia das linguas románicas. É a lingua cooficial na rexión española de Galicia. Galicia ten máis de 2.800.000 de habitantes. Aproximadamente dous millóns de persoas son falantes habituais de galego e medio millón máis emprégao como segunda lingua[5],[6].

O galego é o idioma histórico de Galicia. O documento máis antigo escrito en galego e preservado en Galicia data do ano 1228, do reino de Alfonso IX, e atópase actualmente nos Arquivos da Casa de Alba en Madrid.

O territorio xeográfico da lingua galega está delimitada pola comunidade autónoma de Galicia e as áreas máis occidentais de Asturias, León e Zamora, ademais de tres pequenos lugares de Estremadura. Ademais diso, e polas circunstancias históricas da emigración da poboación galega por todo o mundo, existen áreas nas que hai unha ampla presenza de xente de orixe galego. Esta xente conservou a súa lingua como vehículo comunicativo, non só no ámbito privado, senón tamén no público, a través de publicacións periódicas, literarias ou mesmo na comunicación radiofónica dos países de acollida. Aínda existen grandes comunidades galegofalantes noutras rexións de España (Madrid, Barcelona, País Vasco e Illas Canarias), en Europa (Portugal, Francia, Suíza, Alemaña, Reino Unido e Holanda) e en América (Arxentina, Uruguai, Brasil, Venezuela, Cuba, México e Estados Unidos).

Galicia é, segundo recoñece a Constitución, unha comunidade autónoma que conta con institucións propias: o seu propio parlamento, goberno, corpos de seguridade, televisión e radio públicas, bandeira, etc. O Estatuto de Autonomía de Galicia, aprobado en 1981, recoñece o galego como lingua "propia" de Galicia e idioma cooficial da comunidade, que "todos teñen o dereito de coñecer e usar" e, ao mesmo tempo, responsabiliza os poderes públicos da normalización do galego en todos os ámbitos. A Lei de normalización lingüística, aprobada por unanimidade o 15 de xuño de 1983 no Parlamento de Galicia, garante e ordena os dereitos lingüísticos dos cidadáns, especialmente os referidos aos ámbitos da administración, a educación e os medios de comunicación.

No Estatuto de Autonomía do ano 1981 declárase o galego como lingua "propia" de Galicia e cooficial da comunidade, e outórgaselles ás institucións autónomas a plena competencia no proceso de normalización.

En virtude da Lei de normalización lingüística, a Administración local e a autonómica están obrigadas a escribir todos os seus documentos oficiais en galego; está establecida a presenza do galego en todo o sistema educativo e garántese a promoción lingüística nos países con comunidades galegas emigrantes e nas áreas limítrofes con Galicia nas que se fala o galego.

Dende a morte de Franco, a situación do galego, sobre todo no que fai referencia ao seu status legal e á súa promoción, mellorou notablemente[7]. Así e todo, estas melloras non conseguiron o que realmente importa, xa que aínda non se conseguiu un aumento no uso falado do idioma, e unha plena igualdade xurídica co idioma español.

[5] Información extraída da páxina web da Xunta de Galicia
http://www.xunta.es/linguagalega/datos_basicos_da_lingua_galega
[6] Información extraída da páxina web do Consello da Cultura Galega
http://www.consellodacultura.org/
[7] Proxecto LOIA do Consello da Cultura Galega:
//www.consellodacultura.org/arquivos/cdsg/loia/historia.php?idioma=2&id=76

Competencia lingüística en galego[8]

	Entenden	Falan	Len	Escriben
Censo 2001	99.16%	91.04%	68.65%	57.64%
Censo 1991	96.96%	91.39%	49.30%	34.85%

Particularidades do idioma galego

O galego está estreitamente emparentado co idioma portugués. Tamén ten relación con outras linguas románicas, como o español ou o francés. O galego emprega sete sons vocálicos e dezanove sons consonánticos[9]. O alfabeto galego contén 23 letras (*a, b, c, d, e, f, g, h, i, l, m, n, ñ, o, p, q, r, s, t, u, v, x, z*) e seis dígrafos (*ch, gu, ll, nh, qu, rr*). As letras *ç, j, k, w* e *y* empréganse só nos estranxeirismos. O til (´) emprégase para marcar a sílaba acentuada nas palabras polisílabas e tamén se usa como diacrítico para distinguir pares de vocábulos que se diferencian entre eles na pronuncia porque unha é tónica e a outra átona (*dá*, verbo *dar* / *da*, preposición *de* + artigo *a*), ou porque unha delas ten unha vogal media aberta e a outra ten a correspondente pechada (*vés*, verbo *vir* / *ves*, verbo *ver*). Na escrita, *é* e *ó* representan as vogais medias tanto abertas coma pechadas.

Con respecto á orde das palabras nas oracións ou enunciados en galego, o padrón principal empregado é suxeito, verbo, obxecto. Non obstante, o galego é un idioma bastante libre e é común o uso de elementos clíticos que alteran a estrutura básica. A voz pasiva, formada polo verbo auxiliar *ser* e mais o participio do verbo principal, non se usa normalmente en galego, excepto en rexistros formais (documentos legais, xornalísticos ou científicos). No seu lugar, empréganse outras construcións para expresar a idea de pasividade: invértese a orde habitual das palabras (*Ese libro lino eu cando era pequeno*, *Esa película rodárona na Coruña*), úsase a forma activa do verbo co pronome reflexivo de terceira persoa (*Esa película rodouse na Coruña*), e existe tamén unha construción impersoal na que se usa a forma activa do verbo en terceira persoa do singular sen suxeito explícito, pero acompañada do pronome se (*Véndese viño*).

As interrogativas totais fórmanse normalmente invertendo a orde de suxeito e verbo (*Veu Antón?*). Se queremos resaltala pode engadírselle unha partícula interrogativa final (*Veu Antón ou non?*). A negación exprésase habitualmente poñendo o adverbio non antes do verbo: *Carme non dixo nada interesante*. Como pode verse no exemplo, no galego a "dobre negación" é de regra.

O galego tende á elipse dos pronomes: é posible empregar o verbo conxugado sen necesidade de incluír o pronome persoal que xoga o papel de suxeito.

[8] Táboa extraída do Plan Xeral de Normalización da Lingua Galega. Datos comparativos do Censo do 1991 e dos provisionais do Censo do 2001. Fonte: Instituto Galego de Estatística

[9] http://www.consellodacultura.org/arquivos/cdsg/loia/gramatica.php?idioma=2&seccion=6

A ortografía en galego é máis transparente que en inglés, pero menos que en español ou italiano. Por exemplo, as vogais *e* e *o* poden pronunciarse de maneiras distintas nalgúns dialectos.

Os tres principais bloques dialectais son: (1) galego oriental, que inclúe os dialectos falados fóra da Galicia administrativa, dos que o máis importante hoxe é o galego de Asturias; (2) galego central, onde salientan a área norte ou mindoniense, e a área sur ou lucu-auriense; e (3) galego occidental, onde salientan a área fisterrá no norte e tudense e baixo limega no sur.

Os principais trazos dialectais son:

(1) trazos fonéticos: *gheada* (en lugar do fonema oclusivo velar sonoro /g/ existe un fonema fricativo ou aproximante, con realizacións xordas ou sonoras, en palabras como *gato* ou *pagar*), característica do galego occidental e boa parte do galego central; e *seseo* (presenza de /s/ nas mesmas posicións nas que en galego común hai un /θ/, en palabras como *cen* ou *cazar*), característico do galego occidental;

(2) trazos morfolóxicos: nos substantivos, terminación -án (<latín -ANU & -ANA: *irmán* <latín GERMANU, GERMANA) nos dialectos occidentais, fronte á terminación -ao (<latín -ANU) e -á (<latín -ANA) (*irmao*/*irmá*) dos dialectos dos bloques central e oriental; na formación do plural dos nomes rematados en -n, terminación -óns (<latín -ONES) nos dialectos do bloque occidental, fronte á terminación -ós dos dialectos do bloque central e -ois nos do bloque oriental; nos verbos, o sufixo de persoa -is para a segunda persoa do plural (*andais*) nos dialectos orientais, fronte ao sufixo -des do galego común (*andades*). Os dialectos orientais (especialmente o galego falado en Asturias) posúen outras moitas particularidades.

Avances recentes

Os medios de comunicación e as industrias culturais

Actualmente non hai ningún xornal integramente escrito en galego. En algúns o idioma non está totalmente ausente, aínda que si arrecunchado na información cultural e nas columnas de colaboradores. Na prensa non diaria, salienta un semanario de información xeral (*A Nosa Terra*), que sae regularmente desde hai máis de vinte anos, e un mensual de información e debate (*Tempos Novos*). Cunha difusión máis restrinxida, hai que sinalar os trimestrais *Grial* (de grande tradición), *Encrucillada*, *A Trabe de Ouro* e *Agália*.

No que atinxe á televisión, en 1985 creouse a Compañía de Radio-Televisión de Galicia, de titularidade autonómica, e a partir de aí comezou a emitir a televisión galega, basicamente en galego, cunha notable audiencia e algúns éxitos salientables.

Canto ás emisoras de radio, é sen dúbida a Radio Galega, de titularidade pública, a que amosa un maior compromiso co uso e promoción do idioma galego.

A edición de libros en galego incrementouse de forma moi notable no período que vimos describindo, pasando de 187 títulos editados en galego en 1980 a 1.826 no 2005. Con todo, hai que sinalar algúns problemas, como a abraiante presenza da edición institucional, a excesiva atomización das empresas editoras galegas e a perigosa dependencia do mercado escolar.

Canto á produción musical, destaca a renovada voga da música "de raíces", o que no noso caso quere dicir de inspiración (máis ou menos vaga) popular-tradicional, ou ben sinxelamente "celta"; e a apropiación desde a experiencia galega da música popular internacional contemporánea, isto é, o pop-rock.

Durante os últimos anos, creáronse unha serie de produtos e servizos coa intención de incorporar o galego á sociedade das TIC. Os sistemas operativos, os correctores ortográficos e as aplicacións telefónicas son algúns exemplos[10].

O cultivo da lingua

A Real Academia Galega[11] (RAG) é unha institución dedicada ao estudo da cultura galega, con especial atención ao seu idioma; elabora as normas gramaticais, ortográficas e léxicas e traballa na promoción do idioma galego.

O Consello da Cultura Galega[12] é unha institución estatutaria dedicada á promoción e conservación da cultura galega. A promoción do idioma galego é un dos seus obxectivos. O seu proxecto LOIA[13], levado a cabo pola Sección de Lingua e o Centro de Documentación Sociolingüística do Consello da Cultura Galega, ten como obxectivo presentar en forma condensada para un público moi amplo os elementos fundamentais para iniciarse no coñecemento do idioma galego, a súa singradura histórica, a súa produción cultural, a súa realidade social e as súas perspectivas de futuro. A maioría do material recollido neste capítulo foi extraído da páxina web do proxecto LOIA.

A linguaxe na educación

No Estatuto de autonomía de Galicia de 1981 declárase o galego lingua "propia" de Galicia e oficial, xunto co español. A introdución do idioma galego na educación tivo lugar no ano 1979. A elaboración da Lei de Normalización Lingüística ten como obxectivo lograr que os estudantes teñan as mesmas competencias orais e escritas en galego e castelán.

En Galicia, os nenos e nenas teñen o dereito de recibir a educación primaria na súa lingua materna, e as autoridades educativas están

[10] http://www.xunta.es/linguagalega/o_galego_nas_novas_tecnoloxias
[11] http://www.realacademiagalega.org/
[12] http://consellodacultura.org/
[13] http://www.consellodacultura.org/arquivos/cdsg/loia/index.php?idioma=2

obrigadas a proporcionar "os medios necesarios para promover o uso progresivo do galego na educación", establecendo como obxectivo mínimo que “ao remate dos ciclos en que o ensino do galego é obrigatorio os alumnos coñezan este, nos seus niveis oral e escrito, en igualdade co castelán”.

Desde os comezos da década dos oitenta emprendeuse un labor intensivo de reciclaxe lingüística en galego de profesores dos niveis primario e medio, por medio de cursos de lingua e literatura galegas, aos que ao longo da década asistiron unha boa parte dos profesores en exercicio. Desde os comezos da década dos noventa adóptanse previsións para a creación de equipos de normalización lingüística e a elaboración de plans de normalización nos centros de ensino, e establécense liñas de axuda para a promoción de actividades de fomento do galego.

En xeral, pode dicirse que a día de hoxe logrouse delimitar e centrar as diferentes iniciativas arredor dos que se consideran os dous obxectivos principais deste ámbito: converter o galego en lingua vehicular do sistema educativo e lograr que o alumnado obteña unha competencia lingüística plena nas dúas linguas oficiais (galego e castelán) ao remate do ensino obrigatorio. Con todo, e malia os incuestionables logros (desiguais dependendo do nivel educativo), queda aínda moito camiño por percorrer para que estes obxectivos se acaden realmente.

Aspectos internacionais

O galego é unha das linguas denominadas minoritarias, e foi recoñecida como tal na Carta Europea das Linguas Rexionais ou Minoritarias do Consello de Europa, que "ten como obxectivo protexer e fomentar as linguas rexionais e minoritarias de Europa". A importancia destas linguas está demostrada polo feito de que as falan un total de máis de corenta millóns de cidadáns da UE.

Como lingua minoritaria, o galego foi representado na Oficina Europea de Linguas Minoritarias, creada no ano 1982 por iniciativa do Parlamento Europeo. O obxectivo desta organización paneuropea non gobernamental vén sendo fomentar o respecto cara ás linguas minoritarias protexidas da UE, así como promover a diversidade lingüística.

Tendo en conta todas as linguas faladas en España, o español é a única que conta con status de idioma oficial na UE. Non obstante, en novembro de 2004, o goberno español entregou á UE a tradución da Constitución Europea nos idiomas do estado, que tamén son oficiais nos seus territorios respectivos: o galego, o catalán (chamado catalán ao empregado en Cataluña e nas Illas Baleares, e valenciano cando se emprega na Comunidade Valenciana) e o éuscaro.

En 2005, o Consello de Ministros admitiu a posibilidade de empregar outros idiomas oficiais que non sexan o español nas institucións europeas. Tras a sinatura de acordos administrativos con algunhas institucións da UE, que recoñecen un uso limitado do galego, o status do galego é actualmente o de lingua semioficial, unha lingua de comunicación cos cidadáns. Este status implica que os cidadáns poden dirixirse por escrito en galego a estas institucións (Comisión Europea, Parla-

mento Europeo, Consello, Defensor do Pobo Europeo e Comité de Rexións) e, á vez, teñen o dereito de ser respondidos na mesma lingua. Así mesmo, algunhas publicacións e documentos oficiais tradúcense ao galego.

A proxección internacional do galego é bastante limitada. No mundo empresarial a nivel internacional, o uso do galego é inexistente. De feito, o inglés converteuse na principal lingua de comunicación a nivel escrito e oral. Hoxe en día, dende o punto de vista dos clientes, algunhas grandes compañías internacionais empregan a lingua galega para tratar cos seus clientes galegos, como valor engadido aos seus produtos e supoñendo unha mellora dos seus servizos de atención ao cliente. Algunhas destas empresas son Microsoft, ou Telefónica.

A tecnoloxía lingüística pode afrontar este desafío dende unha perspectiva diferente, ao ofrecer servizos como os de tradución automática ou de recuperación de datos plurilingüe para textos escritos en idiomas estranxeiros, axudando así a reducir as desvantaxes persoais e económicas ás que se enfrontan as persoas cuxa lingua materna non é o inglés.

No que respecta á aprendizaxe de galego como lingua estranxeira, a situación mellora algo. A Comisión Europea está a desenvolver unha política activa en favor do plurilingüismo, co obxectivo de conservar e fomentar a diversidade lingüística en Europa, de impulsar a aprendizaxe de linguas (incluídas as linguas rexionais e minoritarias) e de empregar o plurilingüismo como estímulo para a competitividade. Neste contexto, o Programa de Aprendizaxe Permanente 2007-13 contén una selección de proxectos que fomentan a aprendizaxe de linguas. Entre eles, o centro de recursos plurilingüe en liña Lingu@net Europa Plus[14] proporciona apoio e recursos para o ensino e aprendizaxe en 20 idiomas europeos, incluído o galego. Ademais disto, os estados membros da UE tomaron a importante decisión de incluír o galego, así como o éuscaro e o catalán, na listaxe de linguas ofrecidas nos cursos de idiomas intensivos do programa Erasmus dende o ano académico 2010-2011[15]. Estes cursos financiados pola UE teñen como obxectivo preparar os futuros estudantes Erasmus para o seu período de estudo nas universidades galegas, onde o galego se emprega para a comunicación e como lingua académica.

Os servizos de normalización lingüística das tres universidades galegas, así como aqueles presentes nalgúns concellos, organizan periodicamente cursos de galego. Durante o verán, tamén existe a posibilidade de asistir aos “Cursos de Verán de Lingua e Cultura Galegas para Estranxeiros e para Españois de Fóra de Galicia”.

A Secretaría Xeral de Política Lingüística mantén convenios de colaboración con diferentes universidades de fóra de Galicia co ánimo de crear lectorados e cátedras de galego que potencien e difundan a lingua no eido internacional. Na actualidade existen corenta e sete centros de

[14] http://www.linguanet-worldwide.org/lnetww/gl/home.jsp
[15] http://ec.europa.eu/education/news/news1518_en.htm

estudos galegos en diferentes universidades de Europa, América e Oceanía.

Grazas ao desenvolvemento das novas tecnoloxías é posible aproximarse á aprendizaxe da lingua galega utilizando novas ferramentas dispoñibles na rede, como cursos interactivos en liña: *é-galego, A Palabra Herdada, Galingua.*

O galego na Internet

A presenza do Galego na Internet é bastante limitada (despois de todo, o galego ocupa o posto 160 segundo a clasificación do Ethnologue[16] de linguas dependendo da envergadura do idioma). Non obstante, existen algunhas iniciativas que intentan aumentar a presenza do galego na web. A Galipedia[17] (a Wikipedia en galego), con máis de 75.000 artigos está ao mesmo nivel que algúns idiomas oficiais da UE como o grego ou o letón. Outro exemplo é a iniciativa PuntoGal[18], que intenta obter un dominio na Internet para o idioma e a cultura galegos. A través deste dominio, a sociedade galega pasaría a ter máis visibilidade na rede e en todo o mundo. Google ou Facebook, entre outros, ofrecen unha versión en galego para as súas interfaces de navegación.

O goberno autonómico tamén levou a cabo algunhas iniciativas para dar apoio á creación de páxinas web en galego. Ademais, a páxina web *Mancomun*[19] ofrece unha serie de ferramentas informáticas gratuítas en galego creadas coa axuda da Xunta de Galicia. *Galinux*[20], por exemplo, é unha distribución GNU/Linux en galego deseñada para finalidades educativas.

A web tamén ofrece cada vez un maior número de xornais dixitais en galego (ou xornais españois que contan cunha ferramenta de tradución ao galego), así como algúns cursos en liña para a aprendizaxe do idioma.

[16] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size
[17] http://gl.wikipedia.org/wiki/Portada
[18] http://www.puntogal.org/
[19] http://www.mancomun.org
[20] http://www.galinux.org/

Apoio da tecnoloxía lingüística para o galego

As Tecnoloxías Lingüísticas

As Tecnoloxías Lingüísticas son tecnoloxías da información que están especializadas no tratamento da linguaxe humana. É por iso que estas tecnoloxías tamén se poden agrupar baixo o termo Tecnoloxías da Linguaxe Humana. A linguaxe humana prodúcese tanto en forma oral como escrita. Mentres que a fala é a forma máis antiga e natural de comunicación lingüística, a información máis complexa, así como a maior parte do coñecemento humano transmítense e están documentados en textos escritos. As tecnoloxías da fala e as tecnoloxías textuais procesan o idioma nestas dúas formas. Pero a linguaxe tamén ten aspectos comúns a ambas as dúas formas, tales como os dicionarios, a maioría da gramática, e o significado das oracións. Por tanto, moitas partes da tecnoloxía lingüística non poden ser agrupadas dentro das tecnoloxías da fala ou das tecnoloxías textuais. As tecnoloxías do coñecemento inclúen tecnoloxías que vinculan a linguaxe e o coñecemento. A Figura 1 representa o panorama da tecnoloxía lingüística. Na nosa comunicación, mesturamos a linguaxe con outras formas de comunicación e medios de información. Combinamos a fala con expresións xestuais e faciais. Os textos poden combinarse con fotografías e sons. As películas poden conter linguaxe falada e escrita. Por iso, as tecnoloxías textuais e da fala solápanse e interactúan con moitas outras tecnoloxías que facilitan o procesamento de comunicación multimodal e documentos multimedia.

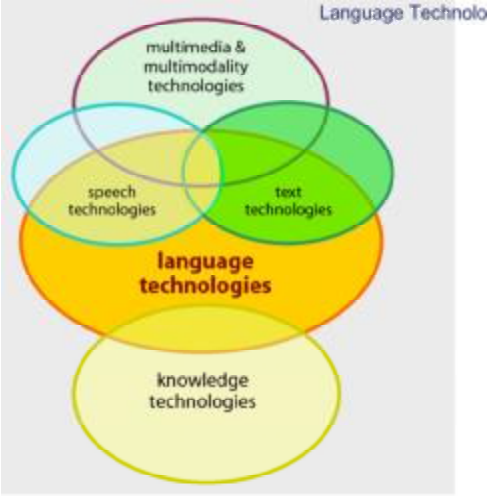


Figura 1: Panorama da tecnoloxía lingüística

As arquitecturas das aplicacións na tecnoloxías lingüísticas

As aplicacións informáticas habituais para o tratamento da linguaxe consisten en varios compoñentes que reflicten diversos aspectos da linguaxe e da tarefa que levan a cabo. A Figura 2 presenta unha arquitectura moi simplificada que pode atoparse nun sistema de procesamento de textos. Os tres primeiros módulos tratan a estrutura e o significado do texto de entrada:

- Procesamento previo: limpar os datos, eliminar formato, detectar a lingua do texto de entrada, etc.
- Análise gramatical: atopar o verbo e os seus obxectos, modificadores, etc.; detectar a estrutura da oración.
- Análise semántica: desambiguación (cal dos significados de *gato* é o axeitado no contexto dado?), resolución de anáforas e expresións referenciais como *ela*, *o coche*, etc.; representación do significado da oración de xeito que se poida facer unha lectura automática.

Figura 2: Unha arquitectura típica de aplicación no procesamento de textos

Os módulos de tarefas específicas levan a cabo diferentes operacións, tales como o resumo automático dun texto de entrada, consultas de bases de datos, e moitas outras. Abaixo, mostraremos as áreas principais de aplicación e destacaremos os seus módulos principais. Novamente, as arquitecturas das aplicacións aparecen moi simplificadas para mostrar a complexidade das aplicacións da tecnoloxía lingüística (TL) dun xeito comprensible para o público xeral.

Tras facer unha introdución sobre as áreas principais de aplicación, faremos un pequeno repaso da situación na investigación aplicada á tecnoloxía lingüística e á educación, concluíndo cun resumo dos pro-

gramas de investigación pasados e actuais. Ao final desta sección ofreceremos a opinión dos expertos sobre a situación das principais ferramentas e recursos da tecnoloxía lingüística no tocante a varios aspectos, como a súa dispoñibilidade, madurez ou calidade. Esta táboa proporciona unha boa visión de conxunto sobre a situación da TL para o galego.

As ferramentas e recursos máis importantes aparecen subliñados no texto e tamén se poden atopar na táboa ao final do capítulo. As seccións que tratan das principais áreas de aplicación tamén conteñen un resumo do mercado activo nos campos respectivos para o galego.

Principais áreas de aplicación

A corrección lingüística

Calquera persoa que empregue unha ferramenta de procesamento de textos, como Microsoft Word, atoparase coa opción de **corrección ortográfica**, que indica os erros de ortografía e propón as súas correccións. Corenta anos despois do primeiro programa de corrección ortográfica deseñado por Ralph Gorin, hoxe en día os correctores lingüísticos non só comparan a lista de palabras extraídas cun dicionario de palabras correctamente escritas, senón que estes son cada vez máis sofisticados. Ademais dos algoritmos que dependen do idioma para o tratamento da morfoloxía (por exemplo, a formación do plural), algúns son agora capaces de recoñecer erros relacionados coa sintaxe, tales como a omisión do verbo ou un verbo que non estea acorde co seu suxeito en persoa e número (por exemplo, “Ela**escriben* unha carta”). Non obstante, para outros tipos de erros comúns, os métodos anteriormente descritos non son suficientes. Por exemplo, botémoslle unha ollada a continuación ao primeiro verso dun poema escrito por Jerrold H. Zar (1992):

> *Eye have a spelling chequer,*
> *It came with my Pea Sea.*
> *It plane lee marks four my revue*
> *Miss Steaks I can knot sea.*

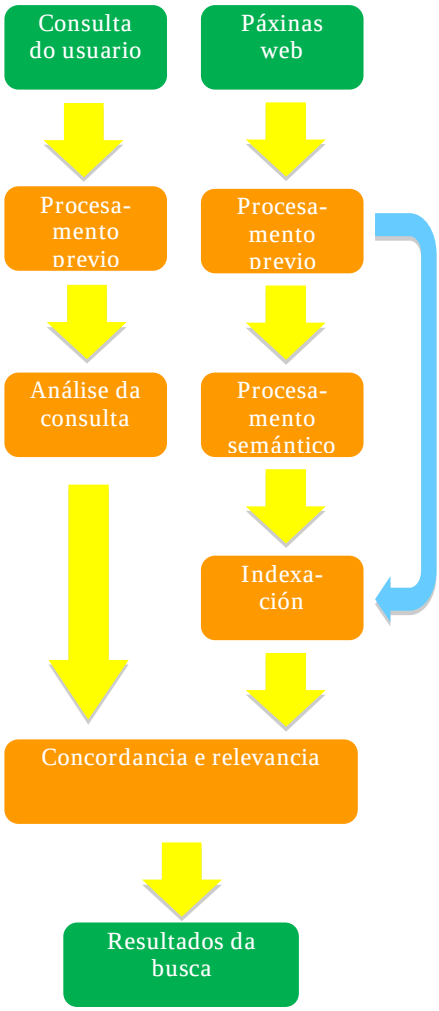
A meirande parte dos correctores ortográficos (incluído Microsoft Word) non atoparán erros neste poema, porque se fixan principalmente nas palabras de forma illada. Non obstante, para detectar os erros chamados homófonos (por exemplo, "*Eye*" no canto de "*I*"), o corrector lingüístico ten que considerar o contexto no que se atopa a palabra. No caso do galego, mesmo a corrección ortográfica require unha análise do contexto en moitos casos. Un caso típico sucede cando o erro ortográfico transforma unha palabra noutra que tamén existe. No seguinte exemplo, a primeira frase contén un erro frecuente (problemas con acentos ortográficos). A segunda frase é a versión corrixida da primeira.

*A casa do meu tío **e** a casa da miña avoa.*

*A casa do meu tío **é** a casa da miña avoa.*

Para corrixir automaticamente estes erros, non é suficiente facer unha comprobación de cada palabra no dicionario, xa que todas as palabras da primeira frase son correctas de forma illada. Isto ou ben require a

formulación de normas gramaticais específicas para cada lingua (é dicir, un alto grao de experiencia e traballo manual), ou o uso dun modelo lingüístico estatístico. Estes modelos calculan a probabilidade de que se dea unha palabra nun contexto específico (é dicir, as palabras anteriores e as seguintes). Por exemplo, "é a" é unha secuencia de palabras moito máis probable que "e a". Ao empregar unha grande cantidade de datos (correctos) lingüísticos (é dicir, un corpus), pódese obter automaticamente un modelo lingüístico estatístico.



Ata o de agora, este tipo de mecanismos foron principalmente desenvolvidos e avaliados con datos do idioma inglés. Non obstante, non sempre son aplicables a outras linguas, como por exemplo as linguas sintéticas ou as linguas con flexión sintáctica, como o galego. Para estes idiomas máis complexos, un corrector lingüístico avanzado de alta precisión requirirá o desenvolvemento de métodos máis sofisticados e que implican unha análise lingüística máis profunda.

O emprego dos correctores lingüísticos non está limitado ás ferramentas de procesamento de textos, senón que tamén se aplican no *software* de autoría. Xunto co aumento de produtos técnicos, durante as últimas décadas tamén se incrementou a cantidade de documentación técnica. Por temor a recibir reclamacións de clientes por danos e prexuízos como resultado dunhas malas instrucións, ou da súa mala comprensión, as compañías comezaron a centrarse cada vez máis na calidade da documentación técnica, coa intención, ao mesmo tempo, de chegar ao mercado internacional. Os avances no procesamento da linguaxe natural levaron ao desenvolvemento de *software* de autoría, que axuda ao redactor da documentación técnica a empregar vocabulario e estruturas sintácticas coherentes seguindo certas normas e restricións terminolóxicas (corporativas).

Só unhas poucas empresas e provedores de servizos lingüísticos ofrecen produtos nesta área para o galego. O software *Imaxin*[21] é un exemplo, xa que ofrece servizos en liña gratuítos de tradución e corrección gramatical. O software *OrtoGal*, do Seminario de Lingüística Informática (SLI)[22] da Universidade de Vigo, ofrece un servizo de corrección gramatical e ortográfica. Tamén existen extensións de software para OpenOffice, como *Golfiño*[23], creado por *Imaxin Software* e promovido pola Xunta de Galicia.

Ademais dos correctores ortográficos e do *software* de autoría, a corrección lingüística tamén é importante no eido da aprendizaxe de linguas asistida por ordenador, así coma nos motores de busca na web, onde se aplica para corrixir automaticamente as consultas realizadas, como as suxestións de Google "Quizais quixo dicir...".

A busca na Web

A busca na web, en intranets, ou en bibliotecas dixitais é probablemente a tecnoloxía lingüística máis empregada hoxe en día á vez que é a menos desenvolvida. O motor de busca Google, que se creou no ano

[21] http://www.imaxin.com/
[22] http://webs.uvigo.es/sli/index_en.html
[23] http://www.mancomun.org/descargarprogramas/detalledeproduto/nova/golfino/

1998, emprégase hoxe en día para o 80% das buscas que se realizan en todo o mundo. Ata o de agora, non houbo cambios significativos con respecto á primeira versión no referente á súa interface de busca nin ao xeito de presentar os resultados obtidos. Na versión actual, Google ofrece unha corrección gramatical para as palabras escritas incorrectamente, mesmo para a versión en galego, e, no ano 2009, incorporáronse posibilidades de busca semántica básica ao seu conxunto de algoritmos, que melloran a precisión da busca mediante unha análise do significado dos termos da consulta dentro dun contexto. O caso do éxito de Google demostra que, dispoñendo dunha grande cantidade de datos, e con técnicas eficientes para a indexación destes datos, un método baseado principalmente en estatísticas pode producir resultados satisfactorios.

Non obstante, cando hai unha solicitude de información máis complexa, é fundamental integrar un coñecemento lingüístico máis profundo. Nos laboratorios de investigación, os experimentos levados a cabo empregando tesouros e recursos lingüísticos ontolóxicos como *WordNet* amosaron melloras ao permitir a posibilidade de atopar unha páxina baseándose en sinónimos dos termos empregados na busca, ou mesmo en termos non relacionados directamente. Novamente, estes avances esixen uns recursos lingüísticos específicos. O Centro Ramón Piñeiro para a Investigación en Humanidades[24] elaborou un *WordNet* para o galego. O *WordNet* galego chámase GALWORDNET.

A próxima xeración de motores de busca terá que incluír unha tecnoloxía lingüística moito máis sofisticada. Se unha busca consiste nunha pregunta, ou noutro tipo de frase que non se trate dunha listaxe de palabras clave, é precisa unha análise desta frase, a nivel sintáctico e semántico, así como a dispoñibilidade dun índice que permita unha rápida recuperación de documentos relevantes, para poder obter respostas relevantes. Por exemplo, imaxinemos que o usuario introduce a busca: "Dáme unha lista de todas as compañías que foron absorbidas por outras compañías nos últimos cinco anos". Para obter un resultado satisfactorio, é necesario realizar unha análise sintáctica para analizar a estrutura gramatical da frase e determinar que o usuario está buscando compañías que foron absorbidas, e non compañías que absorberon outras compañías. Así mesmo, a expresión "últimos cinco anos" tamén ten que ser procesada para saber a que anos se refire.

Figura 3: A arquitectura dunha busca na web

Finalmente, a solicitude procesada contrástase cunha grande cantidade de datos non estruturados para atopar a información ou as informacións que o usuario está a solicitar. Este proceso coñécese habitualmente como recuperación da información, e implica a busca e a clasificación de documentos relevantes. Ademais, á hora de xerar un listado de compañías, tamén necesitamos extraer a información que fai referencia ao nome dunha compañía dentro dunha cadea de palabras concreta situada nun documento. Este tipo de información está dispoñible grazas aos chamados recoñecedores de nomes de entidades.

Unha tarefa aínda máis difícil é intentar facer coincidir unha busca con documentos escritos noutro idioma. Para a recuperación de informa-

[24] http://www.cirp.es

ción plurilingüe, temos que traducir automaticamente a consulta a todas as posibles linguas fonte e volver a pasar a información recuperada á lingua meta. A crecente porcentaxe de datos dispoñibles en formatos non textuais enfoca a demanda cara a servizos que permiten a recuperación de información multimedia (é dicir, a busca de información en imaxes, audio e vídeo). Para os arquivos de audio e vídeo, introdúcese un módulo de recoñecemento do discurso, que converte o contido do discurso en texto ou nunha representación fonética, e pode facer coincidir coa consulta realizada polo usuario.

Non nos consta que exista tecnoloxía lingüística en compañías dedicadas á busca plurilingüe e á recuperación de información, tanto na Internet como en sistemas de información internos, para o galego.

A interacción da fala

A tecnoloxía de interacción da fala é a base para a creación de interfaces que permiten ao usuario interactuar con máquinas empregando a lingua falada no canto de, por exemplo, un teclado, un rato ou unha pantalla gráfica. Hoxe en día, estas interfaces de usuario baseadas en voz empréganas habitualmente as empresas para proporcionar aos seus clientes, empregados, ou socios, servizos parcialmente ou totalmente automatizados por vía telefónica. Os sectores empresariais que dependen en gran medida das interfaces de usuario baseadas en voz son: a banca, a loxística, o transporte público e as telecomunicacións. Outros usos da tecnoloxía de interacción da fala son as interfaces de certos aparellos, como os sistemas de navegación integrados nos automóbiles, ou o emprego da lingua falada como alternativa ás modalidades de entrada e saída de información nas interfaces gráficas de usuario, como por exemplo os *smartphones*.

En esencia, a interacción da fala consta das seguintes catro tecnoloxías:

- O recoñecemento automático da fala (RAF) é o responsable de determinar que palabras foron emitidas tendo en conta a secuencia de sons pronunciados por un usuario.
- A análise sintáctica e a interpretación semántica ocúpanse de analizar a estrutura sintáctica da locución do usuario e da súa posterior interpretación en función da finalidade do sistema respectivo.
- A xestión do diálogo é necesaria para determinar, pola parte do sistema coa que interactúa o usuario, que acción debe levarse a cabo tendo en conta o *input* do usuario e a funcionalidade do sistema.
- A tecnoloxía de síntese de diálogo (texto a voz, TAV) emprégase para transformar as palabras que se emitiron nesa locución en sons que xerarán como resultado unha información de saída para o usuario.

Un dos maiores retos é que o sistema de RAF recoñeza as palabras pronunciadas por un usuario do xeito máis preciso posible. Isto require ou ben unha restrición da variedade de posibles pronuncias do usuario, para así contar cun conxunto limitado de palabras clave, ou a creación manual de modelos de linguaxe que cubran unha grande variedade de

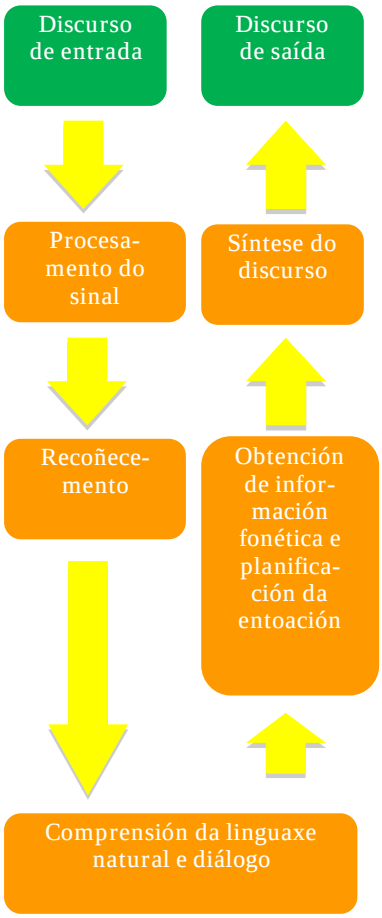


Figura 4: Unha arquitectura de diálogo simple baseado no discurso

pronuncias de usuarios da linguaxe natural. Mentres que o anterior dá como resultado un uso bastante ríxido e inflexible dunha interface de usuario baseada en voz, e posiblemente provoque unha mala aceptación por parte do usuario, a creación, optimización e mantemento de modelos de linguaxe chegan a aumentar os custos considerablemente. Non obstante, as interfaces de usuario baseadas en voz que empregan modelos de linguaxe e que inicialmente permiten que o usuario exprese con flexibilidade as súas intencións (como, por exemplo, nun saúdo como "En que podo axudarlle") amosan tanto un maior grao de automatización como unha maior aceptación por parte do usuario e, polo tanto, poden considerarse vantaxosas mesmo cun enfoque de diálogo dirixido menos flexible.

No que respecta á saída de información das interfaces de usuario baseadas en voz, as empresan a miúdo adoitan empregar fórmulas previamente gravadas por locutores profesionais (a ser posibles corporativos). Para as locucións estáticas, nas que a redacción non depende do contexto particular de uso, ou dos datos persoais do usuario determinado, isto dará lugar a unha experiencia de usuario satisfactoria. Non obstante, canto máis dinámico sexa o contido que a locución deba ter en conta, a experiencia de usuario sufrirá máis dunha mala prosodia como resultado da concatenación de arquivos de audio individuais. Pola contra, os sistemas actuais de texto a voz demostran ser superiores, a pesar de que se poden mellorar, no referente á naturalidade prosódica das locucións dinámicas.

Con respecto ao mercado das tecnoloxías de interacción da fala, a última década caracterizouse por unha forte estandarización das interfaces entre os diferentes compoñentes tecnolóxicos, así como polas normas para a creación de aparatos de software particulares para unha determinada aplicación. Tamén existiu unha forte consolidación de mercado nos últimos dez anos, especialmente nos campos do RAF e o TAV. Neste caso, os mercados nacionais dos países do G20 (é dicir, países economicamente fortes cunha poboación considerable) están dominados por menos de 5 reprodutores en todo o mundo, dos cales *Nuance* e *Loquendo* destacan en Europa, así como para o galego (*Loqueando*); non obstante, algunhas compañías locais máis pequenas están comezando a facer competencia, tales como Verbio[25], que é unha versión da Universitat Politècnica de Catalunya e que ten a súa propia tecnoloxía da fala, ou a galega 2Mares[26]

En relación á tecnoloxía e ás destrezas de coñecemento da xestión do diálogo, os mercados están fortemente dominados por reprodutores nacionais, que a miúdo se tratan de PEMEs.
A maioría das compañías no mercado do TAV español (algunhas ofrecen galego) son fundamentalmente desenvolvedoras de aplicacións. Os principais reprodutores no mercado español son: Indsys[27] (Sistemas de diálogo intelixente), Fonetic[28], Ydilo[29], NaturalVoz[30], e 2Mares.

[25] http://www.verbio.com/
[26] http://www.2mares.com/
[27] http://www.indisys.es/default.aspx
[28] http://www.fonetic.es/
[29] http://www.ydilo.com/esp/index.php
[30] http://www.naturalvox.com/

Pero máis aló do estado actual da tecnoloxía, xurdirán importantes cambios debido ao hábito, cada vez máis profuso, de empregar os teléfonos intelixentes (*smartphones*) como unha nova plataforma para xestionar as relacións cos clientes (ademais doutros canais como o teléfono, internet e o correo electrónico). Esta tendencia tamén afectará ao uso da tecnoloxía para a interacción da fala. Por unha banda, diminuirá a largo prazo a demanda de interfaces de usuario baseadas en telefonía. Por outra banda, o emprego da lingua falada cobrará unha importancia considerable nos *smartphones* como modalidade de entrada de información fácil de empregar polos usuarios. Esta tendencia tamén está reforzada pola mellora perceptible de precisión no recoñecemento independente da voz do locutor nos servizos de ditado de discurso que xa se ofrecen como servizos centralizados para os usuarios de *smartphones*. Dada esta "externalización" da tarefa de recoñecemento á infraestrutura das aplicacións, o emprego específico nas aplicacións de tecnoloxías lingüísticas básicas adquirirá probablemente máis importancia que na actualidade.

A tradución automática

A idea de empregar ordenadores dixitais para a tradución de linguas naturais foi desenvolvida no ano 1946 por A. D. Booth e, na década dos 50, foi acompañada dun financiamento importante para a investigación nesta área, que comezou novamente na década dos 80. Non obstante, a Tradución Automática (TA) aínda non consigue cumprir as grandes expectativas que se esperaban nos seus inicios.

No seu nivel básico, a TA simplemente substitúe palabras nunha lingua natural por palabras noutra. Isto pode resultar útil en disciplinas que posúen unha linguaxe moi restrinxida e con fórmulas establecidas, tales como os informes meteorolóxicos. Non obstante, para que exista unha boa tradución dos textos menos estandarizados, as unidades textuais máis grandes (frases, oracións, ou mesmo pasaxes completas) teñen que coincidir coas súas unidades homólogas máis próximas na lingua meta. Neste caso, a principal dificultade reside no feito de que a linguaxe humana é ambigua, e presenta retos a varios niveis, como por exemplo a desambiguación do sentido da palabra a nivel léxico ("Gato" pode significar animal ou aparello mecánico), ou a vinculación de frases preposicionais a nivel sintáctico como en:

> O policía observou ao home co telescopio.
> [*The policeman observed the man with the telescope.*]

> O policía observou ao home co revólver.
> [*The policeman observed the man with the revolver.*]

Unha das maneiras para abordar esta tarefa está baseada en regras lingüísticas. Para traducións entre linguas moi relacionadas, unha tradución directa pode resultar viable en casos como o exemplo presentado anteriormente. Non obstante, a miúdo os sistemas baseados en regras (ou baseados no coñecemento) analizan o texto de entrada e crean

unha representación simbólica intermediaria, a partir da cal se crea o texto na lingua meta. O éxito destes métodos depende en boa medida da dispoñibilidade de lexicóns extensos que contan con información morfolóxica, sintáctica e semántica, así como de grandes conxuntos de regras gramaticais coidadosamente deseñadas por lingüistas especializados.

Comezando a finais da década de 1980, e nunha situación na que a informática ía avanzando e se volvía menos cara, empezou a mostrarse cada máis interese nos modelos estatísticos de TA. Os parámetros destes modelos estatísticos derívanse da análise de corpus textuais bilingües, como o corpus paralelo Europarl, que contén as actas do Parlamento Europeo en 11 linguas europeas. Cunha cantidade de datos suficiente, a TA estatística funciona bastante ben á hora de obter un significado aproximado dun texto nunha lingua estranxeira. Non obstante, a diferenza dos sistemas baseados no coñecemento, a TA estatística (ou baseada en bases de datos) a miúdo obtén un resultado gramaticalmente incorrecto. Por outra banda, ademais de que grazas a ela requírese menos esforzo humano para realizar unha escritura gramaticalmente correcta, a TA baseada en datos tamén inclúe particularidades da linguaxe que non se inclúen nos sistemas baseados no coñecemento, como por exemplo as expresións idiomáticas.

Xa que os puntos fortes e febles da TA baseada no coñecemento e da TA baseada en datos son complementarios, hoxe en día os investigadores buscan enfoques híbridos que combinen as metodoloxías de ambas as dúas. Isto pode levarse a cabo de diferentes maneiras. Unha delas é empregar tanto os sistemas baseados en coñecemento como os sistemas baseados en datos e facer que un módulo de selección elixa o mellor resultado para cada oración. Non obstante, no caso das oracións máis longas, non existe un resultado perfecto. Unha solución mellor sería combinar as mellores partes de cada oración extraídas de múltiples resultados, o cal pode resultar bastante complexo, xa que as partes correspondentes obtidas de múltiples alternativas non sempre son obvias e necesitan ser aliñadas.

O desenvolvedor de software líder a nivel internacional Lucy Software ten unha importante filial en España, Lucy Iberica[31], anteriormente coñecida como Translendium. Lucy Ibérica encárgase do desenvolvemento de pares de linguas que inclúen o español, así como de todos os pares de linguas que implican calquera outra lingua ibérica (catalán, portugués, galego e éuscaro). O sistema de Lucy está baseado en regras gramaticais. A Xunta de Galicia[32] ofrece un servizo de tradución na Internet que emprega a tecnoloxía de Lucy Iberica. A pesar das importantes investigacións levadas a cabo sobre os sistemas baseados en datos e os sistemas híbridos a nivel nacional e internacional, ata o de agora esta tecnoloxía tivo menos éxito no eido comercial que no eido da investigación.

Apertium é unha plataforma de tradución automática libre que proporciona un motor de tradución automática independentemente da lingua

[31] http://www.lucysoftware.com/
[32] http://www.xunta.es/tradutor/

da que se trate. Esta plataforma foi deseñada polo grupo Transducens da Universitat d'Alacant, e posteriormente desenvolvida no marco do proxecto de financiamento nacional Opentrad. Entre os actuais sistemas de TA que empregan a tecnoloxía Apertium atopamos interNOSTRUM (español-catalán), desenvolvido por Transducens, o Traductor Universia (español-portugués), Matxin (éuscaro-español), desenvolvido polo grupo IXA[33] da Euskal Herriko Unibertsitatea, e Imaxin Software (galego-español). É posible empregar a tecnoloxía Apertium para construír sistemas de TA para unha variedade de pares de linguas (existen máis de 20 ata a data); con ese fin, Apertium emprega formatos estándar baseados en XML para codificar os datos lingüísticos necesarios (facéndoo a man ou convertendo os datos existentes) que se compilan, empregando as ferramentas proporcionadas, nos formatos de alta velocidade que emprega o motor.

En tanto exista unha boa adaptación no referente á terminoloxía específica do usuario e unha integración do fluxo de traballo, hai un amplo consenso á hora de establecer que o uso da TA pode aumentar a produtividade de maneira significativa. Así e todo, considérase que a calidade dos sistemas de TA aínda ten un grande potencial de mellora. Entre os retos de mellora que se presentan, cabe citar a adaptabilidade dos recursos lingüísticos a unha disciplina específica ou área do usuario, e a integración dentro de fluxos de traballo existentes con bases terminolóxicas e memorias de tradución. Ademais disto, aínda faltan moitos pares de linguas.

A tecnoloxía lingüística

A creación de aplicacións de tecnoloxía lingüística implica unha serie de subtarefas que non sempre se desenvolven no plano da interacción co usuario, senón que proporcionan importantes prestacións de servizo "baixo a carcasa" do sistema. Polo tanto, estas aplicacións constitúen asuntos importantes de investigación, e chegaron a converterse en subdisciplinas individuais no ámbito académico dentro da Lingüística Informática.

A resposta a preguntas converteuse nunha área de investigación, para a cal se construíron corpus anotados e se empezaron a levar a cabo competicións científicas. A idea é pasar da busca baseada en palabras clave (para a cal o motor responde cunha colección enteira de documentos potencialmente relevantes) a unha situación na que o usuario realiza unha pregunta concreta e o sistema proporciona unha soa resposta: "A que idade puxo Neil Armstrong pé na lúa?" - "38". Mentres que isto está obviamente relacionado coa área principal da busca na web, hoxe en día a resposta a preguntas é principalmente un termo xenérico para as preguntas de investigación tales como que *tipos* de preguntas deberían distinguirse e como deberían tratarse, como poden analizarse e compararse un conxunto de documentos que potencialmente conteñen a resposta (ofrecen respostas contraditorias?), e como pode extraerse de xeito fiable unha información específica -a resposta- dun documento, sen ignorar excesivamente o contexto.

[33] http://ixa.si.ehu.es/Ixa

Isto está, á vez, relacionado coa tarefa de extracción de información (EI), unha área que se volveu moi popular e influente na época do "cambio estatístico" na Lingüística Computacional a principios da década dos 90. A EI ten como obxectivo identificar extractos de información específicos dentro dun tipo específico de documentos; un exemplo disto podería ser o descubrimento dos axentes clave na absorción de empresas tal como informan os artigos xornalísticos. Outra hipótese na que se traballou é os informes sobre atentados terroristas, onde o problema reside en estruturar o texto seguindo un modelo que especifique o autor do crime, o obxectivo, a hora e localización do atentado, e os resultados do mesmo. Cubrir modelos sobre dominios específicos é a principal característica da EI, que por esta razón é outro exemplo dunha tecnoloxía "non visible" que constitúe unha área de investigación ben delimitada pero que, a efectos prácticos, necesita estar integrada dentro dun contorno de aplicacións axeitado.

Existen dúas áreas "dubidosas" que ás veces desempeñan o papel de aplicacións independentes, e outras veces a de compoñentes de apoio "baixo a carcasa" do sistema, que son o resumo do texto e a produción do texto. O resumo, obviamente, refírese á tarefa de converter un texto longo en curto, e ofrécese, por exemplo, como función dentro de MS Word. Funciona en gran medida sobre unha base estatística, identificando primeiro as palabras "importantes" nun texto (é dicir, por exemplo, palabras que son moi frecuentes neste texto, pero notablemente menos frecuentes no uso xeral da lingua) e, a continuación, determinando as oracións que conteñen moitas palabras importantes. Logo, estas oracións márcanse no documento, ou extráense del, e cóllense para compoñer o resumo. Neste caso, que é sen dúbida o máis común, o resumo é igual á extracción da oración: o texto redúcese a un subconxunto das súas oracións. Todos os resumidores comerciais fan uso desta idea. Un enfoque alternativo, ao que se dedicaron algunhas investigacións, é o de sintetizar *novas* oracións, é dicir, construír un resumo de oracións que non aparecen necesariamente nesa forma no texto orixinal. Isto require unha certa comprensión máis profunda do texto e, polo tanto, resulta moito menos robusto. En definitiva, un xerador de texto na maioría dos casos non é unha aplicación independente senón que está integrado nun contorno de software máis grande, como no sistema de información clínica onde se recollen, almacenan e procesan os datos dos pacientes, e a xeración de informes é só unha de entre moitas funcións.

No caso do galego, a situación nestas áreas de investigación está moito menos desenvolvida que no caso do inglés, onde, desde a década de 1990, a resposta a preguntas, a extracción de información e a síntese foron obxecto de numerosas competicións, principalmente as organizadas por DARPA/NIST en Estados Unidos. Estas melloraron notablemente a tecnoloxía punta, pero sempre se fixo fincapé no idioma inglés; algunhas competicións engadiron pistas plurilingües, pero nunca se empregou a lingua galega. Como consecuencia, apenas existen corpus anotados ou outros recursos para estas tarefas. Os sistemas de resumo, cando empregan métodos puramente estatísticos, adoitan ser en boa medida independentes da lingua, e polo tanto están dispoñibles algúns prototipos de investigación. Para a xeración de texto, os compoñentes reutilizables tradicionalmente estaban limitados aos módulos de desenvolvemento da superficie (as "gramáticas de xeración"); unha

vez máis, atopamos que a maioría do software dispoñible é para o inglés.

Ademais dos sistemas experimentais desenvolvidos polos grupos de investigación, non existen PEMEs que ofrezan este tipo de servizos. Desde o ano 2000 ata hoxe, o goberno español vén apoiando, dentro do Plan Nacional de Investigación e Tecnoloxía, varios proxectos na área das tecnoloxías da fala plurilingües: TEHAM, AVIVAVOZ, e BUCEADOR. O seu principal propósito era mellorar a calidade do recoñecemento da fala, da tradución da fala e da síntese de texto a voz en todas as linguas oficiais de España: éuscaro, galego, catalán e español.

A tecnoloxía lingüística na educación

A tecnoloxía lingüística é un eido sumamente interdisciplinario, no que participan expertos lingüistas, informáticos, matemáticos, filósofos, psicolingüistas e neurólogos, entre outros. Como consecuencia, a actual formación básica dun lingüista informático en España debe levarse a cabo no marco dun título en Filoloxía ou Lingüística, que inclúa á Lingüística Informática como materia troncal, ou nas facultades de Informática. Entre as universidades que ofrecen a primeira opción están: A Universitat de Barcelona, a Universitat Pompeu Fabra, a Universitat Oberta de Catalunya e a Universidade de Vigo. Por outra banda, as principais facultades de informática que ofrecen como materia a Lingüística Informática son: A Universidade Politécnica de Madrid, a Universidade Carlos III, a Universidade Autónoma de Madrid, a Universitat d'Alacant, a Universidade Nacional de Educación a Distancia, e a Euskal Herriko Unibertsitatea. Outros casos, como a Universidade Complutense, combina ambos os dous.

Os cursos de posgrao ofrecen unha formación profesional máis específica. Existen varios programas de doutoramento que ofrecen másteres ou materias relacionadas co procesamento da fala e da linguaxe. Algunhas universidades, como a Universidade Politécnica de Cataluña, tamén participan no Máster Europeo en Linguaxe e Fala, avalado pola ELSNET (Rede de Excelencia Europea en Tecnoloxías Lingüísticas). Existen varios másteres en diversas universidades, a nivel nacional ou a nivel europeo; por exemplo, a Universitat Autònoma de Barcelona ofrece o Máster Internacional en Procesamento de Linguaxe Natural e Tecnoloxías Lingüísticas, en colaboración con universidades estranxeiras. Noutros másteres ou cursos de doutoramento tamén se imparten módulos sobre tecnoloxías lingüísticas, especialmente en Tradución (por exemplo, nas universidades Autónoma de Barcelona, Alacante, Castellón, Politécnica de Valencia e Granada).

Existen máis de 30 grupos de investigación en España repartidos nas diferentes universidades, que traballan no recoñecemento da fala, o procesamento de linguaxe natural, a tradución de texto a texto, e a síntese da fala. A Sociedade Española para o Procesamento da Linguaxe Natural (SEPLN) é unha organización sen ánimo de lucro con máis de 300 membros, tanto do ámbito académico como da industria, que foi creada en 1984 co propósito de promover e difundir as actividades relacionadas coa docencia, a investigación e o desenvolvemento do PLN,

tanto a nivel nacional como internacional. A SEPLN organiza seminarios, simposios e conferencias, e promove a colaboración con institucións nacionais e internacionais.

Tamén organiza unha conferencia anual, á que cada ano asisten máis investigadores que traballan sobre o PLN, tanto de España como do estranxeiro. A asociación tamén edita unha revista periódica e mantén un servidor web con información sobre cuestións relacionadas co procesamento de linguaxe natural, así como un foro aberto para os membros.

A Rede Temática en Tecnoloxías da Fala (RTTF) española[34] é un foro común onde os investigadores (actualmente máis de 250) en tecnoloxías da fala xuntan esforzos e comparten experiencias para:

- Promover a investigación nas tecnoloxías da fala para atraer os investigadores novos neste campo mediante a formación, os intercambios de estudantes, as bolsas e os premios.
- Atraer inversións para a investigación empresarial mediante o desenvolvemento de aplicacións innovadoras que ofrezan novas oportunidades de negocio.
- Progresar no establecemento de alianzas e na integración dos membros da rede para manter o liderado de España na investigación do español, e impulsar tamén os idiomas cooficiais como o catalán, o éuscaro e o galego.

Dende o ano 2000, a RTTF vén promovendo as "Xornadas en Tecnoloxía da Fala" bianuais. Este seminario ten como obxectivo converterse nun punto de encontro para presentar e discutir os resultados da investigación sobre a fala e as tecnoloxías lingüísticas nos idiomas ibéricos. Tamén busca promover a colaboración entre a industria e as universidades. Lévanse a cabo unha ampla variedade de actividades, como presentacións de relatorios técnicos, conferencias maxistrais, presentación de informes dos proxectos e actividades dos laboratorios, demostracións, e presentacións recentes de teses de doutoramento.

Programas de tecnoloxía lingüística

Os Ministerios de Educación e Ciencia e Innovación españois apoian a investigación no eido das tecnoloxías da información a través de programas nacionais de investigación. Estes programas impulsaron numerosos proxectos de investigación e colaboración con centros de investigación e empresas internacionais. A base do desenvolvemento da tecnoloxía e das aplicacións comerciais para o procesamento automatizado da lingua española creouse en parte como resultado destes proxectos.

O Centro para o Desenvolvemento da Tecnoloxía Industrial (CDTI) é unha entidade pública española dependente do Ministerio de Ciencia e Innovación, cuxo obxectivo é axudar ás empresas españolas a mellorar o seu nivel tecnolóxico. O CDTI avalía e financia proxectos de I+D a través de programas como CENIT e AVANZA.

[34] http://www.rthabla.es

O programa CENIT (Consorcio Estratéxico Nacional para a Investigación Tecnolóxica) busca estimular a cooperación en I+D entre o sector privado, as universidades, as organizacións e centros públicos de investigación, os parques de ciencia e tecnoloxía e os centros tecnolóxicos, así como impulsar a cooperación público-privada en I+D. Os proxectos do CENIT duran polo menos catro anos e teñen un presuposto mínimo de 5 millóns de euros ao ano, durante os cales recibirán un financiamento mínimo do 50% do sector privado. Polo menos o 50% do financiamento público destinarase aos centros públicos de investigación ou aos centros tecnolóxicos. A tecnoloxía da información e da comunicación é unha das áreas de prioridade do programa. Os proxectos levados a cabo nesta área inclúen ás veces investigación sobre tecnoloxías lingüísticas.

O obxectivo do plan AVANZ@ é achegar a sociedade da información aos cidadáns ordinarios, así como aos sectores público e privado. A promoción do uso das tecnoloxías TIC terá importantes repercusións en todos os sectores en xeral dentro de España e, polo tanto, no seu estado de innovación. Os obxectivos do plan inclúen incrementar a porcentaxe de empresas que utilizan o comercio electrónico, promover o uso da factura electrónica, ampliar o sector público electrónico mediante a posta en funcionamento dunha tarxeta de identidade electrónica e de rexistro electrónico, alcanzar un índice dun ordenador con acceso á Internet por cada dous alumnos nas escolas, e duplicar o número de fogares con acceso á Internet. Unha das súas prioridades é facilitar o uso de novas tecnoloxías aos anciáns e ás persoas minusválidas como medio ideal para a integración social, evitar a marxinación e mellorar a súa calidade de vida. As ferramentas da tecnoloxía lingüística fáciles de usar ofrecen a principal solución para satisfacer este obxectivo, por exemplo, proporcionando unha síntese da fala para as persoas invidentes.

A Xunta de Galicia apoia a investigación a través do Plan Galego de Investigación, Desenvolvemento e Innovación Tecnolóxica (PGIDIT). A tecnoloxía lingüística non é unha área de prioridade pero, ao longo dos anos, os grupos de investigación das universidades e algunhas empresas conseguiron bolsas para levar a cabo investigacións e desenvolvementos na TL.

Dispoñibilidade de ferramentas e recursos para o idioma galego

A seguinte táboa proporciona unha visión xeral da situación actual do apoio das tecnoloxías lingüísticas para o idioma galego. A valoración dos recursos e ferramentas existentes está baseada en estudos estimados levados a cabo por expertos, que empregaron os seguintes criterios (comprendidos entre 0 e 6).

1. **Cantidade:** Existe unha ferramenta e/ou recurso para a lingua en cuestión? Cantas máis ferramentas e/ou recursos existan, maior será a valoración.
    - 0: ausencia total de ferramentas e/ou recursos
    - 6: moitas ferramentas e/ou recursos, grande variedade
2. **Dispoñibilidade**: As ferramentas e/ou recursos son accesibles (é dicir, de código aberto e de uso gratuíto en calquera plata-

forma) ou só están dispoñibles por un alto prezo ou baixo condicións moi restrinxidas?

- 0: practicamente todas as ferramentas e/ou recursos están só dispoñibles pagando un prezo alto.
- 6: unha grande cantidade de ferramentas e/ou recursos é gratuíta, dispoñible baixo licencias de código aberto ou licencias Creative Commons, que permiten a súa reutilización e readaptación.

3. **Calidade**: En que medida as mellores ferramentas, aplicacións, e recursos dispoñibles satisfán os criterios respectivos de rendemento das ferramentas e dos indicadores de calidade dos recursos? Son estas ferramentas e recursos actuais e mantéñense actualizadas?
   - 0: ferramenta e/ou recurso de pouca importancia
   - 6: ferramenta de alta calidade, anotacións de calidade humana nun recurso.
4. **Cobertura**: En que grao as mellores ferramentas cumpren os criterios respectivos de cobertura (estilos, xéneros, tipo de texto, fenómenos lingüísticos, tipos de información de entrada e de saída, número de linguas coas que traballa un sistema de TA, etc.)? En que medida representan os recursos a lingua meta ou as sublinguas?
   - 0: recurso ou ferramenta especial, casos concretos, moi pouca cobertura, só para empregarse en casos moi específicos e non xerais.
   - 6: recurso cunha ampla cobertura, ferramenta moi robusta, amplamente aplicable, traballa con moitas linguas.
5. **Madurez**: Pode considerarse a ferramenta e/ou recurso maduro, estable, listo para o mercado? Están as mellores ferramentas e/ou recursos dispoñibles listos para usar ou teñen que adaptarse? É o rendemento deste tipo de tecnoloxía axeitada e lista para o seu uso en produción ou é só un prototipo que non se pode empregar para sistemas de produción? Un indicador pode ser se os recursos e/ou ferramentas son aceptadas pola comunidade e empregados con éxito en sistemas de TL.
   - 0: prototipo preliminar, sistema en fase de desenvolvemento, en etapa de probas preliminares, exemplo de recurso.
   - 6: compoñente inmediatamente integrable e aplicable.
6. **Sostibilidade**: En que medida pode manterse ou integrarse a ferramenta e/ou recurso nos sistemas informáticos actuais? Cumpre a ferramenta e/ou recurso un certo nivel de sostibilidade no que respecta á documentación e manuais, explicación para os casos de uso, interfaces gráficas de usuario, etc.? Emprega ou usa contornos de programación (como Java EE) estándar ou de acordo coas boas prácticas? Existen normas ou cuasi-normas para a investigación e a industria? Se existen, cumpre a ferramenta e/ou recurso con elas (formatos de datos, etc.)?
   - 0: formatos de datos totalmente restrinxidos, cun fin determinado, e datos API.
   - 6: perfecto cumprimento coas normas, totalmente documentado.

7. **Adaptabilidade**: En que medida poden adaptarse ou estenderse as mellores ferramentas e recursos a novas tarefas, dominios, xéneros, tipos de texto, casos de uso, etc.?
    - 0: practicamente imposible adaptar a ferramenta e/ou recurso a outra tarefa, mesmo imposible con grandes cantidades de recursos ou de persoas/mes dispoñibles.
    - 6: nivel moi alto de adaptabilidade; adaptación moi sinxela e eficientemente posible.

Táboa de ferramentas e recursos

	Cantidade	Dispoñibilidade	Calidade	Cobertura	Madurez	Sostibilidade	Adaptabilidade
Tecnoloxía lingüística (ferramentas, tecnoloxías, aplicacións)							
Tokenización, Morfoloxía (tokenización, etiquetado gramatical, análise/xeración morfolóxica)	4	5	4	5	4	4	4
Análise (análise sintáctica superficial ou profunda)	4	5	5	4	3	4	4
Semántica da oración (desambiguación do sentido das palabras, estrutura argumental, roles semánticos)	2	1	3	2	2	1	2
Semántica textual (resolución de correferencias, contexto, pragmática, inferencias)	1	1	3	2	2	2	1
Procesamento de discurso avanzado (estrutura textual, coherencia, estrutura retórica/TER, argumentación, patróns textuais, tipos textuais, etc.)							
Recuperación da Información (indexación de textos, RI multimedia, RI en varios idiomas)	2	1	2	2	1	2	1
Extracción de Información (recoñecemento de entidades, extracción de eventos e relacións, recoñecemento de opinións e sentimentos, minería/análitica de textos)	3	1	3	1	2	1	1
Xeración da linguaxe (xeración de oracións, informes e textos)							
Resumo, resposta a preguntas, tecnoloxías avanzadas de acceso á información	2	1	1	2	1	1	1
A tradución automática	3	5	4	5	5	4	4
Recoñecemento da fala	3	2	5	5	5	5	5
Síntese da fala	4	3	5	5	5	5	4
Xestión do diálogo (capacidades de diálogo e modelado orientado ao usuario)	1	0	1	1	0	0	0
Recursos lingüísticos (recursos, datos, bases de coñecemento)							
Corpus de referencia	5	4	5	5	5	5	4
Corpus sintáctico (corpus etiquetados sintacticamente, corpus de dependencias)	1	1	2	2	2	2	1
Corpus semánticos	1	1	1	1	1	1	1
Corpus de diálogo							
Corpus paralelos, memorias de tradución	3	5	5	5	5	5	5
Corpus orais (datos da fala sen procesar, datos da fala etiquetados e anotados, datos do diálogo da fala)	3	4	4	5	4	4	5
Datos multimedia e multimodais (datos de texto combinados con audio/vídeo)	2	1	4	2	3	3	3
Modelos de linguaxe	2	1	5	4	4	4	4
Lexicóns, terminoloxías	5	4	5	4	5	4	4
Gramáticas	2	2	2	2	2	2	2
Tesouros, WordNets	4	5	4	4	3	3	3
Recursos ontolóxicos para o coñecemento mundial (por exemplo, modelos superiores, datos vinculados)	2	1	1	1	1	1	1

Conclusións

A situación do galego no que respecta ao apoio das tecnoloxías lingüísticas dá lugar a un optimismo cauto. Grazas ao apoio dalgúns proxectos de investigación levados a cabo anteriormente, hoxe en día existen en España un panorama de investigación e unha industria da tecnoloxía lingüística emerxentes, que desenvolven produtos e servizos para o idioma galego. A industria está conformada por PEMEs, a maioría das cales xurdiron a raíz dun proxecto ou un grupo de investigación.

No caso do galego, existen varias tecnoloxías e recursos, pero moitas menos que para o inglés. Así e todo, mesmo no caso do inglés e doutras linguas principais, o apoio da tecnoloxía lingüística hoxe en día aínda está lonxe de ofrecer o soporte necesario que precisa unha verdadeira sociedade plurilingüe do coñecemento.

Nesta serie de libros brancos, levouse a cabo un primeiro esforzo de avaliar a situación xeral de moitas linguas europeas no tocante ao apoio da tecnoloxía lingüística dun xeito que permitise a comparación e identificación exhaustiva de carencias e necesidades.

Para o galego, os principais resultados en relación ás tecnoloxías e recursos son os seguintes:

- O procesamento da fala actualmente parece estar máis desenvolvido có procesamento do texto escrito. As tecnoloxías de acceso á información avanzada aínda se atopan nos seus comezos e, para o caso do galego en particular, apenas existen.
- Canto máis coñecemento lingüístico e semántico inclúe unha ferramenta, máis carencias aparecen (véxase, por exemplo, a recuperación de información contra a semántica textual) e son necesarios máis esforzos para dar apoio ao procesamento lingüístico exhaustivo.
- A investigación tivo éxito á hora de deseñar software de alta calidade, pero moitos dos recursos carecen de estandarización, é dicir, mesmo se existen, non sempre teñen a sostibilidade necesaria; precísanse iniciativas e programas concertados para estandarizar os datos e os formatos de intercambio.
- Para o galego existe un amplo corpus textual de referencia (cunha mestura equilibrada de diversos xéneros), así como outros corpus especializados, mais estes non son facilmente accesibles nin baratos.
- Aínda que existen algúns corpus específicos de alta calidade, non se dispón dun corpus extenso anotado sintacticamente.
- Existen moi poucos corpus anotados con información sintáctica, semántica ou discursiva e, polo tanto, a situación empeora cando se precisa información lingüística e semántica máis profunda.
- O procesamento da fala está actualmente máis desenvolvida ca o PLN para os textos escritos.

- Existen corpus paralelos entre o galego e o español, e estes foron empregados para o desenvolvemento de sistemas de tradución automática. Non obstante, hai unha ausencia de corpus paralelos entre o galego e outras linguas.
- Os datos multimedia constitúen tamén unha enorme carencia.

É por isto que queda claro que é preciso centrar máis esforzos na creación de recursos para o galego, así como na investigación, innovación e desenvolvemento. A necesidade de grandes cantidades de datos e a enorme complexidade dos sistemas informáticos que incorporan tecnoloxías lingüísticas tamén obrigan a desenvolver novas infraestruturas para o intercambio e a cooperación.

META-NET

META-NET é unha Rede de Excelencia financiada pola Unión Europea. Actualmente está formada por 44 membros, que representan a 31 países europeos, os cales se enumeran a continuación. META-NET está construíndo a Alianza Tecnolóxica Plurilingüe de Europa (META), unha crecente comunidade de organizacións e profesionais das tecnoloxías lingüísticas en Europa.

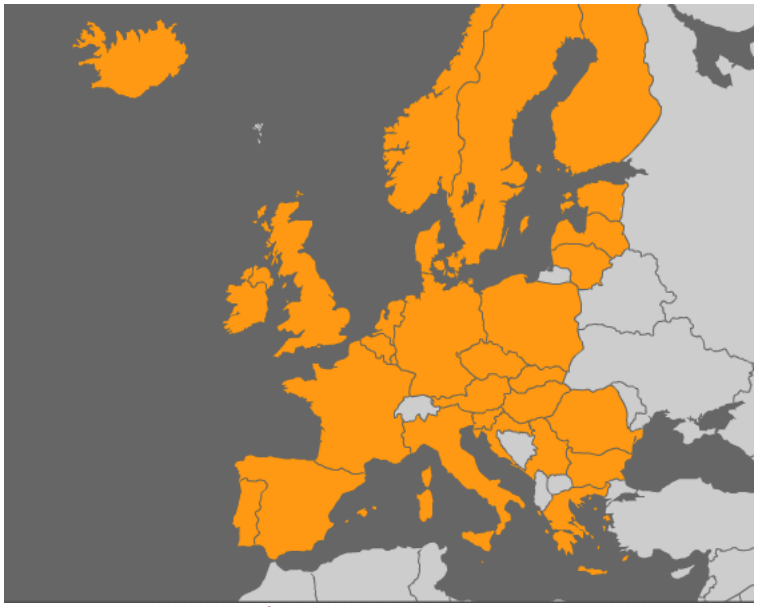


Figura 5: Países representados en META-NET

META-NET coopera cunha ducia de iniciativas importantes como CLARIN, que está a axudar ás ciencias sociais a establecer o eido das Humanidades Dixitais en Europa. META-NET dedícase a fomentar os fundamentos tecnolóxicos para establecer e manter unha verdadeira sociedade europea da información plurilingüe que

- permita a comunicación e cooperación entre as linguas,
- asegure aos usuarios de calquera idioma o acceso á información e ao coñecemento,
- ofreza funcionalidades avanzadas de tecnoloxías da información en rede a todos os cidadáns e a un prezo asumible.

META-NET fomenta e promove as tecnoloxías plurilingües para todas as linguas europeas. As tecnoloxías permiten a tradución automática, a produción de contidos, o procesamento da información e a xestión do coñecemento para unha ampla variedade de aplicacións e de disciplinas. A rede pretende mellorar os enfoques actuais para que se poida establecer unha mellor comunicación e cooperación entre linguas. Todos os europeos teñen o mesmo dereito á información e o coñecemento, independentemente da súa lingua.

As tres liñas de acción de META-NET

META-NET presentouse o 1 de febreiro de 2010 co obxectivo de avanzar na investigación sobre a tecnoloxía lingüística. A iniciativa apoia unha Europa unida nun único mercado dixital e espazo de información. META-NET vén levando a cabo diversas actividades que promoven os seus obxectivos. META-VISION, META-SHARE e META-RESEARCH son as tres liñas de acción da rede.

META – A Alianza Tecnolóxica Plurilingüe de Europa

Figura 6: Tres liñas de acción en META-NET

META-VISION fomenta a construción dunha comunidade dinámica e influínte arredor dunha visión compartida e unha axenda estratéxica de investigación (AEI). O principal enfoque desta actividade é construír unha comunidade dedicada ás tecnoloxías lingüísticas en Europa que sexa coherente e cohesiva, reunindo aos representantes de grupos de interesados sumamente divididos e diversos. Durante o primeiro ano de META-NET, as presentacións no Foro FLaReNet (España), as Xornadas sobre tecnoloxía lingüística (Luxemburgo), JIAMCATT 2010 (Luxemburgo), LREC 2010 (Malta), EAMT 2010 (Francia) e ICT 2010 (Bélxica) centráronse na divulgación. Segundo as previsións iniciais, META-NET xa contactou con máis de 2.500 profesionais no eido das tecnoloxías lingüísticas para compartir con eles os seus obxectivos e visións. META-NET compartiu os resultados iniciais do seu proceso de elaboración dunha visión común no evento META-FORUM 2010, celebrado en Bruxelas, que contou con máis de 250 participantes. Nunha serie de sesións interactivas, os participantes intercambiaron opinións acerca das visións presentadas pola rede.

META-SHARE crea un portal aberto e distribuído para compartir e intercambiar recursos. A rede *peer-to-peer* de repositorios conterá datos da linguaxe, ferramentas, e servizos web documentados con metadatos de alta calidade e organizados en categorías estandarizadas. Pódese acceder facilmente aos recursos e buscalos de maneira uniforme. Os recursos dispoñibles inclúen materiais gratuítos e de código aberto, así como elementos restrinxidos, dispoñibles no mercado, mediante o seu pago. META-SHARE diríxese aos datos lingüísticos existentes, ferramentas e sistemas informáticos, así como aos produtos novos e emerxentes que se precisan para construír e avaliar novas tecnoloxías, produtos e servizos. A reutilización, combinación, adaptación e remodelación das ferramentas e dos datos lingüísticos xogan un papel crucial. META-SHARE converterase nunha parte fundamental do mercado das tecnoloxías lingüísticas para os desenvolvedores, expertos en localización, investigadores, tradutores e profesionais da linguaxe, tanto de empresas pequenas, como medianas e grandes. META-SHARE aborda o ciclo completo de desenvolvemento da tecnoloxía lingüística, dende a investigación ata os produtos e servizos innovadores. Un aspecto clave desta actividade é establecer META-SHARE como unha parte importante e valiosa dunha infraestrutura europea e global para a comunidade dedicada á TL.

META-RESEARCH constrúe pontes cara a campos tecnolóxicos relacionados. Esta actividade busca provocar avances noutros campos e

sacar partido da investigación innovadora que poida beneficiar á tecnoloxía lingüística. En concreto, esta actividade pretende dotar de máis semántica á tradución automática (TA), optimizar a división do traballo na TA híbrida, explotar o uso do contexto na tradución automática por ordenador e preparar unha base empírica para a TA. META-RESEARCH traballa con outros campos e disciplinas, como a aprendizaxe automática e a comunidade dedicada á web semántica. META-RESEARCH céntrase na recollida de datos, preparación de conxuntos de datos e organización de recursos lingüísticos para fins de avaliación; elaboración de inventarios de ferramentas e métodos; e organización de obradoiros e eventos de formación para os membros da comunidade. Esta actividade identificou claramente aspectos da TA onde a semántica pode influír nas boas prácticas actuais. Ademais, a actividade creou recomendacións sobre como tratar o problema da integración da información semántica na TA. META-RESEARCH está a finalizar un novo recurso lingüístico para a TA, o corpus anotado para unha mostra de TA híbrida, que proporciona datos para os pares de linguas inglés-alemán, inglés-español e inglés-checo. META-RESEARCH tamén desenvolveu software que compila corpus plurilingües que están ocultos na Internet.

Composición da Rede de Excelencia META-NET

País	Membro (Afiliación)	Contactos
Austria	Universität Wien	Gerhard Budin
Bélxica	Universidade de Antwerp	Walter Daelemans
	Universidade de Leuven	Dirk van Compernolle
Bulgaria	Academia de Ciencias de Bulgaria	Svetla Koeva
Croacia	Universidade de Zagreb	Marko Tadic
Chipre	Universidade de Chipre	Jack Burston
República Checa	Universidade Charles en Praga*	Jan Hajic
Dinamarca	Universidade de Copenhague	Bente Maegaard, Bolette Sandford Pedersen
Estonia	Universidade de Tartu	Tiit Roosmaa
Finlandia	Universidade Aalto*	Timo Honkela
	Universidade de Helsinki	Kimmo Koskenniemi, Krister Linden
Francia	CNRS, LIMSI*	Joseph Mariani
	ELDA*	Khalid Choukri
Alemaña	DFKI*	Hans Uszkoreit, Georg Rehm
	RWTH Aachen*	Hermann Ney
Grecia	ILSP, R.C. “Athena”*	Stelios Piperidis
Hungría	Academia de Ciencias de Hungría	Tamás Váradi
	Universidade Técnica de Budapest	Géza Németh, Gábor Olaszy
Islandia	Universidade de Islandia	Eirikur Rögnvaldsson
Irlanda	Dublin City University*	Josef van Genabith
Italia	Consiglio Nazionale Ricerche*	Nicoletta Calzolari
	Fondazione Bruno Kessler*	Bernardo Magnini
Letonia	Tilde	Andrejs Vasiljevs
	Universidade de Letonia	Inguna Skadina
Lituania	Instituto do Idioma Lituano	Jolanta Zabarskaitë

Luxemburgo	Arax Ltd.	Vartkes Goetcherian
Malta	Universidade de Malta	Mike Rosner
Países Baixos	Universiteit Utrecht*	Jan Odijk
Noruega	Universidade de Bergen	Koenraad De Smedt
Polonia	Academia de Ciencias de Polonia	Adam Przepiórkowski
	Universidade de Łódź	Piotr Pezik
Portugal	Universidade de Lisboa	Antonio Branco
	Instituto de Engenhería de Sistemas e Computadores	Isabel Trancoso
Romanía	Academia de Ciencias de Romanía	Dan Tufis
	Universidade Alexandru Ioan Cuza	Dan Cristea
Serbia	Universidade de Belgrado	Dusko Vitas, Cvetana Krstev, Ivan Obradovic
Eslovaquia	Academia de Ciencias de Eslovaquia	Radovan Garabik
Eslovenia	Instituto Jozef Stefan*	Marko Grobelnik
España	Barcelona Media*	Toni Badia
	Universidade Técnica de Cataluña	Asunción Moreno
	Universidade Pompeu Fabra	Núria Bel
Suecia	Universidade de Goteburgo	Lars Borin
Reino Unido	University of Manchester	Sophia Ananiadou

O * representa aos membros fundadores.

Como participar?

META-NET e META ofrecen diversas oportunidades de participación. Visita www.meta-net.eu para máis información sobre futuros eventos e actividades.

	Instituto Pupin	SanjaVranes