

# Ferramentas tecnoloxía lingüística

José Ramon Pichel Campos

imaxin|software



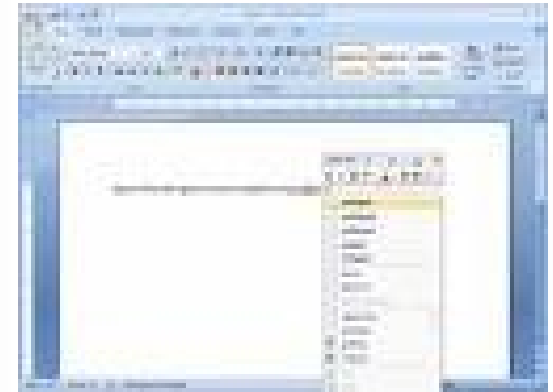
jornadas de lingua

de  
27/01 até 07/02  
Ourense 2009

A lingüística computacional é un campo multidisciplinar da lingüística e a informática que utiliza a informática para estudar e tratar a linguaxe humana.

Para logralo, tenta modelar de forma lóxica a linguaxe natural desde un punto de vista computacional.

Non se centra en ningunha das áreas da lingüística en particular, senón que é un campo interdisciplinar, no que participan lingüistas, informáticos especializados en intelixencia artificial, psicólogos cognoscitivos e expertos en lóxica, entre outros.



Algunhas das áreas de estudo da lingüística computacional son:

- \* Corpus lingüístico asistido por ordenador.
- \* Deseño de analizadores sintácticos (en inglés: parser), para linguaxes naturais.
- \* Deseño de etiquetadores ou lematizadores (en inglés: tagger), tales como o POS-tagger.
- \* Definición de lóxicas especializadas que sirvan como fonte para o Procesamiento de Linguaxes Naturais.
- \* Estudo da posible relación entre linguaxes formais e naturais.
- \* Tradución automática.



## Breve historia TIC e galego 01.

1994. O galego e as novas tecnoloxías. Mesa redonda CAF da FI da UDC

1995. Plataforma polo galego na informática (1º web feito na Galiza DES). Trillas formatadas

1996. Ciberirmandades da fala, ciberlingua, Vieiros [www.vieiros.com](http://www.vieiros.com)

1997. Allegue, Corrixe, imaxin software

1999. Codigo Cero [www.codigocero.com](http://www.codigocero.com), Software libre (Grupos de usuarios, corrector myspell Ramom Flores, etc)

2001. Portal galego da Lingua [www.agal-gz.org](http://www.agal-gz.org)

2002. Web social Blogs <http://todonada.blogspot.com>, VA-CA(Ridiculismo), Galiza Indymedia, blogaliza.

2003. Artigos tecnoloxía e lingua (Ramom Flores, Fernando Garea, Dario Janeiro, Suso Baleato, José Ramom Pichel)

## Breve historia TIC e galego 02.

2004-2006. Tradutores automáticos propietarios e de código aberto (Traduza-g, Tradutor Es-Ga, Opentrad).

2005-2007. Universalización redes sociais (blogs, facebook, tuenti, orkut, etc), software libre

2006-2008. Tradutores para diferentes pares de linguas en código aberto Opentrad: pares es-gl, gl-es, es-ca, ca-es, pt-gl, gl-pt.

2008-2009: corrector de linguaxe non sexista Exeria para OpenOffice.org, software libre para PEMES traducidos ao galego, Galinux (sistema operativo Linux en galego). Guia de estilo para as traducións, glosario de termos tecnolóxicos.

## empresa e lingua 01.

imaxin|software é unha empresa creada hai doce anos por catro titulados superiores en informática da FI da UDC.

Somos unha empresa de desenvolvemento de servizos e solucións avanzadas multilingües de software.

Vendor de *Microsoft Corporation*, sendo a única empresa galega e das poucas españolas que desenvolve software directamente para a multinacional americana.

Colaboramos desde o ano 2000 en proxectos de I+D en colaboración con universidades e centros tecnolóxicos.

Somos 26 profesionais, con mais do 70% titulados superiores e técnicos en informática, filoloxía, lingüística computacional e pedagogía.

Certificados en CMMI-3 na área de produción (só o 0,67% das empresas españolas teñen esta certificación internacional, en total 25 empresas en España)

## areas de negocio 02.

imaxin|context

Realización de solucións multilingüe de procesamento lingüístico e documental. imaxin|software aplica a tecnoloxía lingüística a través de ferramentas informáticas na xestión de contidos, xestión documental e xestión do coñecemento, mellorando a produtividade e optimizando a explotación dos recursos intanxibles da empresa.

Procesamento lingüístico: correctores ortográficos, sintácticos e de estilo, correctores ortográficos en rede, dicionarios electrónicos, **tradutores automáticos (<http://opentrad.imaxin.com>)**.

Procesamento documental: buscadores documentais, buscadores semánticos, sumarizadores de documentos, clasificadores automáticos documentais, extractores de información.

presentación de **empresa** 01.

01. 02. 03. 04. 05. 06. 07. 08. 09. 10.

# necesidades en **tecnoloxía lingüística dunha lingua**

## 04.

Básicos:

Recursos: lexicóns, dicionarios terminolóxicos, dicionarios multilingües.

Ferramentas: stemmers, etiquetadores, desambiguadores automáticos, gramáticas de dependencias, lematizadores, parsers superficiais.

Aplicacións: correctores ortográficos, correctores gramaticais, tradutores automáticos RBMT.

# necesidades en **tecnoloxía lingüística dunha lingua**

## 04.

Medios:

Recursos: corpus aliñados, corpus etiquetados, corpus comparábeis, thesaurus, WordNET.

Ferramentas: sumarizadores de textos, detectores de idiomas, detectores de entidades (nomes propios-datas).

Aplicacións: correctores ortográficos en rede, investigadores de información, clasificadores de información, recuperadores de documentos, extractores de información, tradutores automáticos EBMT.

# necesidades en tecnoloxía lingüística dunha lingua

## 04.

Avanzados:

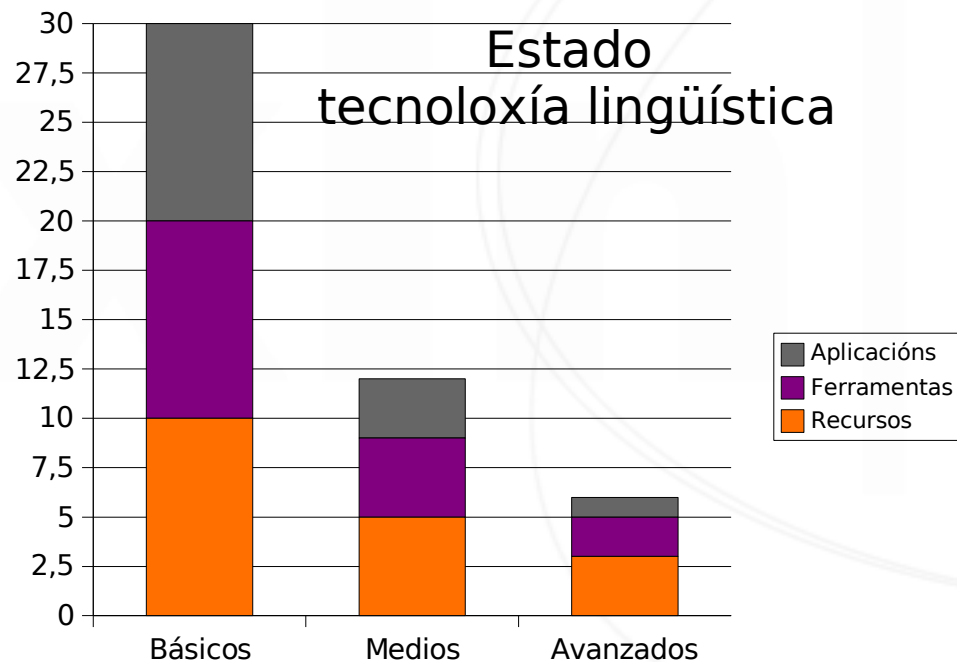
Recursos: corpus comparábeis.

Ferramentas: analizadores de discurso, chunkers, parsing profundo.

Aplicacións: tradutores automáticos EBMT e SMT (hibridación), clasificadores automáticos, recuperadores de documentos, e-learning con tecnoloxía lingüística, conversores texto-voz, analizadores de voz, correctores de linguaxe non sexista.

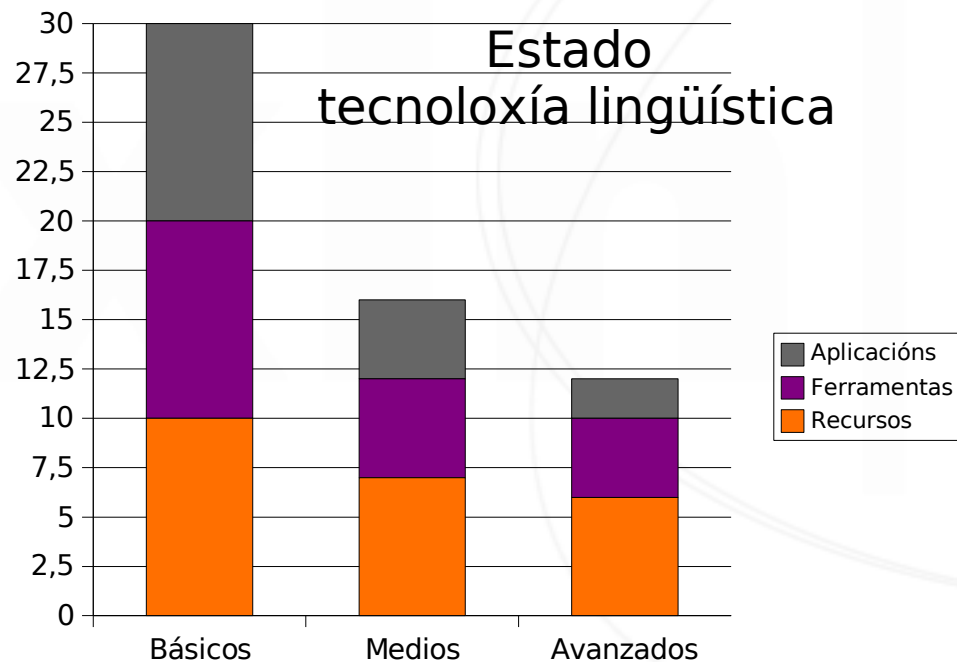
# estado da tecnoloxía lingüística en galego-portugués-español

Galego

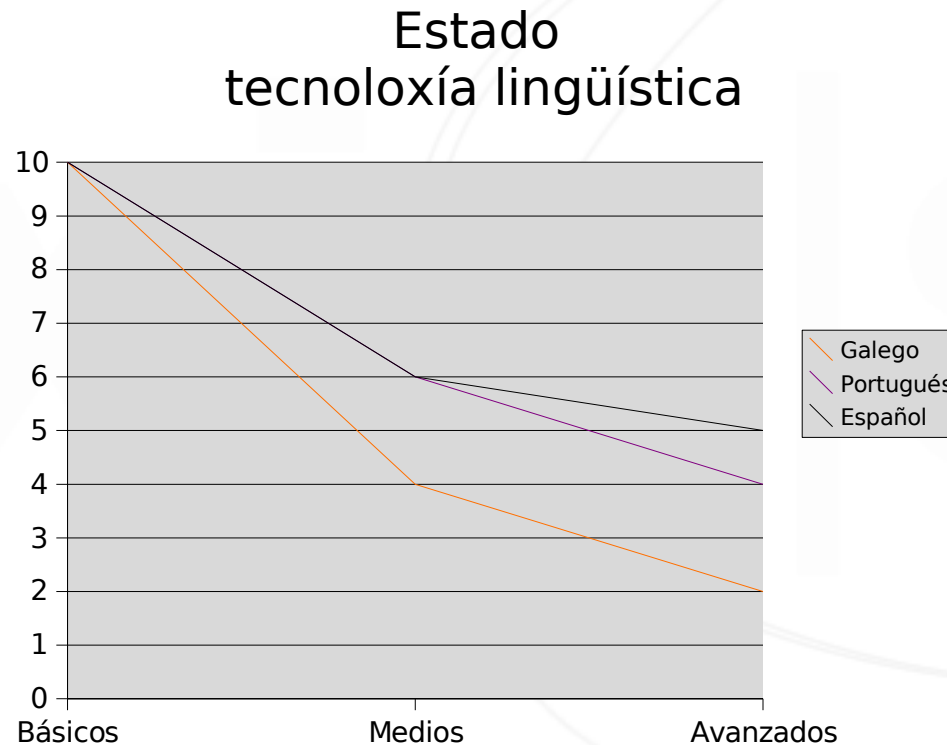


# estado da tecnoloxía lingüística en galego-portugués-español

Portugués-español



# estado da tecnoloxía lingüística en galego-portugués-español



## principais proxectos de tecnoloxía lingüística 04.

### Básicos (Galego)



Recursos: lexicóns (Opentrad, OpenOffice, [VOLg](#)), dicionarios terminolóxicos (SNL-USC, TERMIGAL), multilingües (Opentrad) e online ([IrIndo](#), [Xunta](#), [CRTVG](#), [Estraviz](#), [Galizionario](#))

Ferramentas: stemmers (UDC), etiquetadores (OT, Freeling), desambiguadores automáticos ([FreeLing galego](#)), lematizadores (OT, Freeling), parsers superficiais (Golfinho)

Aplicacións: Corrector imaxin [Galgo](#)

[Proofing Tools de galego](#) para Office 2000, XP, 2003, 2007

Corrector gramatical en código aberto [Golfinho](#) para OpenOffice.org

## principais proxectos de **tecnoloxía lingüística** 04.

Corrector ortográfico de **galego** para OpenOffice.org

Corrector ortográfico online (**Ortogal**)

## principais proxectos de **tecnoloxía lingüística** 04.

Implantación de OpenTrad <http://opentrad.imaxin.com> para o Sistema de Edición dixital da Voz de Galicia e o Sistema de Xestión documentación da Voz de Galicia <http://www.lavozdeg Galicia.com>

Implantación de OpenTrad para a redacción O Correo Galego e o diario Galicia Hoxe

Integración das melloras realizadas pola comunidade OpenTrad-Apertium nos paquetes. <http://www.traducindote.com/>  
<http://ousli.org/content/view/65/1/>

Participación no desenvolvemento de OpenTrad (español-galego), Europentrad (hibridación tecnoloxía RBMT e CBMT), EixOpentrad (galego-português).

## principais proxectos de tecnoloxía lingüística 04.

### Básicos (Portugués)



Recursos: lexicóns , dicionarios terminolóxicos, multilingües e online ( [Priberam](#), [Infopédia](#), [Michaelis](#), [Universal](#), [Vocabulário Acad. Bras.](#), [IDicionário Aulete](#), [KingHost](#))

Ferramentas: stemmers, etiquetadores, desambiguadores automáticos, lematizadores ([LX-Lem](#)), parsers superficiais ([CoGrOO](#)), analizadores morfolóxicos ([WebJspell](#)), bases de datos morfolóxicas ([MorDebe](#)), conxugadores verbais ([GConjugue](#))

Aplicacións: Correctores [FLIP](#)

Proofing Tools de português para Office 2000, XP, 2003, 2007

## principais proxectos de **tecnoloxía lingüística** 04.

Corrector gramatical en código aberto “CoGrOO” para OpenOffice.org

Corrector ortográfico de portugués para OpenOffice.org ([Vero](#))

Tradutores de código aberto ([Opentrad](#))

# necesidades en tecnoloxía lingüística dunha lingua

04.

Medios (Galego) :



Recursos: corpus paralelos e etiquetados ([Cluvi](#), UVigo), [TILGA](#) (ILGA-Imaxin), [CORGA](#), WordNET (Centro Ramón Piñeiro??), [base de datos terminolóxica](#) (UVigo)

Ferramentas: sumarizadores de textos (imaxin software), detectores de idiomas (Consortio Opentrad), detectores de entidades (nomes propios-datas) (Consortio Opentrad).

Aplicacións: correctores ortográficos en rede ([Imaxin](#)), investigadores de información, clasificadores de información ([Rede PL&IR](#)), recuperadores de documentos, extractores de información, tradutores automáticos EBMT (Consortio OpenTrad, ETSI Uvigo)

# necesidades en tecnoloxía lingüística dunha lingua

04.

Medios (Português):



Recursos: [Corpus Informatizado do Português Medieval](#) (Univ.Nova de Lisboa), [REDIP Corpus \(iLteC\)](#), [WordNet](#),  
córpora monolingües ([Corpus do Português](#), [Projecto AC/DC](#)),  
lexicográficos ([Corpus Lexicográfico](#), Univ.de Aveiro)

Ferramentas: detección de entidades ([Lingua::PT::ProperNames](#),  
Univ.do Minho), detectores de idiomas ([Lingua::Identify](#), Univ.do  
Minho), analizadores automáticos ([VISL](#), Syddansk Universitet)

Aplicacións: correctores ortográficos e sintácticos en rede ([Priberam](#)),  
pesquisadores de información, clasificadores de información,  
recuperadores de documentos, extractores de información

# necesidades en tecnoloxía lingüística dunha lingua

04.



Avanzados:

Recursos: corpus comparábeis (USC+Imaxin)

Ferramentas: analizadores de discurso, chunkers, parsing profundo ([Parser Multilingua](#), USC), extractores terminolóxicos multilingües ([GaleXtrac](#), USC)

Aplicacións: tradutores automáticos EBMT e SMT (hibridación), clasificadores automáticos, recuperadores de documentos, e-learning con tecnoloxía lingüística ([Português para nós](#)), sintetizadores de voz ([Cotovía](#), ETSI Uvigo+C.Ramón Piñeiro), aplicación de linguaxe non sexista ([Exeria](#)), Dicionario de Dicionarios e [Tesouro Medieval Informatizado da Lingua Galega](#) (ILGA+Imaxin).

presentación de **empresa** 01.

01. 02. 03. 04. 05. 06. 07. 08. 09. 10.

# necesidades en tecnoloxía lingüística dunha lingua

04.

Avanzados (Português):



Recursos: corpus paralelos bidireccionais (**COMPARA**),

Ferramentas: analizadores morfolóxicos (**jSpell**), córpora anotados con información lingüística (**CINTIL**, Univ.de Lisboa), ferramentas para tratamento de córpora paralelos (**NATools**, **O Constructor**)

Aplicacións: **Dicionário de Verbos de Português Medieval** (Univ.Nova de Lisboa), sintetizadores de voz (**Lingua::PT::Speaker**, Univ.do Minho)

# principais proxectos de i+d tecnoloxía lingüística 05.

## 2001-2004

PROXECTO I+D: Estudo e adquisición de recursos básicos de lingüística computacional do galego para a elaboración e melloría de aplicacións informáticas de tecnoloxía lingüística

## 2003-2005

PROXECTO I+D: Estudo de necesidades e xeración de recursos e ferramentas intelixentes en xestión da información e enxeñaría lingüística para a melloría das empresas exportadoras galegas

# principais proxectos de i+d 05.

## 2004-2005

2 PROXECTOS PROFIT. Tradución automática de código aberto para as linguas do estado español

Corpus bilingües CLUVI

## 2006-2007

EIXOPENTRAD: Tradución automática avanzada de código aberto para as linguas de Galiza e Portugal

# principais proxectos de i+d tecnoloxía lingüística 05.

2005-2009

**EurOpentrad.** Construción de tradutores automáticos en código aberto entre o inglés-galego, inglés-basco e inglés-catalán para a internacionalización das tres linguas usando a estratexia híbrida RBMT e SMT. <http://www.europentrad.com>

**GalinOpentrad e RecursOpentrad.** Construción de tradutores automáticos en código aberto entre o inglés-galego, inglés-basco e inglés-catalán para a internacionalización das tres linguas usando a estratexia RBMT <http://desarrollo.imaxin.com:7000/>

presentación de **empresa** 01.

01. 02. 03. 04. 05. 06. 07. 08. 09. 10.

# Tecnoloxía en galego 05.

Mancomun.org: software libre en galego para todas as necesidades.

GALINUX

SLI (Seminario Lingüística Informática) da Universidade de Vigo.

CLUVI

Instituto da Lingua Galega, Centro Ramón Piñeiro en Literatura e Humanidades

ETSI Enxeñaría informática (Universidade da Coruña)

ETSI Enxeñaría de telecomunicacións (Universidade de Vigo)

Dimensiona. Traduza-g

Tagen Ata. Parte lingüística corrector de linguaxe sexista exeria.

# Futuro tecnoloxía e lingua? 05.

Uso da tradución automática para xerar mais recursos.

Estamos a un moi bo nivel básico de ferramentas e moi dispersos no seguinte, partindo sempre de iniciativas individuais empresariais.

Falta de financiamento para aplicacións de gama alta.

Falta de recursos básicos e medios para atinxir proxectos grandes (problemas de produción de galego, falta de financiamento)

Linguas minorizadas = grandes custos para soportar os desenvolvementos tecnolóxicos (vantaxe potencial do Brasil e Portugal)

Aposta polo software libre como medio de favorecer as localizacións

Aproveitamento de recursos, ferramentas e aplicacións portuguesas e brasileiras sobretudo agora que chegaron a un acordo ortográfico (países BRIC)

# Futuro tecnoloxía e lingua? 05.

Estandarización da Terminoloxía galega que facilite a autonomía dos creadores.

SL+P+I: Software libre, aproveitamento de recursos do portugués e Innovación.

Que as Universidades e administracións públicas liberten os seus proxectos e recursos.

Polo de agora só o grupo da USC liderado por Paulo Gamallo Otero.

<http://gramatica.usc.es/~gamallo/index.html>

## Moitas grazas

José Ramom Pichel Campos,

Dtor. área de tecnoloxía lingüística

[jramompichel@imaxin.com](mailto:jramompichel@imaxin.com)

981 554 068

imaxin|software



jornadas de lingua

de  
27/01 até 07/02  
Ourense 2009