Estrategia para desarrollar tecnología de la lengua en lenguas con pocos recursos.

Euskara y quechua

IXA - Hinantin

ixa
Hinantin
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
NAZIOARTEKO BIKAINTASUN CAMPUSA
CAMPUS DE EXCELENCIA INTERNACIONAL

Presentación del grupo IXA

Grupo consolidado EJ tipo A (2010-2015)
Premio Abbadie 2014- Diputación de Gipuzkoa

- **61 miembros (33 doctores)**
  - **43 informáticos, 14 lingüistas,**
  - **3 técnicos de apoyo a la investigación**
- **30 profesorado permanente (UPV/EHU)**
- **9 post-doc**
- **16 contratos PDI**
- **3 estudiantes de grado**
- **3 técnicos de apoyo**

Presentación del grupo

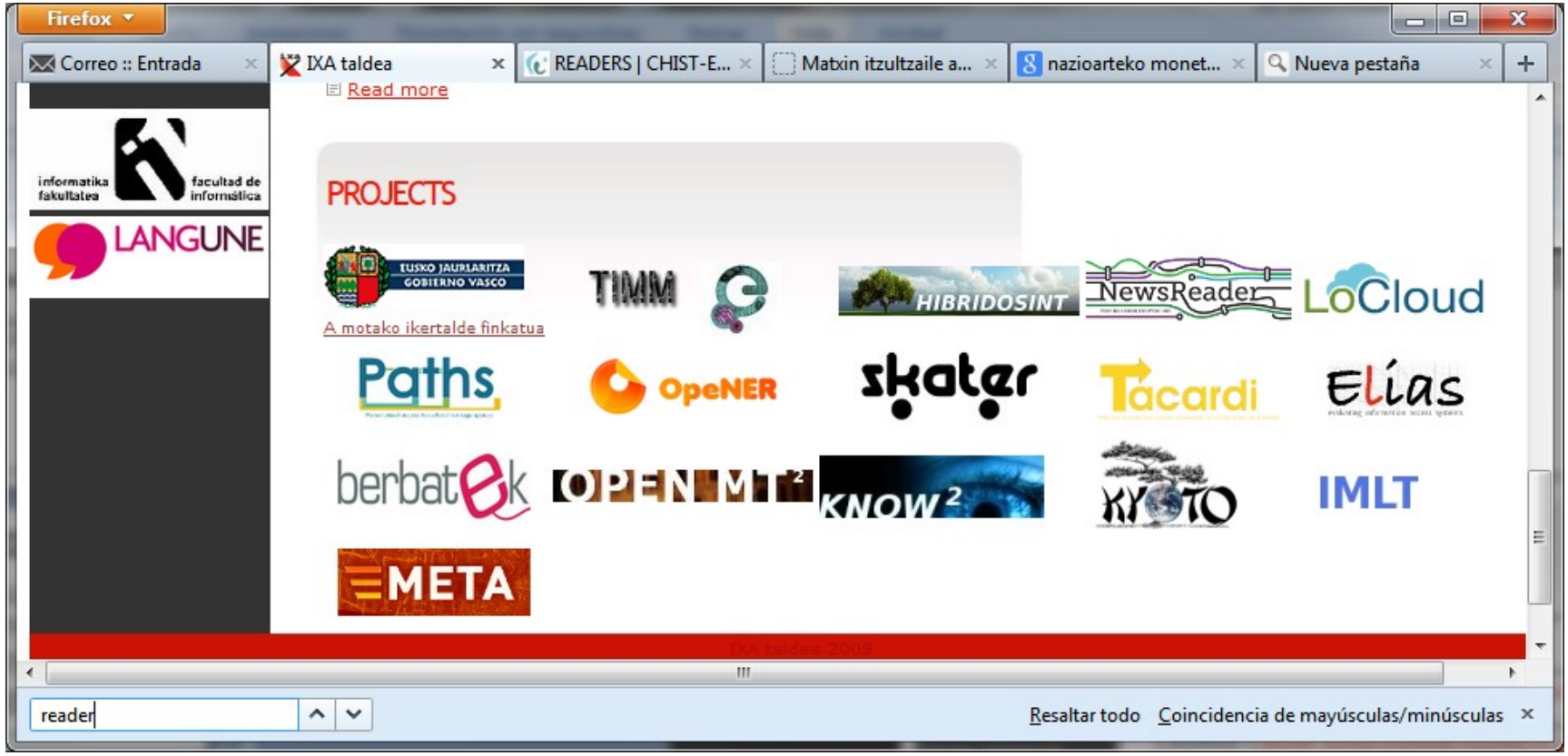
Grupo consolidado EJ tipo A (2010-2015)
Premio Abbadie 2014- Diputación de Gipuzkoa

- **61 miembros (33 doctores)**
  - **43 informáticos, 14 lingüistas,**
  - **3 técnicos de apoyo a la investigación**
- **30 profesorado permanente (UPV/EHU)**
- **9 post-doc**
- **16 contratos PDI**
- **3 estudiantes de grado**
- **3 técnicos de apoyo**

- **Máster**:
  HAP Análisis y Procesamiento del lenguaje/
  LAP Language Analysis and Processing
  (interdepartamental - 5 departamentos UPV/EHU)
- **Máster Erasmus-Mundus** (2013-2018)
  Language Technology and Communication
  (http://lct-master.org/)
  Universidades: Saarbrucken (coordinadora), Groningen, Trento, Lorraine, Malta, Charles de Praga, Melbourne, Shangai y UPV/EHU
- **Programas de doctorado**:
  HAP Análisis y Procesamiento del lenguaje
  (11 tesis leídas en los últimos 5 años)

- Aprox. 50 artículos anuales en congresos y revistas internacionales

- Proyectos activos en este momento:
  - Comunidad Europea: 6
  - Ministerio de Economia y Competitividad: 3
  - Eusko Jaurlaritza/Gobierno Vasco:
    1 ETORTEK (2012-2014)
    Grupo Consolidado (2010-2015)
  - Con otras entidades del entorno: 2


Firefox
Correo :: Entrada
IXA taldea
READERS | CHIST-E...
Matxin itzultzaile a...
nazioarteko monet...
Nueva pestaña
Read more
informatika fakultatea
facultad de informática
LANGUNE
PROJECTS
EUSKO JAURLARITZA GOBIERNO VASCO
A motako ikertalde finkatua
TIMM
HIBRIDOSINT
NewsReader
LoCloud
Paths
OpeNER
skater
Tacardi
Elías
berbatek
OPENMT²
KNOW²
KYOTO
IMLT
META
IXA taldea 2009
reader
Resaltar todo
Coincidencia de mayúsculas/minúsculas

- Ingeniería lingüística
  - Centrada en el euskera
    ... pero, también castellano, inglés, ...
  - Texto escrito
- Niveles de trabajo
  - Recursos: corpus, diccionarios, bases de conocimiento
  - Herramientas: procesamiento lingüístico (morfológico, sintáctico, semántico...)
  - Aplicaciones:

Gestión de contenidos:
Medicina, Turismo, Financiero, Cultural

Terminologia/ Lexikografia

Traducción automática

Enseñanza asistida por ordenador

Recursos y herramientas básicas

Aplicaciones

Traducción Automática:
- basados en reglas, corpus, híbridos
- Análisis morfológico, sintactico, semántico
- Análisis superficial, profundo

Gestión de contenidos: dominios financiero, turismo, medicina, cultural, …

Enseñanza: ayuda al aprendizaje de idiomas, especializado, …

Terminología/Lexicografía

Herramientas

Analizadores morfológicos, sintácticos, semánticos, discursivos, …

Basados en reglas, estadísticos, híbridos

Distintos dialectos y lenguas:
- Euskara, Inglés, Castellano, Quechua,

Recursos básicos

Diccionarios, Bases de datos léxicas

Corpus etiquetados:
- paralelos, monolingües
- morfológico, sintáctico, semántico, pragmático, ….

Recursos básicos
(http://ixa.si.ehu.es/Ixa/Produktuak)

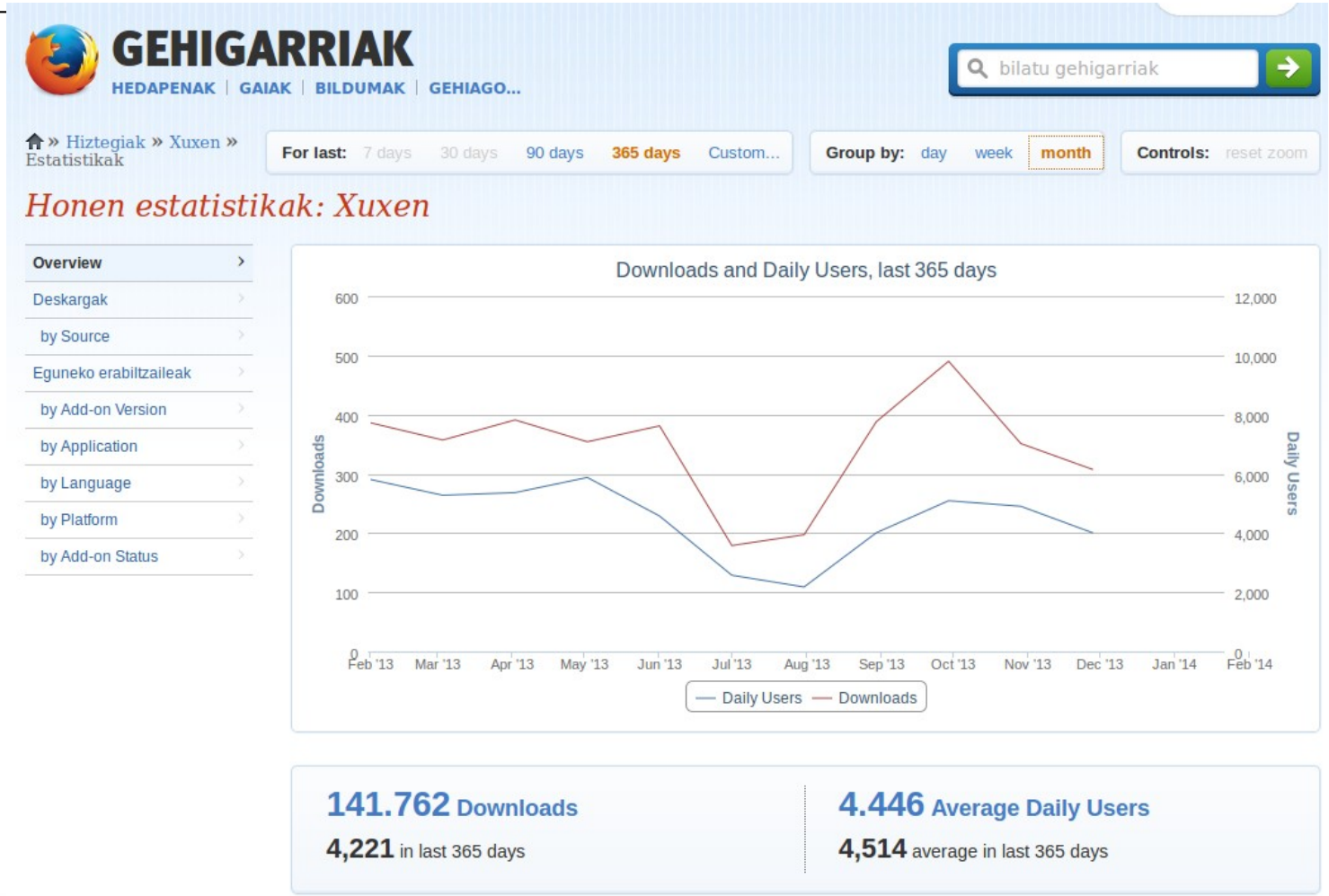
- EDBL:
  - Base de Datos léxica para el euskara
- EPEC:
  - Corpus de Referencia del euskera para el desarrollo de Aplicaciones lingüísticas basadas en análisis morfosintactico.
- ztC:
  - Corpus de Ciencia y Tecnología.
- EUSEMCOR:
  - Corpus de Referencia del euskera para el desarrollo de Aplicaciones lingüísticas basadas en análisis semántico.
- EuskalWordNet:
  - base de datos léxico semántica._

- MORPHEUS:
  - analizador morfológico
- EUSTAGGER:
  - lematizador/etiquetador morfosintáctico
- IXAti:
  - Analizador de sintagmas para el euskera (*chunker*)
- Malt-IXA:
  - analizador sintáctico
- EIHERA:
  - reconocedor/clasificador de entidades
- UKB:
  - Metodo gráfico de desambiguación de acepciones de palabras
- WSD-IXA:
  - desambiguador de acepciones

Aplicaciones (1)

(http://ixa.si.ehu.es/Ixa/Produktuak)

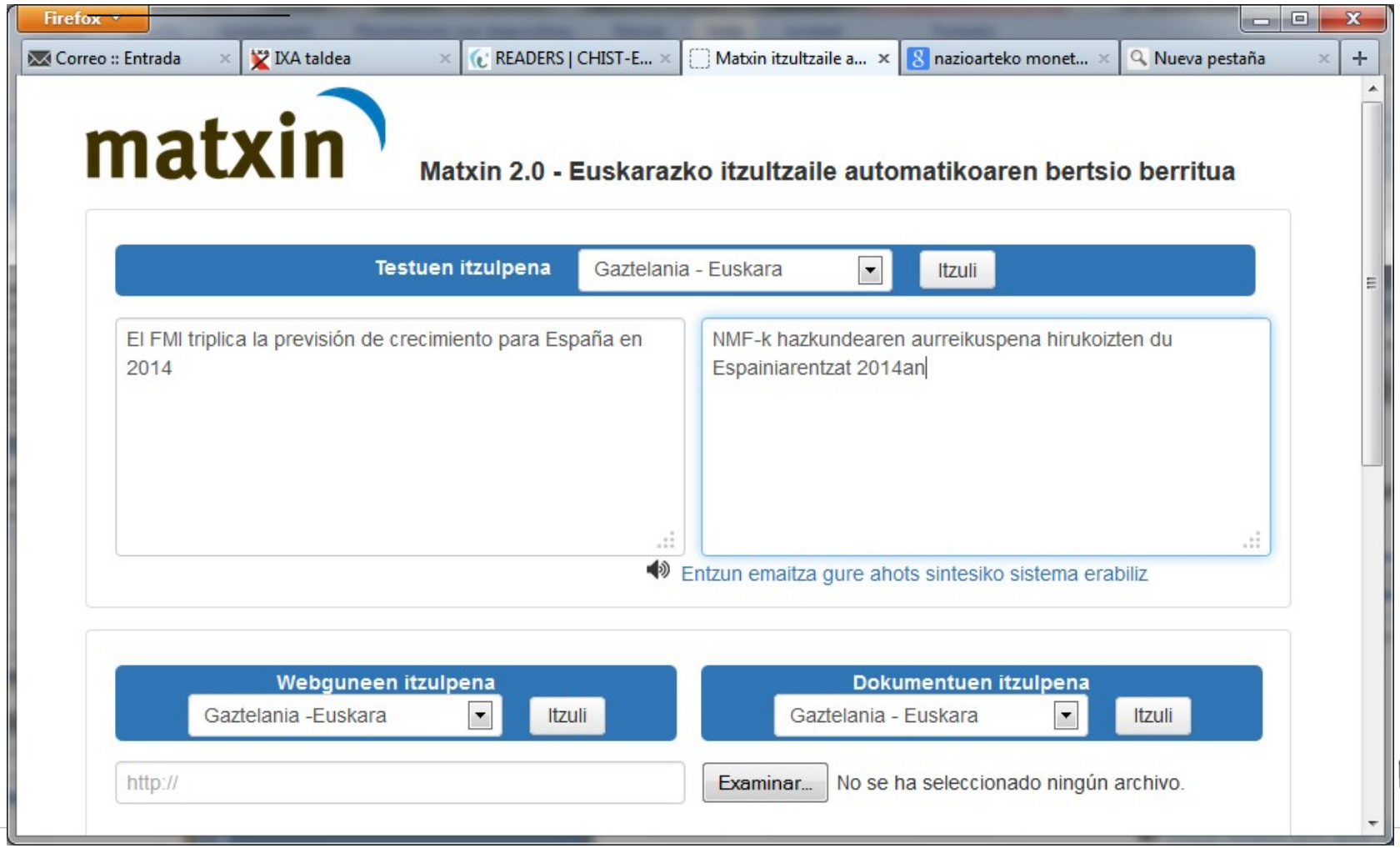
- **Xuxen: corrector ortográfico**



Aplicaciones (2)

(http://ixa.si.ehu.es/Ixa/Produktuak)

- **Matxin: Traductor español--> euskara**



- **_Proyectos europeos:**
  - **PATHS:** Personalised Access To cultural Heritage Spaces
  - **LoCloud:** Local content in a Europeana (cloud)
  - **OpeNER:** Open Polarity Enhanced Named Entity Recognition
  - **NewsReader:** Building structured event Indexes of large volumes of financial and economic Data for Decision Making
  - **Readers:** Evaluation And DEvelopment of Reading System
  - **QTLeap:** Quality Translation by Deep Language Engineering Approaches

- **Proyectos europeos: Gestión de contenidos**

  – **PATHS:** Personalised Access To cultural Heritage Spaces
  2011/02/01 - 2014/31/01
  – Participantes: University of Sheffield, MDR partners, i-Sieve, AskPlan
  – Dominio de aplicación: contenidos digitales de naturaleza cultural
  – Objetivo: Facilitar al usuario la tarea de explorar a través del material cultural disponible proponiéndole caminos de visita. Se trabaja sobre los documentos de museos y librerias digitales integrados en la red Europeana.eu, the European Library, Museum and Archive.

- Proyectos europeos: Gestión de contenidos

  - **LoCloud**: Local content in a Europeana cloud
    2013/03/02 - 2016/03/01
  - Dominio de aplicación: contenidos digitales de naturaleza cultural
  - Participantes: 32 partners, 28 paises
  - Objetivo: Desarrollar la tecnología y servicios en la nube para ayudar a instituciones locales de tamaño medio a añadir sus recursos digitales y hacerlos accesibles on-line via Europeana.eu , the European Library, Museum and Archive

- Proyectos europeos: Gestión de contenidos

  - **OpeNER**: Open Polarity Enhanced Named Entity Recognition

    2012/07/01 - 2014/06/30
  - Participantes: Vicomtech, VUA, CNR, Synthema, Olery.
  - Dominio de aplicación: turismo
  - Objetivo: Herramientas para el procesamiento lingüístico de libre distribución y fáciles de usar
  - Foco: análisis de opinión y polaridad de documentos
  - Ejemplo: extraer la opinión de los clientes sobre ciertos recursos (hoteles, …) en la Web

- Proyectos europeos: Gestión de contenidos

  – **NewsReader**: Building structured event Indexes of large volumes of financial and economic Data for Decision Making 2012/07/01 - 2014/06/30

  – Dominio de aplicación: financiero y económico

  – Participantes: VUA, Fondazione Bruno Kessler, Lexis Nexis (Holanda), ScraperWiki (Liverpool), UPV/EHU.

  – Objetivo: Síntesis de la información textual contenida en grandes volúmenes de documentos como base para la toma de decisión en el ámbito financiero y económico

- Proyectos europeos: Gestión de contenidos

  - **Readers**: evaluation and development of reading systems
    2013/07/01 - 2015/06/30
  - Dominio de aplicación: general
  - Participantes: UNED, Synapse Développement (France), Universidad de Edimburgo, UPV/EHU.
  - Objetivo: Extracción de conocimiento de manera automática a partir de grandes cantidades de texto no estructurado.

- Proyectos europeos: Traducción automática

  – **QTLEAP**: Quality Translation by Deep Language Engineering Approaches 2013/11/01 - 2016/10/31

  – Dominio de aplicación: financiero y económico

  – Participantes: Saarbrucken, Lisboa, DFKI, Universidad Charles de Praga, Universidad de Sofia, UPV/EHU

  – Objetivo: Mejorar la traducción automática mediante procesamiento semántico y sintaxis profunda

- Dominio de aplicación: informes médicos
- Participantes:
  - Hospital Galdakao-Usansolo,
  - BIOEF (Fundación Vasca de Innovación e Investigación Sanitarias),
  - I3B (Instituto Ibermática Innovación),
  - IXA Taldea.
- Objetivos:

  OSAKU: Codificación automática de diagnósticos médicos (CIE-9)

  SENDAUR: Detección de efectos adversos a medicamentos.

  Además, traducción semiautomática de SNOMED CT al euskara

Colaboraciones

(http://ixa.si.ehu.es/Ixa/Estekak))

- Consorcio ETORTEK:
  - VICOMTech, Elhuyar, Robotiker, Aholab
- Consorcio Proyecto traducción automática:
  - UPC, Universitat d'Alacant, Universidad del País Vasco, Universidad de Vigo, Fundación Elhuyar, Eleka S.L., imaxin |software
- Proyectos MEC:
  - Universidad Politécnica de Cataluña, Universidad de Barcelona, UNED, Universidad de Alicante, Universidad de Vigo, Pompeu Fabra,
- Consorcio proyecto "Observatorio léxico"
  - Euskaltzaindia, UZEI, Fundación Elhuyar
- Pertenencia a redes
  - TIMM (Tratamiento de Información Multilingüe y Multimodal),
  - RTH (Red temática de tecnologías del habla)

- Information and Communication Technologies
  - Advanced Computing
    - ICT-7 Advanced Cloud Infrastructures and Services
    - ICT-9: Tools and Methods for Software Development
    - …
  - Content technologies and information management
    - ICT-15 Big data and Open Data Innovation and take-up
    - ICT-17 Cracking the language barriers
    - ICT-20 Technologies for better human learning and teaching
    - ICT-21 Advanced digital gaming/gamification technologies
    - ICT-22 Multimodal and Natural Computer Interaction

Hinantin

Dos proyectos entre la Universidad Nacional San Antonio Abad del Cusco (UNSAAC) y la Universidad del País Vasco (UPV/EHU).

1) "***Primera aproximación al procesamiento automático del Quechua. Corpus, morfología y léxico. Bases para un corrector ortográfico***", financiado por el Ministerio de Asuntos Exteriores y Cooperación de España en 2012 (AECIP),

2) "**RUNASIMI Recursos básicos para el procesamiento automático de la lengua quechua: base de datos léxica y corpus textual**", financiado por la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) en 2013.

·Colaboración con **Institut für Computerlinguistik** de la Universidad de Zurich (UZH)

Hinantin

- Formación inicial de un grupo de investigadores en la ciudad del Cusco conformado por docentes y estudiantes de la Universidad Nacional San Antonio Abad del Cusco.

- Recopilación y organización de un corpus textual inicial.
- Versión inícial de la Base de Datos Léxica del Quechua.
- Primera versión de un corrector ortográfico.
- Aproximaciones iniciales
    - conversión de texto a voz,
    - un analizador sintáctico
    - un sistema de enseñanza de la lengua.

- Formación inicial de un grupo de investigadores en la ciudad del Cusco conformado por docentes y estudiantes de la Universidad Nacional San Antonio Abad del Cusco.

- Recopilación y organización de un corpus textual inicial.
- Versión inícial de la Base de Datos Léxica del Quechua.
- Primera versión de un corrector ortográfico.
- Aproximaciones iniciales
    - conversión de texto a voz,
    - un analizador sintáctico
    - un sistema de enseñanza de la lengua

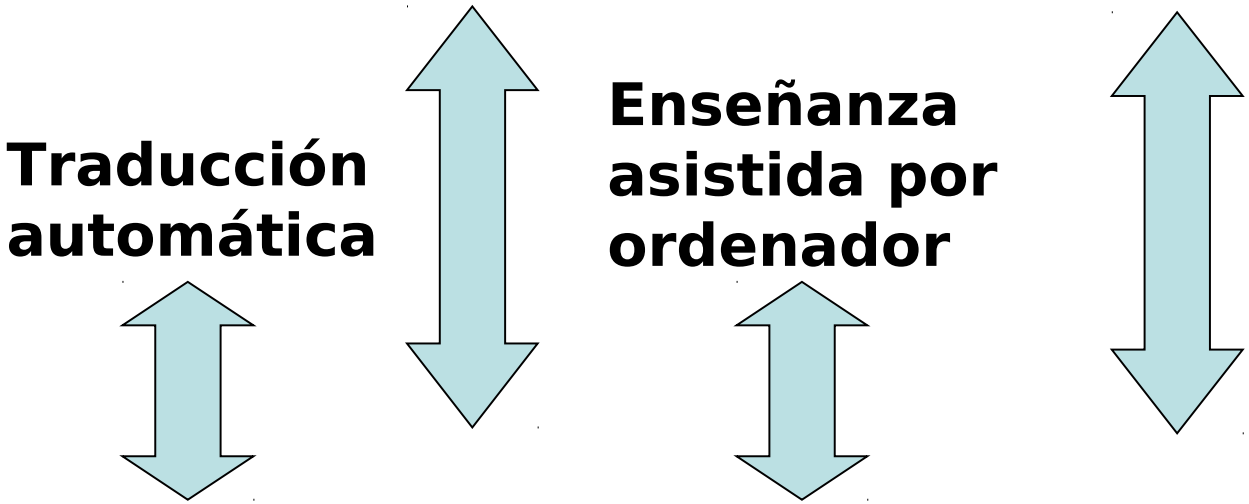
http://hinantin.com/

- Ingeniería lingüística
  - Centrada en el QUECHUA
    ... pero, también castellano, inglés, ...
  - Texto escrito
- Niveles de trabajo
  - Recursos: corpus, diccionarios, bases de conocimiento
  - Herramientas: procesamiento lingüístico (morfológico, sintáctico, semántico...)
  - Aplicaciones:

Hinantin

Gestión de contenidos:
Medicina, Turismo, Financiero, Cultural

Terminologia/ Lexikografia



Recursos y herramientas básicas

Aplicaciones

Traducción Automática:
- basados en reglas, corpus, híbridos
- Análisis morfológico, sintactico, semántico
- Análisis superficial, profundo

Gestión de contenidos: dominios financiero, turismo, medicina, cultural, …

Enseñanza: ayuda al aprendizaje de idiomas, especializado, …

Terminología/Lexicografía

Herramientas

Analizadores morfológicos, sintácticos, semánticos, discursivos, …

Basados en reglas, estadísticos, híbridos

Distintos dialectos y lenguas:

Euskara, Inglés, Castellano, Quechua,

Recursos básicos

Diccionarios, Bases de datos léxicas

Corpus etiquetados:
- paralelos, monolingües
- morfológico, sintáctico, semántico, pragmático, ….

¿Qué hacemos?

Estrategia para desarrollar tecnología de la lengua en lenguas con pocos recursos.

Euskara y quechua

IXA - Hinantin

ixa
Hinantin
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
NAZIOARTEKO BIKAINTASUN CAMPUSA
CAMPUS DE EXCELENCIA INTERNACIONAL