RUNASIMI

RECURSOS BÁSICOS PARA EL PROCESAMIENTO AUTOMÁTICO DE LA LENGUA QUECHUA: BASE DE DATOS LÉXICA Y CORPUS TEXTUAL

UNSAAC (Cusco, Peru)
Ixa Taldea (UPV/EHU)
2013/2014

Introducción

- Este proyecto ha pretendido dinamizar el grupo de ingeniería lingüística de Cusco cuyo objetivo es abrir una línea de investigación en el tratamiento automático del quechua en Cusco (procesamiento del lenguaje natural, PLN).

- El objetivo principal consiste en el fortalecimiento institucional del grupo de trabajo local (Hinantin), para que, a partir de esta iniciativa, se promuevan la utilización de las nuevas tecnologías en el tratamiento del lenguaje.

Precedentes

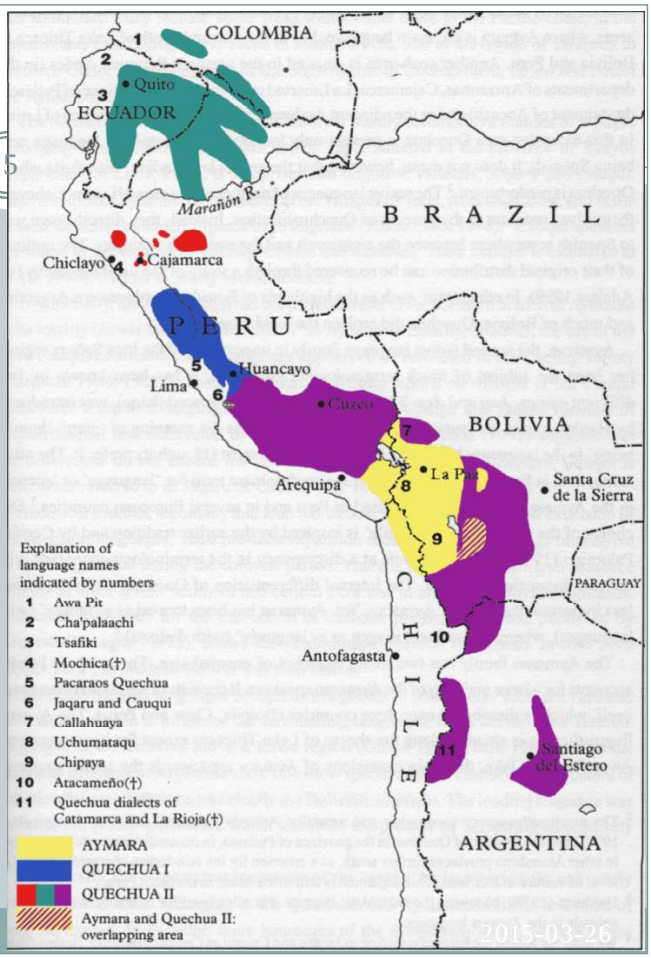
- Primer contacto con Juan Cruz (2010), por medio de profesores de la facultad que viajan a Cusco en el contexto de un máster (*Master Universitario en Ingeniería Computacional y Sistemas Inteligentes*) que se impartía allá.
- Proyecto de fin de máster de Rosmary Jiménez (UNSAAC), dirigido por Iñaki Alegria y Olatz Arregi (UPV/EHU).
- Proyecto de coooperación financiado por la AECID (2011), donde se empieza a trabajar con las personas del grupo en Cusco; de la EHU viajamos allá en varias ocasiones, impartimos cursos, etc.

Grupos de trabajo

- UNSAAC
  - Richard Castro
  - Juan Cruz (coord.)
  - Rosmary Jiménez
  - José Martí
  - Luis Palma
  - Hugo Quispe
  - ...
- Ixa (UPV/EHU)
  - Olatz Arregi
  - Xabier Artola (coord.)
  - Kepa Sarasola

Quechua

- El quechua o Runasimi fue la lengua oficial dentro de la cultura Inca.
- Existen entre 8 a 10 millones de hablantes como macro lengua (Colombia, Ecuador, Perú, Bolivia, Chile y Argentina).
- En la actualidad, según la UNESCO, el quechua se encuentra en “situación crítica”.
- Es una familia de lenguas, con diferentes dialectos y modos de escritura.
- Variedad cusqueña o Quechua II-C (Alfredo Torero) con 1.500.000 hablantes.
- Dentro de la variedad cusqueña existen dos modos de escritura: trivocálico y pentavocálico.
- El quechua es una lengua aglutinante.



Objetivos

- Consolidación del grupo de Ingeniería Lingüística del Cusco (Hinantin)
  - Docentes de la UNSAAC, estudiantes pre-doctorales, magísteres, estudiantes de grado y lingüistas quechua hablantes.
  - Apoyo: grupo Ixa (UPV/EHU) e Institute of Computational Linguistics (UZH).
- Creación de recursos básicos para el PLN:
  - Base de datos léxica (QLDB)
    - Componente básico de los sistemas de PLN.
    - 20.000 entradas.
  - Corpus textual
    - Recopilación de textos para diferentes tareas del tratamiento automático de la lengua.
    - 200.000 palabras.

Base de datos léxica (QLDB)

- Recopilación de las unidades léxicas de la lengua junto con sus propiedades, a fin de ser utilizada en diferentes tareas de PLN, especialmente en la morfología computacional.
- Diseño e implementación (MySQL).
- Número total de entradas : 9.772.
- Fuentes:
  - Analizador morfológico de Annette Ríos.
  - Diccionario de la Academia Mayor de la Lengua Quechua (AMLQ).
- Dificultad de obtener otras fuentes en formatos fácilmente importables.
- QLDB

Corpus textual

- Quechua cusqueño principalmente.
- 85 documentos: 618.913 palabras (Biblia).
- Se ha tokenizado y lematizado el corpus.
- Se ha indexado el corpus.
- Se está preparando una aplicación web de consulta del mismo.
- Dificultad de obtener documentos en formato digital.
- Falta de corpus paralelos.

Impacto del proyecto

- Las personas beneficiarias
  - Se ha consolidado el grupo con los participantes en el proyecto, que continua con las tareas comenzadas y que está preparando un proyecto de envergadura para presentarlo en la universidad (UNSAAC).
- La comunidad universitaria
  - Cuatro personas de Cusco han efectuado estancias en la UPV/EHU (Ixa, Aholab).
  - Hugo Joel Quispe está realizando su tesis doctoral en la UPV/EHU.
  - Richard Castro ha solicitado la matrícula en el máster Erasmus Mundus LCT (en conjunción con el máster HAP/LAP).
  - Rosmary Jiménez presentó su proyecto fin de máster en la Facultad de Informática.
  - Las personas que han venido a la UPV/EHU han impartido sendos seminarios durante su estancia en el grupo de investigación Ixa.

Impacto del proyecto

- La investigación
  - Las diferentes tareas realizadas han dado lugar a la publicación de varios artículos de investigación y presentaciones en congresos internacionales.
  - El trabajo llevado a cabo en el proyecto se imbrica perfectamente en el área de investigación del grupo Ixa al que pertenecemos los solicitantes.

Seminario en la FISS

Mintegia: Kitxuaren prozesamendurako lehen hurbilketa (2012/11/15)

2012/11/12-n argitaratuta

Bada ia urtebete Ixa Taldea ea Cuscoko UNSAAC unibertsitateko Juan Cruz ikertzailearen artean kitxuaren prozesamenduari ekiteko lanean hasi ginela. Euskara eta kitxua biak baliabide gutxiko hizkuntzak direnez eta morfologia antzekoa dutenez, euskara normalizatzeko eta bere erabilera errazteko azken 20 urtetan hemen egin ditugun tresnak eta aplikazioak baliagarri izan daitezke kitxuaren kasuan ere. Madrileko "Ministerio de Asuntos exteriores y Cooperación"-en proiektu bat izan dugu 2012 urtean: Lehen urratsak Quechua-ren prozesaketa automatikoan. Corpus, morfologia eta lexikoa. Proiektu horren barruan Kepa, Xabier eta Olatz Cuscon egon gara urtean zehar, eta irailetik hona bisitan dauzkagu Hugo Quispe eta Richard Castro. Hugo datu base lexikal bat garatzen ari da kitxuarako, eta Richard hizketa sortzeko beste sistema bat eraiki du Bilboko Aholab laborategian. Richard-ek datorren astean Iberspeech2012 kongresuan demo bat aurkeztuko du hizketa sortzeko eginda zeukaten beren lehen sistemarekin. Osteguneko mintegi-saioan proiektu honen barruan egin dena azalduko dugu.

Gaia: Kitxuaren prozesamendurako lehen hurbilketa
(Primera aproximación al procesamiento automático del Quechua)
Hizlaria: Hugo Quispe, Richard Castro (UNSAAC unibertsitatea), Olatz Arregi, Xabier Artola eta Kepa Sarasola (Ixa Taldea)
Eguna: azaroaren 15ean, osteguna
Ordua: 16:00-17:00
Tokia: 3.2 aretoa. Informatika Fakultatea

Laburpena:
El Quechua o "Runasimipi" como lengua oriunda de la cultura Inca en el Perú, es una familia de lenguas en Latinoamérica. La situación actual de

http://www.unibertsitatea.net/blogak/ixa/

Proyecto de futuro

- Solicitud de un proyecto a la UNSAAC
  - Se pretende solicitar un proyecto de varios años en la UNSAAC para poder financiar las actividades del grupo y a ser posible liberar a dos personas para que lideren los trabajos a realizar.
  - Se ampliarán la QLDB y el corpus textual, se mejorarán las herramientas construidas, y se recopilará un corpus hablado del quechua. Además se construirán nuevas herramientas básicas tales como analizadores morfológicos y sintácticos, bases de conocimiento, corpus paralelos etc.
  - El proyecto está prácticamente redactado pero hay algún problema de forma que están intentando solucionar.
  - En todos los contactos que hemos mantenido con el vicerrectorado de la UNSAAC, se nos ha dicho que hay grandes posibilidades de conseguir financiación para el proyecto.

Conclusiones

- En el proyecto nos hemos aproximado bastante a los objetivos fijados: se han construido diversas herramientas, se han formado personas, se ha constituido un (embrión de) grupo de investigación (ver http://hinantin.com)...
- Dadas las circunstancias específicas del personal perteneciente al grupo (trabajo fuera de la universidad, poco tiempo libre, jubilación del responsable, etc.), el ritmo de trabajo no es el que nos hubiera gustado.
- ¿El futuro?
  - Depende sobre todo del grupo de Cusco.
  - La obtención del proyecto a solicitar en la UNSAAC sería un paso importante.