Collaborative corpus building for minoritized languages using wiki-technology.

Documenting the Asturian language

Johann Ari Larsson  
Computer Science Department  
Brandeis University  
johann@cs.brandeis.edu

Roser Sauri  
Computer Science Department  
Brandeis University  
roser@cs.brandeis.edu

Xulio Viejo  
Department of Spanish Philology  
Oviedo University  
jviejo@uniovi.es

Overview

The Eslema project focuses on systematically documenting Asturian in an effort to conserve, and increase the generational transmission of this minoritized Romance language. For such a small community of roughly 300,000 speakers, research funding and resources are in short supply. Thus, we have built a wiki-based workspace that enables the entire community to collaboratively collect, annotate, and share texts online. This will benefit both Eslema, allowing to deliver it at a minimum cost, and the Asturian community, causing a stronger presence of Asturian in the information technologies area and, as a consequence, boosting the confidence of speakers in their language, which will hopefully contribute to slow down the serious process of substitution it is currently undergoing.

The Asturian Language

- Currently spoken in:  
  - Most of the Principality of Asturies  
  - The provinces of Lleóns and Zamora in Portugal  
  - The district of Miranda do Douro  

- Asturian is also known as:  
  - Astur-Leonese  
  - Bable  
  - Mirandes (only in Portugal)

- Speakers number (estimated): 100,000 speakers  
- (over a population of 1,000,000 people)

Eslema is the first corpus created for Asturian (http://www.uniovi.es/eslema)  
- Its goals are:  
  - Documenting the linguistic tradition of Asturian and its historic evolution.  
  - Contributing to the processes of codification and normalization of this language, and therefore aiming at reversing the current diglossic situation.

- Given the limited resources available to a minority language such as Asturian, Eslema suffers from lack of funding and scarcity of collaborators. This situation has prompted us to look for innovative ways in building the corpus, among which the use of wiki-technology.

Why wikis?

- Favorable technological situation  
- Despite Asturian rural context, there has been a significant increase in the consumption of information technology in Asturies; e.g., Internet access in Asturian homes:  
  - (2002): 21.38%  
  - (2006): 41.45% (Spanish average: 39%)
- Favorable social conditions  
- As speakers of a minority language, they feel fairly committed to keeping Asturian alive.  
- There is interest in that Asturian be present in modern technological environments of international projection; e.g.,  
- Wikipedia: Out of 265 languages present in Wikipedia, Asturian is ranked:  
  - 21st place (from Feb, 2009)  
- Using wiki-technology for collaboratively building Eslema

Using wiki-technology for collaboratively building Eslema

WikiDesignPlatform (WDP)  
- A website focusing on fast co-construction and sharing of free-form textual documents (wikis)  
- Philosophy: together we make it better than individually:  
- A simple interface and markup language enables experts and non-technical users to easily participate  
- Widely adopted to facilitate mass-collaboration, e.g., Wikipedia.

WikiDesignPlatform (WDP)  
- The out-of-the-box standard wiki is not structured to fully support a task of this caliber.  
- On the other hand, the WikiDesignPlatform  
  - Provides a set of navigation, communication, and awareness devices that can be coupled with the standard wiki so as to custom-build workspaces for collaborative tasks of different kind.  
- Developed at Brandeis University (Larusson and Sauri, 2005)  
- So far, WDP-based wikis have been deployed to support various collaborative learning activities. Herein, Eslema exploits its capabilities within a research-oriented setting.

References

